Optimal elicitation of probabilistic information from experts

ANDREW CURTIS[1,2] & **RACHEL WOOD**[1,3]

[1] *Schlumberger Cambridge Research, High Cross, Madingley Road, Cambridge CB3 0EL, UK*
(e-mail: curtis@cambridge.oilfield.slb.com)

[2] *Grant Institute of Earth Science, School of GeoSciences, University of Edinburgh, West Mains Rd., Edinburgh EH9 3JW, UK*

[3] *Dept. of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, UK*

Abstract It is often desirable to describe information derived from the cumulative experience of experts in a quantitative and probabilistic form. Pertinent examples include assessing the reliability of alternative models or methods of data analysis, estimating the reliability of data in cases where this can not be measured, and estimating ranges and probable distributions of rock properties and architectures in complex geological settings.

This paper presents a method to design an optimised process of elicitation (interrogation of experts for information) in real time, using all available information elicited previously to help in designing future elicitation trials. The method maximises expected information during each trial using experimental design theory. We demonstrate this method in a simple experiment in which the conditional probability distribution or relative likelihood of a suite of 9 possible 3-D models of fluvial-deltaic geologies was elicited from a geographically remote expert. Although a geological example is used, the method is general and can be applied in any situation in which estimates of expected probabilities of occurrence of a set of discrete models are desired.

Introduction

The generic task of gathering quantitative, probabilistic information on any topic usually requires the interrogation of experts in that field. Usually such an interrogation revolves around some model or representation of the topic of interest. Expert judgement and intuition is used to place constraints both on the range of complexity that should be included in the model, and on the range of model parameter values to be considered. Everyone can recall situations in which experts (including ourselves) have poorly estimated such information, or in which information from multiple experts was inadequately assimilated, so resulting in inefficiency or errors.

Poor results usually occur because all humans – experts included – are subject to natural biases when trying to estimate probabilities or risks mentally. Some of these are shared with other mammals, others only with primates, and some appear to be exclusively human. These biases, and techniques to ameliorate their effects, have been the subject of much study and debate in psychological and statistical research literature, and form the field known as *Elicitation Theory*, that is, the theory concerning the interrogation of subjects for information. Surprisingly however, few elicitation techniques appear to be used in the geosciences, and so one aim of this paper is to review some of the most pertinent elicitation literature.

In order to develop the theory further, we here consider explicitly a generic situation often encountered in the hydrocarbon exploration and production industry, where estimates of the detailed stratigraphic architecture and properties of reservoir rocks in 3-dimensions (henceforth, the reservoir *geology* in *3-D*) are required, using only remote sensing methods which have either lateral or vertical spatial resolution that is poorly matched to this task (e.g., seismic and well log data). Wood & Curtis (this volume) demonstrate how quantitative geological information that is independent from the remotely-sensed data, called *prior information*, can be used to constrain probabilistically the architecture of the geology some distance from a single vertical well at which (in that study) the only available data was collected. How best to elicit the prior information that is used implicitly by experts in placing such constraints is the subject of this study. Once captured, such information might be reused in multiple similar situations.

Consider a case where an expert geologist knows the typical characteristics of the geology in an area of interest, e.g., the likely range either of sinuosity of river channels in a fluvial system, or of different reef types within carbonate strata. Assume also that a computer program exists that creates synthetic geological cross-sections given a set of input parameters (e.g., Tetzlaff & Priddy 2001; Griffiths *et al.* 2001; Burgess this volume; Tetzlaff this volume). Boschetti & Moresi (2001) demonstrated how a genetic algorithm (GA) might be used to find values of the computer program's input parameters that produce the desired characteristics in the modelled geology (see also Wijns *et al.* this volume). The expert simply ranks sets of cross-sections from different modelling runs in order of how well they match the desired characteristics; the GA iterates through further generations of modelling runs, requesting further rankings from the expert, until sets of input parameters are found such that geologies with the desired characteristics are modelled.

Both the strength and weakness of the method of Boschetti & Moresi (2001) is that the interpreter never need produce quantitative estimates of *how* well or poorly any model run matches the desired characteristics – only relative ranks of the models are requested. This property renders the method easy to use and widely applicable, but has the potential drawback that it cannot deliver quantitative, probabilistic information about the range of model parameters that fit the desired characteristics. It also implies that only optimisation algorithms that do not require probabilistic inputs (such as the GA) can be used.

We extend these methods by asking the expert to provide a minimal set of probabilistic information from their prior experience. This allows the construction of fully normalised, probability distributions representing geological prior information.

We begin this paper with an overview of background material. This includes a discussion of the general nature of geological prior information that one might collect and use quantitatively, and a description of a particular geological system about which we require information (the problem addressed later in this paper). Elicitation theory has a large associated literature in engineering, statistical and psychological literature, so we also provide a brief overview of some of this work in the background material section. In the subsequent section we present a new method for collecting the required information. This method employs techniques from the field of Statistical Experimental Design, and

these are described in a separate section. We then demonstrate the benefits of this approach in a simple geological elicitation experiment, and discuss how to apply the method to solve the problem posed in the background material section below.

Overview of Background Material

The nature of geological prior information

Geological prior information is often qualitative. This is usually because it is based on interpretations, best guesses from past individual experiences, and intuition. Although useful to obtain other qualitative results, such information cannot be used directly in any other probabilistic methodology without implicitly or explicitly quantifying this qualitative knowledge (Wood & Curtis this volume).

Additionally, almost all geological prior information is partly subjective. This is a direct consequence of the fact that prior information usually comprises generalisations of previous experience. There are exceptions to this: for example, in a seismic survey, pertinent prior information might include previously acquired surface gravity measurements, or measurements derived from wells, and that information may be objective. Such objective information is relatively easy to capture; however, less easy to capture is geological prior information derived from tacit experience. For example, information about formations and rock characteristics expected in any particular type of geology usually consists of patterns observed previously in outcrops, which have been generalised by experts. Each expert may generalise their various experiences of different outcrops in markedly different ways. Therefore, estimates of prior probabilities are themselves necessarily inaccurate due to uncertainty and biases implicit within each expert from whom prior information was elicited. Hence, it is necessary to understand the kinds of biases and uncertainties that humans tend to exhibit when estimating probabilities in order that the effects of these might be mitigated. Later in this section we summarise some of the psychological and statistical literature examining such behavioural tendencies.

Problem Description

We now describe the particular type of problem for which prior information is elicited in this paper. Define the rock in two layers intersected by a vertical well to be of types T_1 and T_2, respectively (Fig. 1(a)). We consider the most general case where the range of rock types for T_1 is continuous rather than discrete (e.g., T_1=%carbonate minerals), and similarly for T_2.

Let us describe the Earth model by a vector **m** containing the unknown model parameters: **m**=[T_1, T_2]. Vector **m** is clearly 2D, and a value for **m** corresponds to a single point on the 2D plane in Figure 1(a). If we now consider a well that intersects three unknown rock types then **m**=[T_1, T_2, T_3] corresponding to a single point in the 3D space shown in Figure 1(b). By analogy, if four rock types were intersected then **m**=[T_1, T_2, T_3, T_4] corresponds to a single point in 4D hyperspace (hyperspace is simply space in more than 3D – Figure 1(c)). Generally, if we have an unknown model that is described by *N* parameters (*N* rock types in this example) then **m**=[$T_1, T_2, \ldots, T_N$] corresponding to a point in *N*-dimensional hyperspace. We call the plane, space, or hyperspace the *model space M*.

Each time the number of parameters (layers) is increased by one, the cost of exploring the model space *M* increases. Consider the number of points required to sample the model space to some level of detail. Assuming that 10 values of each model parameter would provide a sufficiently detailed exploration then for the 2D problem in Figure 1 we require 10^2 = 100 samples, for the 3D problem we require 10^3 = 1000 samples, etc. If a well in the field intersected even only 20 rock types, the 20-dimensional hyperspace spanned by the unknown model parameters is surely incomprehensibly large to all humans and exploration to the required level of detail would require 10^{20} samples.

Estimating the true Earth model (the true rock types defined by a single point in hyperspace) requires that we have sufficient information to be sure that all other points represent untrue models. As the number of other points increases, so the amount of information required increases. The property of increasing hyperspace volume and the

corresponding increase in information required to estimate the true model as its dimensionality increases can be defined formally and is referred to as the 'curse of dimensionality' (Scales 1996). Curtis & Lomax (2001) show that in such large hyperspaces intuitive or innate notions of probabilities derived from 1D-3D space no longer hold. In turn this implies that simply asking a subject to estimate probability distributions in such spaces will surely lead to poor results.

Consider estimating rock types and architectures in a 2D cross section through two vertical wells (Figure 2(a)) – a similar problem is addressed in Wood & Curtis (this volume). Given no information at all about the rocks intersected by the wells, any point in space on the cross section could contain any rock type. One method of parameterising the model space is to discretise the cross section as shown in Figure 2(b). Each discrete cell might contain only one model parameter, 'rock type', that we wish to estimate. This results in an N-dimensional model vector $\mathbf{m}$, and even if only 10 cells are used both horizontally and vertically, then N=100 and the resulting 100-dimensional model space (Fig. 2(c)) is unimaginably large. If the model was 3D and only ten cells were used in the third dimension then the model space would be 1000-dimensional.

When we include all parameters in which we might be interested in a reservoir situation, it is always the case that even after constraints have been imposed from the most comprehensive geophysical data, there will remain lines, surfaces or hypersurfaces in M on which any model is equally likely. These lines or surfaces span what is called the *null space* in M – the part of M about which we have no information that differentiates between models that are more or less likely to occur in the Earth. Thus, the null space represents fundamental uncertainty in estimated Earth models. No amount of geophysical data can remove all null space directions, so other information is always necessary to constrain our estimates. Since such information is independent of the geophysical survey data, it is called *prior* information.

The problem that we address in this paper is that of collecting and usefully representing reliable geological prior information that reduces the portion of model space M that represents feasible Earth models (Fig. 2(d)), even before any (current) geophysical data is collected. This has two principal, tangible benefits: first, this information will reduce the null space post-data acquisition, thereby improving Earth model estimates and

reducing uncertainty. Second, as this information is available prior to acquiring data it can be used to design an optimal data acquisition survey. Data need only be collected if it fills gaps or reduces large uncertainties in the prior information; the optimal dataset to collect will be that which fulfils as many of these requirements as possible, and at either minimum or acceptable cost.

Elicitation Theory

The problem tackled in the field of Elicitation Theory is to design the best way to interrogate experts or lay-people (henceforth, *subjects*) in order to obtain accurate information about a topic in question. Extensive psychological research has shown that this is a difficult problem: simply asking even expert subjects to provide a (numerical) probability estimate results in poor probability assessments. The reason is that people find providing such estimates difficult and hence tend to use heuristics to help themselves; these in turn introduce biases and poor calibration (Kahneman *et al.* 1982). Baddeley *et al.* (this volume) review various biases that are observed commonly in humans. Below we briefly summarise some of these and highlight various key works from Elicitation Theory.

Individuals' Biases and Heuristics

At least two main types of bias can be distinguished (Skinner 1999): *motivational bias* (caused by the subject having some vested interest in biasing the results) and *cognitive bias* (caused by incorrect processing of information available to the subject). The former can often be removed by explaining that unbiased information is critical, and by using all possible means to remove the particular motivations for biasing the results. The latter are typically the result of using heuristics, and suitable elicitation methods can sometimes correct part of such biases.

At least four types of heuristics causing cognitive biases are commonly encountered: availability, anchoring and adjustment, representativeness, and control (Kahneman *et al.*

1982). *Availability* is the heuristic of assessing the probability of an event by the ease with which occurrences of the event are brought to mind and is biased by prominence rather than frequency of different information (e.g., bicycle accidents are more frequent, incur a greater number of fatalities overall, but are less easy to recall than aeroplane crashes). *Anchoring and adjustment* is a single heuristic that involves making an initial estimate of a probability called an anchor, and then revising it up or down in the light of new information (e.g., information about opinions of others on the matter in question). This typically results in assessments that are biased towards the initial anchor value. The *representativeness* heuristic is where people use the similarity between two events to estimate the probability of one from the other (e.g., we may use knowledge of an individual's personality traits to estimate their political persuasions). The *control* heuristic is the tendency of people to act as though they can influence a situation over which they have no control (e.g., buying lottery tickets with personally chosen rather than random numbers).

Well-known consequences of these heuristics are the gamblers fallacy, the conjunction fallacy, base-rate neglect, probability matching and overconfidence (see Baddeley *et al.* (this volume) for definitions). Of all of these biases, the most prevalent may be overconfidence and base-rate neglect (Baecher 1988). Overconfidence is particularly a problem for extreme probabilities (close to 0% and 100%) which people find hard to assess.

The biases described above imply that the results of elicitation from each individual expert may need to be calibrated. This in turn requires some statistical model of the elicitation process. The most commonly used and referred to model in the literature is that of Lindley *et al.* (1979). This model requires that there be an objective assessor who will consolidate the results derived from subjective experts. We will not make explicit use of this model, however, as it is not clear why an assessor should be any more objective than the expert.

Other work that attempts to recalibrate experts' judgement includes that of Lau & Leong (1999) who created a user-friendly JAVA interface for elicitation that includes graphical illustrations of possible biases and any inconsistencies in elicited probability

estimates (see the review of graphical methods by Renooij 2001). The interface then enters into dialogue with the expert until consistency is achieved.

It should be noted that some of the heuristics described above can perform extraordinarily well in some situations (e.g., Gigerenzer & Goldstein 1996; Juslin & Persson 2002). In practical situations, however, it is not clear from the results alone whether the heuristics work well or not since there is no objective answer with which to compare them. Hence, one role of the elicitor is to try to reduce the use of heuristics as much as possible.

Elicitation protocols and strategies

There are no universally accepted protocols for probability elicitation and there is relatively little formal empirical evaluation of alternative approaches. There are, however, three common assessment protocols (Morgan & Henrion 1990). These generally include 5 phases: motivating the experts with the aims of the elicitation process, structuring the uncertain quantities in an unambiguous way, conditioning the expert's judgement to avoid cognitive biases, encoding the probability distributions and verifying the consistency of the elicited distributions.

Within each protocol the elicitor must decide exactly what problems to tackle or what questions to ask in order to maximise information about the topic of interest. Coupé & van der Gaag (1997) showed how a sensitivity analysis might sometimes be carried out in order to see which elicited probabilities would have most influence on the output of a Bayesian belief network. Using their terminology, later in this paper we extend their "1-way sensitivity analysis" by introducing a new method of analysing the sensitivity test results that allows optimal decisions to be made regarding which probabilities should be elicited.

We have discussed only a few of the many references in the field of elicitation theory. It is clear that eliciting prior information is not a trivial problem. Below we make the first attempt (to our knowledge) to optimise the elicitation process in real time using all information available. This must be an optimal strategy in principle. However, in practise the details of our particular method may be improved in future: for example, we make no

attempt to correct biases in the subject's judgement, but optimise the elicitation process within the framework of the biases described above. It is likely, therefore, that our method can be improved by combining it with bias-reducing techniques such as those of Lindley *et al.* (1971) or Lan & Leong (1999).

Methodology

We now present a method to derive, parameterise, and describe geological prior information. We distinguish two main types of geological prior information: *static* information is information about reservoir architectures and their properties as they exist today, and *dynamic* information about the various processes that created the present characteristics of the rock (Wood and Curtis, 2004). Such a classification implicitly incorporates any information about past geological events as these can only be estimated using a combination of static and dynamic information. The method proposed includes both types of information.

Geological information is introduced in two steps, each described in a separate subsection below. First, geological modelling software that encapsulates dynamic prior information is used to remove large portions of model space M that are geologically infeasible. Second, within the remaining portions of M, a new method uses an interpreters' or geologists' (henceforth called experts', or subjects' [of an experiment]) experiences, from previously observed static information, to assess the relative likelihood of occurrence of different models, and to convert the experts' knowledge into prior probabilities. In order to estimate these probabilities we set up an elicitation procedure that is described in the third subsection below. In the following section we describe how this elicitation procedure was optimised. The result of these steps is a normalised probability distribution representing geological prior information.

Geological modelling software

Software exists that models the architecture of existing rock structures either based on statistics from existing, observed outcrops (static information e.g., Wood & Curtis this volume), or based on geological processes of deposition, erosion, transport, resedimentation and subsequent diagenesis acting over geological time scales (dynamic information e.g., Tetzlaff & Priddy 2001; Griffiths *et al.* 2001; Tetzlaff (this volume); Burgess (this volume)). In this study we will use an example of the latter type of package, the Geological Process Model (GPM: Tetzlaff & Priddy 2001), but any other preferred software may be used to replace this without any other change in our methodology.

The GPM software has been developed by Dan Tetzleff of Schlumberger, based on initial research conducted at Stanford (Tetzlaff & Harabaugh 1989; Griffiths *et al.* 2001; Tetzlaff & Priddy 2001) but with many recent additions and enhancements. GPM currently models at least (a) depositional sedimentation, (b) compaction and diagenesis, and (c) erosion and subsequent re-deposition within siliciclastic settings.

> **Assumption 1**: We make the assumption that whatever software is used embodies at least sufficient geological processes to model the range of possible geologies that are likely to be encountered within the true geology in question – at least to the level of detail relevant to the problem of interest.

The validity of results obtained will be conditional on this assumption. The software may, additionally, be able to model some geologies that are infeasible in the Earth (due either to infeasible parameter values input to the software, or due to errors in the software), or model some geologies that are feasible in the Earth but are impossible in the true geological domain of interest, without damaging the principles of our method.

Since the components of GPM used naturally model only deposition and some diagenetic processes, our use of this package combined with Assumption 1 implicitly implies the following second assumption:

> **Assumption 2**: Either (i) the geology has not been significantly modified by unmodelled tectonic or diagenetic processes, or (ii) we are only interested in the non-tectonic components of the architecture and rock types, and will compensate for the tectonic modifications in a separate modelling or data decomposition step.

Any deviation from these assumptions will introduce errors into the prior probability distributions derived, and may require a different package to be used.

GPM requires various input parameters in order to create 3-D geological scenarios. Denote these parameters by **q**, where **q** is *K*-dimensional if there are *K* parameters, and let their range of values define parameter space Q. Then GPM can be regarded in purely functional terms as a black box that translates parameters **q** into Earth models **m** (Fig. 3). This black box embodies much dynamic prior information about depositional and diagenetic processes. Hence, if we allow parameters **q** to vary over their entire feasible ranges, the range of models produced by GPM represents a range of possible 3-D geological models. By Assumption 1, this range spans the range of possible real geologies, at least to the desired level of detail.

Although there may be huge variation in the models produced, this range is still very much smaller than the *N*-dimensional model space produced by simply discretising the reservoir into *N* cells (as described earlier, $N=1000$ in a relatively coarse model) with no additional information about the distribution of rocks within the cells (compare Figures 2(c) and (d)). In the latter context, both credible and incredible models are allowed, e.g., a model where every consecutive cell has alternately siliciclastic and carbonate rock types in a checkerboard pattern is feasible. In contrast, GPM would not be able to reproduce such a model; the model is therefore deemed impossible and removed from further consideration. Thus, if we consider only models that can be produced by GPM we have automatically greatly reduced the possible model space simply by the introduction of dynamic geological prior information. Such information is difficult to represent without using either GPM or another modelling package explicitly.

Defining prior probabilities

GPM reduces our range of possibilities from an *N*-dimensional hypervolume to some manifold that is at most *K*-dimensional, where *K* is usually much lower than *N* (Fig. 3). In this context a manifold can simply be thought of as a lower-dimensional structure embedded within a higher dimensional space (Fig. 2(d)). An example of a manifold is the

familiar Swiss Roll structure; intrinsically this is a 2D plane (K=2) that is 'rolled up' within a 3D space (N=3).

By Assumption 1, the manifold defined by GPM spans at least the range of all possible geological models. For reasons described earlier, some additional models spanned by the manifold may be impossible. Even within the range of models that are possible, some will be more or less likely to occur in practise. We would therefore like to define over this manifold a probability distribution that describes the likelihood of occurrence of each of the models spanned, in order to further augment the prior information injected into subsequent investigations.

In principle there are at least two methods that can be used to create this distribution. First, we could define the prior probability distribution across parameters **q** directly. Since GPM merely translates **q** into the model space M, we can also use GPM to translate this distribution into the model space. The resulting probability distribution in model space M would represent both prior information embodied within GPM and information from the prior distribution of parameters **q**. In practise this method is problematic: parameters **q** include sedimentation rates, loci of sediment origination, pre-existing topography, compaction heterogeneity, etc. Such information can only be derived from observations of existing geological architectures, properties and observable processes that occur on Earth today. Parameter values are then inferred by geologists who in effect carry out an 'inverse-GPM' in their heads – they unravel the complex dynamic sedimentary and diagenetic processes, usually non-quantitatively, to infer rough ranges on some parameters in **q**. For other parameters in **q** their relationship to observations in the present day will be so complex that no geologist can infer sensible estimates for their ranges, other than by intuition. In turn, it is not clear from where such intuition would have derived since no geologist can have observed true processes that take place over geological time scales, and hence geologists cannot intuitively estimate their effects quantitatively. The only alternative would be to observe present day values for parameters **q** and describe these by a probability distribution, but these values may well not be pertinent over geological time scales.

Here, we opt for a second approach: we directly interrogate experts about the likelihood of occurrence of models produced by GPM. Our approach is to take sample

models from the manifold and interrogate experts about their likelihood of occurrence in the reservoir or geology of interest; we can then interpolate between these models along the manifold, to define the probability of occurrence over the entire section of the manifold bounded by the samples. Interrogating people for information requires an elicitation process; we now describe a new elicitation method suitable for this purpose.

Method to elicit geological information

In light of the brief summary of heuristics and biases given above and described in Baddeley *et al.* (this volume), we need a strategy that does not ask for absolute probabilities, that allows the reliability of probabilistic estimates to be assessed, and that also mediates the effects of individuals' biases and heuristics by sensibly consolidating the input of multiple experts. We therefore include the following elements in our elicitation method:

First, we ask only that experts provide *relative* probabilities of occurrence of sets of a small number of models at a time. This has the advantage that the subject need only have knowledge about the *relative* abundance of Earth structures that are similar to the models in question, rather than knowledge of *all* such structures existing in the Earth. The small number also ensures that they can keep all examples in memory at once, rather than allowing one or more to be displaced by the availability heuristic.

Second, we ensure that 'loops' exist in the models presented within the sets. Loops can be illustrated if we consider only three models, A, B and C. Denote the relative probability of A and B by

$$\Pr(A,B) = \frac{P(A)}{P(B)} \qquad \ldots(1)$$

where $P(A)$ denotes the probability of model A being true. If we ask a subject to estimate Pr(A,B) and Pr(B,C) then we can estimate

$$\Pr(A,C) = \Pr(A,B) \times \Pr(B,C). \qquad \ldots(2)$$

If we also ask the subject to estimate Pr(A,C) then the match between this estimate and the calculated value provides a consistency or reliability check on their probability

estimates. Obviously loops can involve many more models than three. The requirements that probabilities around each loop are consistent, and that other required relationships between valid probabilities are satisfied, are referred to as *coherency conditions* (Lindley *et al.* 1979).

In practise we ensure that loops exist by presenting each subject with a set of L models ($L=5$ in the current study) and asking them first to rank them in order, and second to estimate their relative likelihoods of occurrence. They do this by first assigning the most probable model in their view an (arbitrary) likelihood value of 10, then by assigning the other models (possibly non-integral) likelihoods relative to that model. By simply dividing the different likelihood estimates thus obtained in all possible ways we obtain estimates of the required relative probabilities around all possible loops that can be created with L models, similarly to equations (1) and (2). All probabilities around loops within each set of L models are automatically consistent, but loops between different sets remain as coherency conditions.

Third, we ensure that each set incorporates models that are also a member of other sets, and that at least one loop can be constructed between a set of models that span all other sets, such that relative probabilities around that loop have been constrained. Without this condition it is possible that there exist pairs of models that occur in different sets, and about the relative probabilities of which we obtain no information.

Fourth, in principle we would like absolute probabilities of occurrence of models, not relative probabilities. If the manifold is bounded, then this problem can be solved by introducing the fundamental property of all valid probability distributions, that,

$$\int_{Man} P(\mathbf{m})\, d\mathbf{m} \quad = \quad 1 \qquad \ldots(3)$$

where the domain of integration is the entire manifold. This condition can be used to provide a constant value that normalises the integral on the left of equation (3) to unity and thus provides absolute probabilities.

If the manifold is unbounded then it is only possible to estimate absolute probabilities given some assumptions about the behaviour of the probability distribution approximately outside of the minimum convex hull (on the manifold) containing all assessed models. If the models assessed do not contain all possible degrees of freedom (e.g., if not all parameters in **p** were varied so as to span their true possible ranges) then

the normalisation procedure above will provide only normalised probability distributions that are conditional on those parameters lying within the ranges across which they were varied.

The elicitation process is carried out by email with subjects in remote locations from the person running the elicitation process, henceforth called the *elicitor*. Whenever possible, and certainly on the first two occasions that a particular subject was involved in trials, the subject is in contact by phone with the elicitor during the entire trial. The elicitor ensures that the subject understands what is being asked of him/her, and encourages the subject to talk through exactly what s/he was doing during every stage of the trial.

Our method consists of repeatedly presenting subjects with sets of $L=5$ models selected to fulfil the above criteria and asking each subject to estimate the relative likelihoods of models as described in the second element above. The subjects are presented each set of models as a zipped attachment to a standard email that explains how to assess relative likelihoods. The five models are given in the output file format of GPM; each subject requires a copy of the GPM viewing program, GS, to view each model. Subjects are asked to evaluate the likelihood of occurrence of the final architecture and sedimentary grain size distributions produced in the model. However, they can also use GS to run backwards and forwards through geological time to see how each model was created. Trials continue to be presented to subjects until either sufficient information is obtained, or an equilibrium level of residual uncertainty is reached.

The elicitor takes notes of the subject's description of what s/he is doing from the opening of the standard email through to the final evaluation of all models. The notes focus on how the subject views the models (what angles of view are used, and in what modes of view – both dynamic and map view, either with and without any water layer are possible) and how s/he assesses the models' likelihoods (on what their assessment is based, how and whether they incorporate dynamic information about how each final model was created in their assessment), and any other details mentioned by the subjects.

The four strategic elements described above do not uniquely define the sets of models to be presented to subjects in each trial. Thus, we are left with an experimental design problem: given a particular modelling package (here, GPM), and assuming that we

vary parameters **q** over ranges that include the correct values, exactly which models should we include within each set presented to experts, and how many experts should we use, in order to obtain maximum information about the relative probability of models at reasonable cost? This is a complex question due to the large number of often intricate loops that might be created between sets, and to the unknown degree of variability that might occur between different subjects. In the next section we show how to design such experiments optimally in real time using statistical experimental design theory. To our knowledge this has not been done previously.

Experimental Design

We define the experimental design scenario as follows: let $\mathbf{p} = [P_1, \ldots, P_M]^T$ be the vector of probabilities that we would like to estimate by elicitation. Usually $P_i = P(\mathbf{m}_i)$, (i=1,…,B), where $P(\mathbf{m}_i)$ denotes the probability of model $\mathbf{m}_i$ occurring in reality, and $\mathbf{m}_i$ is the i'th of a total of B models, sampled from the manifold, over which we wish to estimate a prior probability distribution.

Constraints offered by trials

To elicit relative probabilities during each trial we present a subject with a subset of L of the model samples, $\{\mathbf{m}_{k_1}, \ldots, \mathbf{m}_{k_L}\}$, and request estimates of the relative likelihoods of these models as described in the previous section. Denote the elicited estimates thus obtained by $\hat{P}_{K_iK_j}$ where $P_{K_iK_j}$ is defined by,

$$P_{K_iK_j} = \frac{P_{K_j}}{P_{K_i}}, \qquad \ldots(4)$$

where $\hat{P}_{K_iK_j}$ is an estimate of $P_{K_iK_j}$, and where $i = 1,..., L-1$ and $j = i+1,..., L$ form a nested loop. Re-arranging we obtain linear constraints of the form,

$$\hat{P}_{K_iK_j} P_{K_i} = P_{K_j} \quad \Leftrightarrow \quad \hat{P}_{K_iK_j} P_{K_i} - P_{K_j} = 0. \qquad ...(5)$$

Let matrix $\mathbf{A}_1$ and vector $\mathbf{d}_1$ be defined such that all equations of the form (5) that are *expected to be derived from the next* trial may be expressed as,

$$\mathbf{A}_1 \mathbf{p} = \mathbf{d}_1 . \qquad \ldots(6a)$$

If we also have estimates of the uncertainties that might be expected of the vector $\mathbf{d}_1$, expressed as a covariance matrix $\mathbf{C}_1$, then the equations describing constraints available after the next trial become,

$$\mathbf{A}_1^T \mathbf{C}_1^{-1} \mathbf{A}_1 \mathbf{p} = \mathbf{A}_1^T \mathbf{C}_1^{-1} \mathbf{d}_1 \qquad \ldots(6b)$$

In practise we estimate a diagonal $\mathbf{C}_1$ based on results of previous trials (see below). Note that equation (6a) is simply a particular case of the more general equation (6b) with $\mathbf{C}_1 = \mathbf{I}$ where $\mathbf{I}$ is the appropriate identity matrix, and where equation (6a) has been multiplied by $\mathbf{A}_1^T$. We therefore consider equation (6b) only.

We assume that several estimates of relative probabilities may have been elicited in previous trials, providing a set of equations similar to,

$$\hat{P}_{ij} = \frac{P_j}{P_i} \qquad \Leftrightarrow \qquad \hat{P}_{ij} P_i - P_j = 0 \qquad \ldots(7)$$

where *i* and *j* run over all indices for which estimates of $\hat{P}_{ij}$ exist. Let matrix $\mathbf{A}_2$ and vector $\mathbf{d}_2$ be defined such that all equations derived from previous trials may be expressed in the linear form of equation (7) as,

$$\mathbf{A}_2 \mathbf{p} = \mathbf{d}_2 . \qquad \ldots(8a)$$

If uncertainties in $\mathbf{d}_2$ are known and expressed as a covariance matrix $\mathbf{C}_2$ then the corresponding equations derived from previous trials become,

$$\mathbf{A}_2^T \mathbf{C}_2^{-1} \mathbf{A}_2 \mathbf{p} = \mathbf{A}_2^T \mathbf{C}_2^{-1} \mathbf{d}_2 \qquad \ldots(8b)$$

with solution,

$$\mathbf{p} = \left[\mathbf{A}^T \mathbf{C}_2^{-1} \mathbf{A}\right]^{-1} \mathbf{A}^T \mathbf{C}_2^{-1} \mathbf{d}_2 \qquad \ldots(9a)$$

and post inversion covariance matrix,

$$\mathbf{C}_{post} = \left[\mathbf{A}^T \mathbf{C}_2^{-1} \mathbf{A}\right]^{-1} \qquad \ldots(9b)$$

assuming that sufficient constraints exist from previous experiments that equations (8a) or (8b) are strictly overdetermined and hence that matrix inverses in equations (9a) or

(9b) exist. In practise we estimate a diagonal matrix for $\mathbf{C}_2$ by assuming that relative likelihood estimates provided are uncertain with a standard deviation of 50% of each estimate. This would seem to be a conservative estimate.

Note that equation (8a) is simply a particular case of the more general equation (8b) under the substitution $\mathbf{C}_2 = \mathbf{I}$ where **I** is the appropriate identity matrix, and where equation (8a) has been multiplied by $\mathbf{A}_2^T$, and the solution to equation (8a) is given by making the same substitution in equations (9a) and (9b). We therefore only consider equations (8b), (9a) and (9b).

If it is the case that equation (8b) is underdetermined then the system must be regularised to remove non-uniqueness in the solution (9a). This is often achieved either by damping out zero or near-zero eigenvalues, or by smoothing the solution according to some chosen kernel (Menke 1989). However, both of these strategies effectively add additional prior information to the system – information that we do not have. Rather than pursue this line of discussion further, we note that if a single, non-normalised probability value is fixed at an arbitrary value, at least in principle it is relatively easy to carry out sufficient trials that systems (8a) or (8b) become overdetermined. This is achieved by fixing an arbitrary (non-zero) probability value to 1, by including every individual probability in **p** within at least one trial, and by ensuring that at least one particular individual probability in **p** is included in every trial conducted. We therefore assume that previous trials have been carried out sensibly and that there is no need to give further consideration to underdetermined systems.

Design Problem Definition

The experimental design problem to be solved in order to design each trial given all previous information can now be defined precisely:
Let,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \quad , \quad \mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} \quad , \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 \end{bmatrix} \qquad \ldots(10)$$

The design problem is to select the subset of L models $\{\mathbf{m}_{k_1},\ldots,\mathbf{m}_{k_L}\}$ that results in a matrix $\mathbf{A}_1$ and vector $\mathbf{d}_1$ such that the set of equation,

$$\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}\ \mathbf{p} = \mathbf{A}^T\mathbf{C}^{-1}\mathbf{d} \qquad \ldots(11)$$

is expected to place maximum constraint on (result in minimum residual uncertainty in) the probability vector of probabilities **p** of interest.

Optimal Design Algorithm

Since equation (11) is linear, this problem can be solved using linear experimental design theory. Curtis *et al.* (2004) present a deterministic algorithm that approximately solves this kind of problem. The only limitation is that the covariance matrices must be diagonal (that is, measured or estimated correlations between probability estimates are neglected). This limitation is not severe since in practical situations it is unlikely that sufficient trials will be conducted that correlations between individual probability assessments can be measured robustly, and estimates of expected correlations should in principle be based on such measurements. Hence, although inter-probability correlations may exist it is unlikely that accurate descriptions of these will be accessible, so diagonal covariance matrices should suffice for practical purposes.

The algorithm of Curtis *et al.* (2004) initially includes *all M* models within the set that would (hypothetically) be presented to the subject during the next trial (i.e., set $L=M$ above). This results in matrices $\mathbf{A}_1$ and $\mathbf{C}_1$ and vector $\mathbf{d}_1$ defined such that equation (6b) contains *all possible* equations of the form (5). The design problem is then solved by deleting M-L models from this set, so that the remaining L models form a subset that ensures that equation (11) results in low residual uncertainty in **p**.

In essence, residual uncertainty in **p** occurs because matrix $\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}$ has small eigenvalues: uncertainties in the vector **d** propagate into vector **p** with a magnification in proportion to the inverse square root of these eigenvalues (e.g., Menke 1989). Hence, uncertainty in **p** will be minimised when matrix $\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}$ has eigenvalues with the

largest possible magnitudes. Small eigenvalues generally occur because one or more rows of **A** are nearly linearly dependent on the other rows. Since $\mathbf{C}^{-1}$ is diagonal it merely weights rows of **A** relative to each other. Complete linear dependence results in a zero eigenvalue and singularity of the matrix $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}$.

Each equation of the form (5) results in a single row of matrix $\mathbf{A}_1$ in equation (6b). Rows of matrix $\mathbf{A}_1$ that are linearly dependent on other rows of $\mathbf{A}_1$ correspond to redundant constraints resulting from the current trial. Rows of $\mathbf{A}_1$ that are linearly dependent on rows of $\mathbf{A}_2$ correspond to constraints that effectively repeat those offered from previous trials.

The algorithm developed by Curtis *et al.* (2004) removes small eigenvalues from the spectrum by deleting those equations that correspond to rows of $\mathbf{A}_1$ that are most linearly dependent on other rows of matrix **A**. It does this by removing one model at a time from the set that will be presented to the subject. Each removal of a model, $\mathbf{m}_q$ say, results in several rows being deleted from **A**, namely all rows corresponding to equations concerning relative probability estimates involving P_q. The algorithm iterates, and at each iteration that model is deleted for which the corresponding rows are most linearly dependent on other rows in **A**. The measure of linear dependence employed by Curtis *et al.* (2004) accounts also for the relative weighting imposed by $\mathbf{C}^{-1}$. The algorithm stops iterating when the number of models to be presented to the subject reaches the required number, always 5 in our experiments.

While the remaining set of equations result in a matrix **A** that has rows that may not involve the least linear dependency possible, experiments in various domains of application have shown that the algorithm gives designs which significantly increase the smallest eigenvalues in the spectrum relative to those that would pertain if a random experiment was performed (Curtis et al. 2004). Alternative methods include the methods of Curtis (1999a,b) that tend to increase the largest eigenvalues of the spectrum, and those of Rabinowitz & Steinberg (1990), Steinberg *et al.* (1995), Maurer & Boerner (1998), and Maurer *et al.* (2000) that, roughly, focus on increasing all eigenvalues equally. The design problem above is over-determined so all small eigenvalues will usually be included, unregularised, within the solution to equation (11). Since small

eigenvalues are principally responsible for large residual uncertainties in the probabilities estimated, the method of Curtis *et al.* (2004) is well suited to the design problem at hand.

An Elicitation Experiment

We carried out a simple elicitation experiment in order to demonstrate the methodology presented above. In this experiment only a single subject was used, and the manifold over which we estimated the prior probability distribution was 1-dimensional. Our aim in the current experiment was to demonstrate the elicitation technique, and to show that the optimal design method presented above increases information obtained during the elicitation procedure. Our results are not sufficiently prolific to estimate statistically the percentage improvement that one might expect by using the design algorithm rather than conducting random trials. Instead our results simply demonstrate that the algorithm does provide improved estimates of elicited probabilities, and hence encourage more extensive tests of the elicitation methodology in future.

During the experiment we assessed the probability distribution representing the prior likelihood of occurrence of different fluvial deltaic geological models produced by GPM by varying only a single parameter, the 'diffusion coefficient' (explained below). The resulting elicited probability distributions will be conditional both on all of the other parameters required to run GPM being fixed at those values used in this study, and on the diffusion coefficient best representing true sedimentary erosion and re-deposition in the fluvial deltaic systems modelled being within the range of diffusion coefficients spanned by this experiment (the range [0.7 - 30] m^2/yr was used).

When modelling sedimentary erosion and re-deposition, the cumulative effects of various erosional processes that transport sediment approximately (locally) down-dip is modelled by applying a diffusion equation to simulate redistribution of existing sediment (e.g., Tetzlaff & Harbaugh 1989). A simple example would be:

$$\frac{dh}{dt} = D \nabla^2 h$$

where h is the sediment top surface height, D is the diffusion coefficient, and $\nabla^2 h = \left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} \right)$ where *x* and *y* are orthogonal horizontal coordinates is the Laplace operator or second spatial derivative of h. In this equation, the rate of diffusion is moderated both by the local curvature in top sediment topography and by the diffusion coefficient.

Increasing the diffusion coefficient makes sediment redistribution more efficient. For a given set of initial conditions and a fixed period of geological time over which the diffusion process acts, the effects of increasing the diffusion coefficient are to transport more sediment. Since lower diffusion coefficients acting over a longer time period could produce a somewhat similar effect, increasing the diffusion coefficient creates final models where the geological landscapes appear to be more 'mature' to a geologist (compare the left and right images in Fig. 4). Table 1 shows the diffusion coefficient values used in each model $\mathbf{m}_i$ used in this study and for which estimates of $P_i = P(\mathbf{m}_i)$ are required (see below for details).

Experimental description

In this study we elicit information about fluvial-deltaic environments. The single subject was geologically expert, but was not a specialist in this depositional setting. Hence, the knowledge elicited in this paper is the general knowledge of a good geologist about a fairly standard sedimentary environment.

Table 1 shows indices by which each model simulated is referred to in the text, the parameter values used in the GPM simulation for each model, and those models (and their ordering) that were included within each of nine trials conducted on the subject. Trials 1-5 were designed randomly (i.e., models to be included in each of these trials were selected at random, without using the experimental design procedure); trials 6-8 were conducted by running the design procedure before each trial, including all knowledge gained from previous trials within matrix $\mathbf{A}_2$ in equations (6) and (7). Trial 9 was again designed randomly and was used as a control trial as shown below, since the

set of models used in this trial did not appear in any of the first five trials, or in the three designed trials.

Several points should be noted from Table 1. First, we are only interested in estimating prior probabilities conditional on all parameters other than the diffusion coefficient being fixed. Table 1 also shows that in some models, other parameters were varied (models 10, 14 and 15). Models 10, 11, 12 and 13 were additionally performed with boundary conditions that were all closed to sediment and fluid flow; in all other simulations boundaries were open towards the basin, open sideways below sea level, and closed sideways above sea level and landward (see Fig. 4). Hence, all models below the solid line (10-15) do not constitute samples of our conditional distribution, and can be regarded as 'nuisance' models. These simulations were made in order to test the effects of outlying models on elicited probability estimates within the conditional range; the results are not central to this paper and will be analysed elsewhere. Here we only consider relative likelihoods from each trial that pertained to models 1-9. As a consequence, trial 3 is immediately rendered useless for this particular study since it includes only a single model from within the conditional range and hence results in no relative likelihood estimates.

The time simulated in each run was 1000 years. Sediment input had four different grain sizes, 1, 0.1, 0.01 and 0.001 mm, in equal proportions. Each GPM run was produced by a single sediment source point at the centre of the landward boundary of the model, with identical initial topography and in models 1-9 with flow boundary conditions open below sea level and closed above sea level. Sea level was constant throughout each run.

Results

Table 2 shows likelihood estimates of probabilities sampled from the conditional distribution described earlier, from trials involving our single subject. In each trial the subject was asked to order the models in terms of their likelihood of occurrence, and assign a likelihood value of 10 to the most likely model. Other models were then to be

assigned (not necessarily integral) likelihoods relative to this maximum value. Hence, only relative likelihoods (ratios of the values in Table 2) should be interpreted because the absolute values are arbitrarily scaled.

Many inconsistencies are evident in the data. For example, the relative likelihood of models 1 and 2 was estimated three times, in trials 4, 7 and 8, with estimates of $P_{12} = P_2 / P_1$ assuming values 2, 5 and 1.17 respectively. Similarly estimates of the relative likelihood $P_{92} = P_2 / P_9$ assume values 1.29, 3.33, and 3 in trials 6, 7 and 9 respectively. However, some consistencies are also evident in the data. Whenever model 3 is included within a trial it is assigned the maximum likelihood value of 10. When model 3 is not included, the maximum likelihood value is always assigned to models 2 or 4 (when either or both are included). Hence, the conditional distribution appears to assume a maximum somewhere around the diffusion coefficient 1.3 m^2/yr, and this maximum almost certainly occurs between coefficients 1 and 2 (a maximum at D=1 is confirmed in Figure 6 which shows the final estimate of the conditional distribution). Additionally, notice that likelihoods lower than 5 are all confined to models 1, 7, 8 or 9, i.e., towards the two extremes of the range of diffusion coefficient values considered. Therefore, assuming that the trend of diminishing likelihoods continues outside of the range of coefficients included in these trials, we appear to span a range of coefficient values over which the probability distribution reduces to at most one half of its maximum value.

Inconsistencies such as those observed above are not unexpected. The subject in this case did not have a strong quantitative background, and had only a general knowledge of probability theory. General knowledge usually includes only an inkling of the concept of consistency of relative probabilities, so the subject would not have been checking to ensure such consistency. Additionally, it has been shown in numerous previous studies that subjects have most difficulty in assessing extremes in probability – values closest to zero or to one. Probability ratios between high and low probabilities such as P_{12} and P_{92} described above are therefore expected to exhibit maximum inter-trial variation. It is likely, therefore, that the current method could be enhanced by illustrating such

inconsistencies graphically during the elicitation procedure in order that the subject could reconcile their relative probability estimates (e.g., Lau & Leong 1999; Renooij 2001).

Trials 6-8 were designed using the algorithm presented above, using information from previous trials to construct the design of each subsequent trial. Trial 9 on the other hand was designed randomly. We used this as a single control trial: after each optimally designed trial we compare results obtained to those results that would have been obtained if the control trial had been conducted instead, assuming that the same results would have been obtained during the control trial whenever it was carried out. While the percentage improvement (if any) in results observed in this test will certainly not be statistically significant, we simply test whether an improvement is consistent observed.

We first compare the standard deviation (s.d.) of the probability estimates, calculated using equation (9b), using optimal and control trials for each designed trial. The difference in the mean s.d. (control minus optimal) expressed as a percentage of the mean s.d. estimate obtained in each trial was 3.8%, 10% and 2.6% in trials 6, 7 and 8 respectively where a positive percentage represents an improvement. While these improvements in the post-trial s.d.'s are small, they are consistently positive. Also, the total improvement due to the optimal strategy is compounded in each successive trial and hence is 15.7%. The actual percentage values are statistically unconstrained since they depend both on the (random) control trial used, and on the estimates obtained from the subject during the control trial. Hence, positivity of the percentages and hence consistent improvement is probably the limit of interpretation possible given only these results.

In order to search for any improvement in the mean probability estimates, we assume that the probability estimates obtained using all information from trials 1-9 in equation (9a) provide the best estimates available, and these are shown in Figure 6. We calculate the mean absolute difference between these best probabilities and those estimated after each of trials 6-8. We compare these with the differences that are obtained if each of trials 6-8 is, respectively, replaced with the control trial 9. The results expressed as percentage of the probabilities in Figure 6 are 15%, 6% and 2% after trials 6, 7 and 8 respectively. Again, the improvements are all positive indicating improvement in all cases. The decrease in magnitude of the improvement over successive trials is expected: as trials progress we approach an equilibrium level of uncertainty – that which is inherent

within our subject's general knowledge. As we approach that equilibrium, additional trials should on average show diminishing improvement in the mean estimate.

Discussion

The results presented above confirm that the optimal elicitation strategy described earlier improves the information content expected in conducted trials beyond that expected in random trials, at least in the simple experiment that we performed and describe herein. Clearly more detailed and extensive trials are required in order to constrain exactly how much improvement one might reasonably expect using such a strategy over other methods in particular elicitation experiments. However, the strategy makes use of real-time experimental design between elicitation trials including all pertinent information available; this must be strategically optimal, although the details of the algorithm can probably be improved in future (e.g., by including graphical illustrations of probabilistic inconsistencies in real time as described above).

The assignment of uncertainties in matrix **C** is probably the least satisfactory part of the elicitation algorithm described herein. If uncertainties in the P_{K_i}'s are symmetric then uncertainties in $P_{K_iK_j}$'s are usually not symmetric. This fact is responsible for most of the approximations employed: the requirement imposed by most experimental design algorithms (that of Curtis *et al.* (2004) included) to represent uncertainties by symmetric standard deviations implies that uncertainties in equations involving terms like $P_{K_iK_j}$ are necessarily only crude approximations. Although the approximations currently employed are adequate to provide more accurate results than would be provided by random trials, improving this aspect of the algorithm should be the subject of future research.

As in many other studies, our strategy would attempt to compensate for various natural human biases (see review in Baddeley *et al.* (this volume) by including multiple experts. If possible, trials should be continued until an equilibrium level of residual uncertainty is attained. Once this equilibrium has been reached, the addition of further randomly selected experts to those already interrogated should have little effect on the

probability estimates and their uncertainties. This test will diminish some of the effects of over-confidence of individual experts and also in the base-rate neglect heuristic. However, if all or most experts suffer from identical biases in either of these respects, these effects will not be removed.

By solving equations (8b) using equations (9a) and (9b) using data from multiple experts, the strategy also creates a consensus distribution by taking a least-squares fit to the set of individual probability distributions. By differentiating estimates from different experts explicitly within the probability vector and design matrices, we might also allow the design algorithm to select from which expert information should be elicited in each trial. Thus, the algorithm itself might be designed to ameliorate the other common problem of anchoring. Lastly, by careful choice of the set of experts considered such that their ranges of experience overlap minimally we can also diminish the availability bias. Additionally, this would presumably reduce off-diagonal correlation terms in matrix **C** (which are ignored by the current algorithm), since inter-expert correlation is likely to be reinforced if experts have similar background knowledge or experience. Hence, the strategy presented should help to diminish many of the principal biases that commonly affect elicitation procedures.

We return finally to the situation of a well intersecting two rock types as described in the Introduction and in Figure 1(a). Say the geology was known from prior information to have been formed in a siliciclastic, fluvial deltaic environment. A method to quantify an approximate conditional prior distribution for the rock types intersected by the well is the following: Define several sedimentary rock types and for each of models 1-9 assign rock types throughout the model in 3-D. Then calculate the histogram of rock types that could be intersected by a well drilled vertically through that environment at any horizontal location. Calculate the mean weighted histogram of rock types across all models, where each individual model histogram is weighted by (i) the best estimate of the prior probability of that model occurring given in Figure 6 and, if desired, by (ii) the proportion of the diffusion parameter space represented by that model given the sampling of this space shown in Table 1 (i.e., by the inverse sampling density). Normalising this final histogram to have unit area results in a prior probability distribution of the rock

types intersected by the well, conditional on all of the same conditions as those pertaining to the prior distribution in Figure 6.

Conclusions

In many situations central to academic and business research, quantitative information must be elicited from experts. Such information is often valuable, uncertain, and difficult to obtain, and must be derived by directly interrogating people residing in many different geographical locations. In such situations, poor results often occur due to expert over-confidence and other natural human biases. This paper presents a new, optimised method to mediate the effects of such biases in order to gather more accurate and probabilistic information. The information thus obtained forms prior probability distributions for further analysis or decision-making.

The new elicitation strategy involves real-time design of elicitation trials based on all available information. The proposed strategy allows more information to be gained from expert elicitation than would be gained through non-optimally designed interrogation. We demonstrate this improvement in a simple experiment in which the conditional probability distribution (or relative likelihood) of a suite of 9 possible models of fluvial-deltaic geologies was elicited, both optimally and randomly, from a single expert in a location remote from the elicitor. These results will be augmented in future by a more extensive application involving multiple experts in multiple locations, and increased diversity of geological settings.

References

Baecher, G. B. 1988. Judgemental probability in geotechnical risk assessment. *Technical report, prepared for The Office of the Chief, US Army Corps of Engineers.* World Wide Web Address: http://www.ence.umd.edu/~gbaecher/papers.d/judge_prob.d/judge_prob.html.

Baddeley, M., Curtis, A. & Wood, R. 2004. Herding in an uncertain world: the role of prior information. *In this volume.*

Boschetti, F. & Moresi, L. 2001. Interactive inversion in geosciences. *Geophysics***, 66,** 1226-1234.

Burgess, P. 2004. Exploring internal and external forcing of carbonate platform strata with a numerical forward model. *In this volume.*

Coupé, V. M. H. & Van der Gaag, L.C. 1997. Supporting Probability Elicitation by Sensitivity Analysis. *Lecture Notes in Computer Science: Knowledge Acquisition, Modeling and Management,* **1319**, 335-340.

Curtis, A. 1999a. Optimal experiment design: Cross-borehole tomographic examples. *Geophysical Journal International,* **136,** 637-650.

_____1999b. Optimal design of focussed experiments and surveys. *Geophysical Journal International,* **139**, 205-215.

_____ & Lomax, A. 2001. Prior information, sampling distributions and the curse of dimensionality. *Geophysics,* **66,** 372-378.

_____ Michelini, A., Leslie, D. & Lomax, A. 2003. A deterministic algorithm applied to experimental design applied to tomographic and microseismic monitoring surveys. *Geophysical Journal International,* In Press

Griffiths, C. M., Dyt, C., Paraschivoiu, E. & Liu, K., 2001. SEDSIM in Hydrocarbon Exploration. *In:* Merriam, D. F. & Davis, J. C. (eds) *Geological Modeling and Simulation, Sedimentary Systems*. Kluwer Academic/Plenum Publishers.

Juslin, P. & Persson, M. 2002. Probabilities from Examplars (PROBEX): a "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science,* **26**, 563-607.

Kahneman, D., Slovic, P. & Tversky, A. (eds) 1982. *Judgement under Uncertainty: Heuristics and Biases.* Cambridge University Press.

Lau, A.-H. & Leong, T.-Y. 1999. Probes: a framework for probability elicitation from experts. *Amia Annual Symposium American Medical Informatics Association,* 301-305.

Lindley, D. V., Tversky, A. & Brown R. V. 1979. On the reconciliation of probability assessments (with discussion). *Journal of the Royal Statistical Society*, **142,** 146-180.

Maurer, H. & Boerner, D. 1998. Optimized and robust experimental design: a nonlinear application to EM sounding. *Geophysical Journal International*, **132**, 458-468.

_____, _____ & Curtis A. 2000. Design strategies for electromagnetic geophysical surveys. *Inverse Problems*, **16,** 1097-1117.

Menke, W. 1989. *Geophysical Data Analysis: Discrete Inverse Theory* (International Geophysics Series). Revised edition (August 1989). Academic Press, Orlando, Florida.

Morgan, M. G. & Henrion, M. 1990. *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis.* Cambridge University Press, Cambridge.

Rabinowitz, N. & Steinberg, D. M. 1990. Optimal configuration of a seismographic network: a statistical approach. *Bulletin Seismological Society America*, **80**, 187-196.

Renooij, S. 2001. Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review*, **16**, 255-269.

Scales, J. 1996. Uncertainties in seismic inverse calculation. *In*: Jacobson, B. H., Mosegaard, K. & Sibani, P. (eds.) *Inverse Methods*, Springer-Verlag, 79-97.

Skinner, D. C. 1999. *Introduction to Decision Analysis.* Probabilistic Publishing.

Steinberg, D. M., Rabinowitz, N., Shimshoni, Y. & Mizrachi, D. 1995. Configuring a seismographic network for optimal monitoring of fault lines and multiple sources. *Bulletin Seismological Society America*, **85**, 1847-1857.

Tetzlaff, D. & Priddy, G. 2001. Sedimentary process modeling: From academia to industry. **In:** Merriam, D. F. & Davis, J. C. (eds) *Geological Modeling and Simulation, Sedimentary Systems.* Kluwer Academic/Plenum Publishers.

_____ & Harbaugh, J. W. 1989. *Simmulating clastic sedimentation: Computer methods in the geosciences.* Van Nostrand Reinhold, New York, 202 pp.

_____ 2004. Input uncertainty and conditioning in siliciclastic process modelling. *In this volume.*

Wijns, C., Poulet, T., Boschetti, F., Dyt, C., & Griffiths, C. M. 2004. Interactive Inverse Methodology Applied to Stratigraphic Forward Modelling. *In this volume.*

Wood, R. & Curtis, A., 2004. Geological Prior Information, and its application to geoscientific problems. *In this volume.*

Tables

Table 1. *Shows model index numbers (column 1) referred to in the main text, the models included within each trial (columns 2-10) and the values used in each model simulation for the diffusion and transport coefficients, and for basement erodability (columns 11-13). In columns 2-10, notation **T1**, **T2**,… stands for Trial 1, Trial 2,… and the numbers in each column show those five models that were included within each trial. Models presented to the subject in each trial were numbered a-e as shown in this table.*

Model Indices	Models Included									Diffusion Coefficient	Transport Coefficient	Erodability
	T1	T2	T3	T4	T5	T6	T7	T8	T9			
1		b		b			b	b		0.7	1	1
2				a		a	a	a	a	1	1	1
3					a			c	b	1.3	1	1
4				c		b		d		2	1	1
5					b	c		e	c	3	1	1
6					c	d	c			5	1	1
7	d	c	c	d	d		d		d	10	1	1
8		d		e						20	1	1
9	e				e	e	e		e	30	1	1
10	a									1	1	0.7
11			a							0.9	1	1
12	b									1.5	1	1
13	c	a	b							2	1	1
14		e	d							1	0.3	1
15			e							1	2	1

Table 2. *Shows model index numbers referred to in the main text (column 1) and likelihood estimates assigned by the subject during each trial (columns 2-10). Only relative likelihoods (ratios) between different models should be interpreted (see main text). Only models 1-9 are shown since these are samples of our desired conditional distribution (see main text). In columns 2-10, notation* ***T1****,* ***T2****,… stands for Trial 1, Trial 2,… . Trial 3 provided no useful information about the conditional distribution in question.*

Model Indices	**Likelihood Estimates**								
	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**
1		0		5			2	6	
2				10		9	10	7	6
3					10			10	10
4				8		10		9	
5					7	8.5		8.5	9
6					8.5	8	7		
7	2	3		3	8		4		9.5
8		4		1					
9	1				9	7	3		2

Figures

Fig. 1. Representation of rock types intersected by a well in a Cartesian model space. (a), (b) and (c) show a well intersecting 2, 3 and 4 rock types respectively on the left, and the corresponding model space representation on the right.

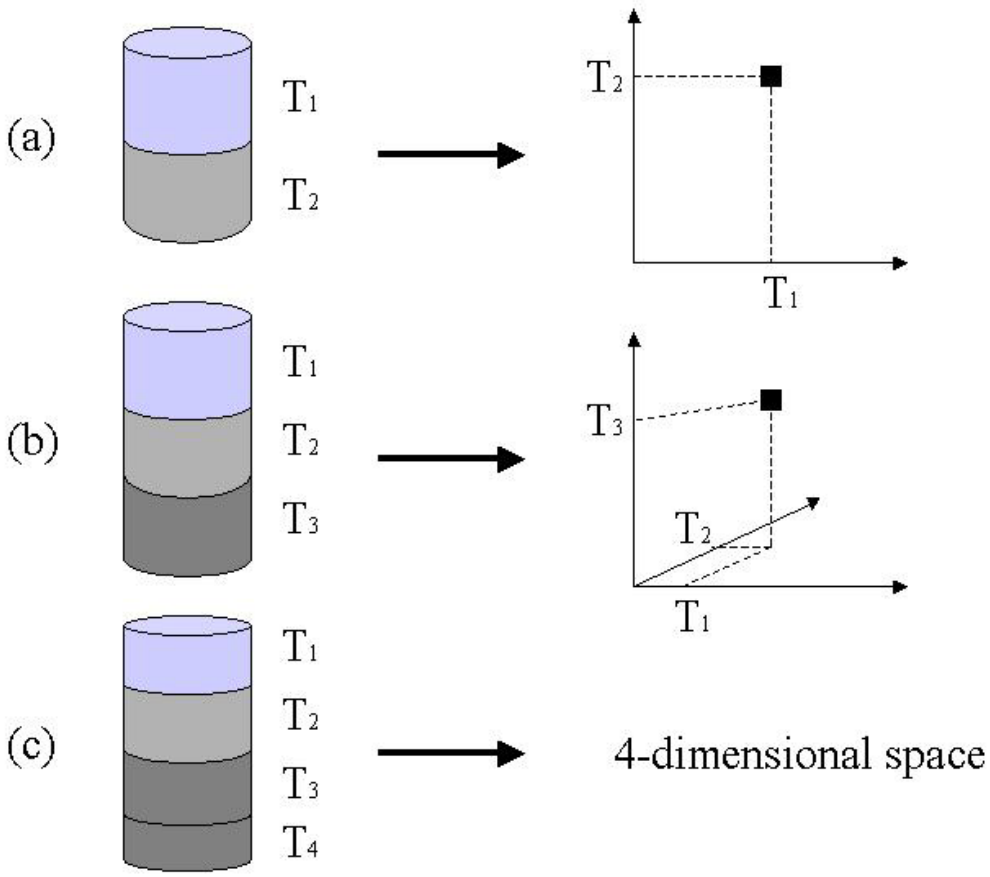

Fig. 1

Fig. 2. (a) Geological cross-section with logs of rock type through two intersecting wells, (b) Possible parameterisation of rock type in the 2D section comprising cells within which any rock type might occur in the Earth. This model space has dimension N. (c) Representation of the N-dimensional model space in Cartesian form. However, if prior information (e.g., from a geological modelling package) exists, the possible model space can be restricted to a manifold shown in (d).

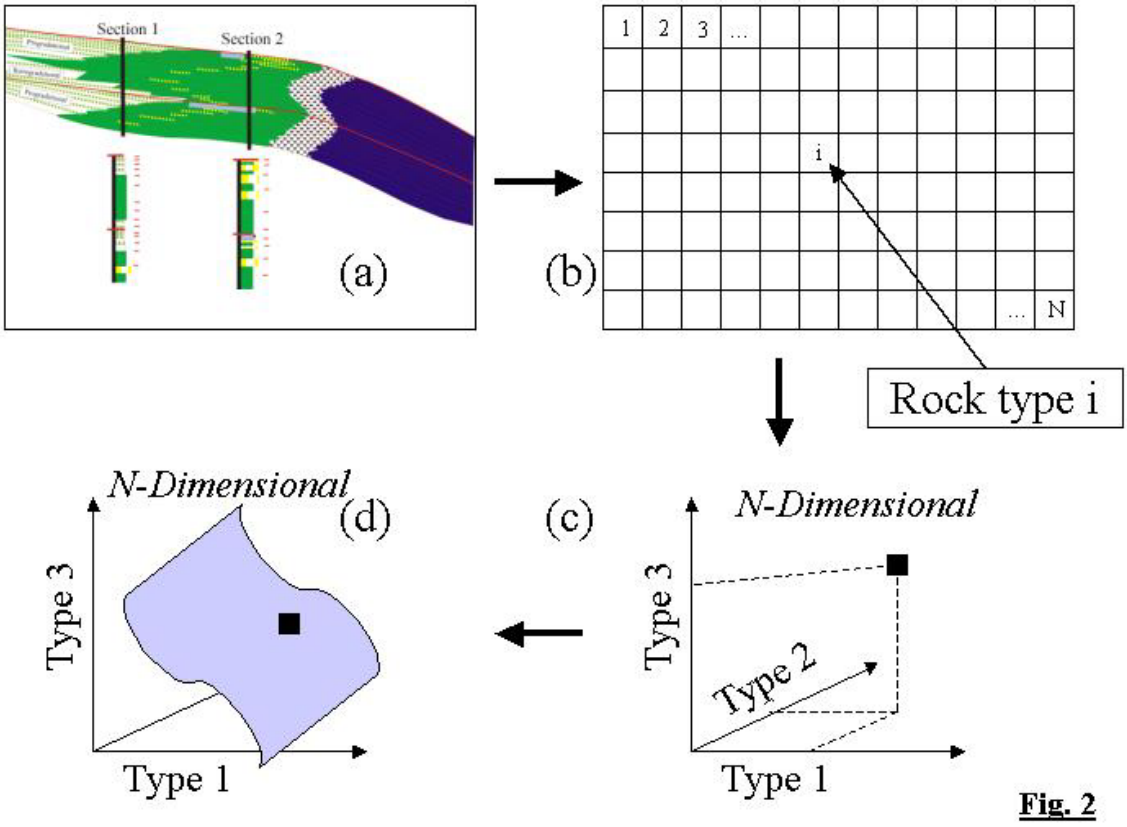

Fig. 2

*Fig. 3. Representation of the relationship between parameters **q** in parameter space **Q** that are input to the Geological Process Model GPM, and the output geological sections parameterised by models **m** in model space **M**.*

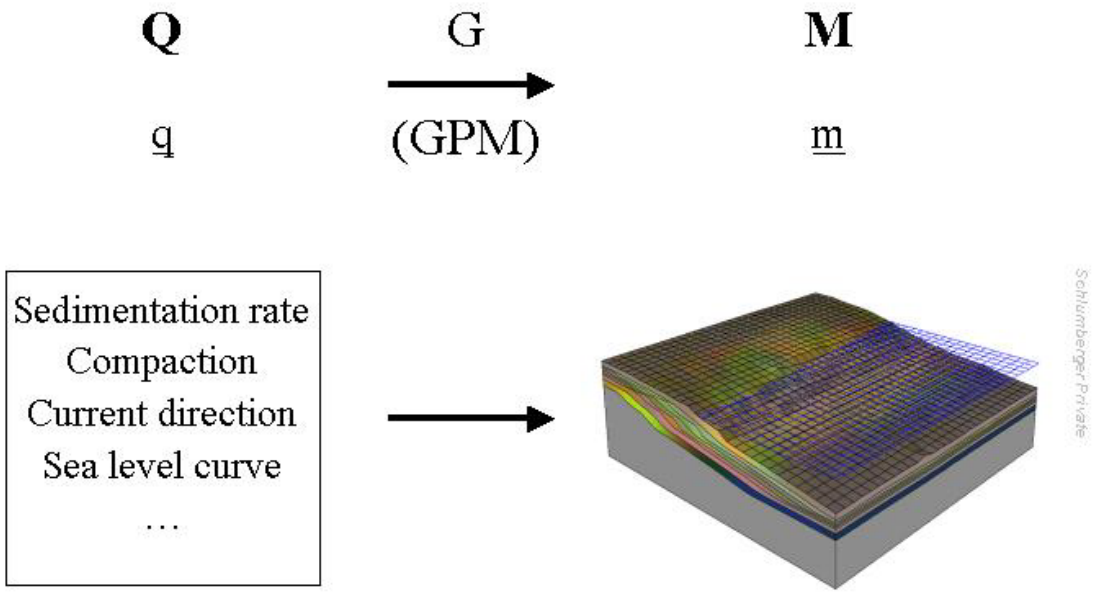

Fig. 3

Fig. 4. Four plots of two output models from GPM. The left plots were generated using a diffusion coefficient of 1, the right plots using a diffusion coefficient of 30. Top plots show the top surface with the water layer removed, lower plots show cross-sections through the centre of each model showing the prograding sequence developed in the sub-surface. Colours represent deposited rocks of different grain sizes. Plain light grey represents basement rock.

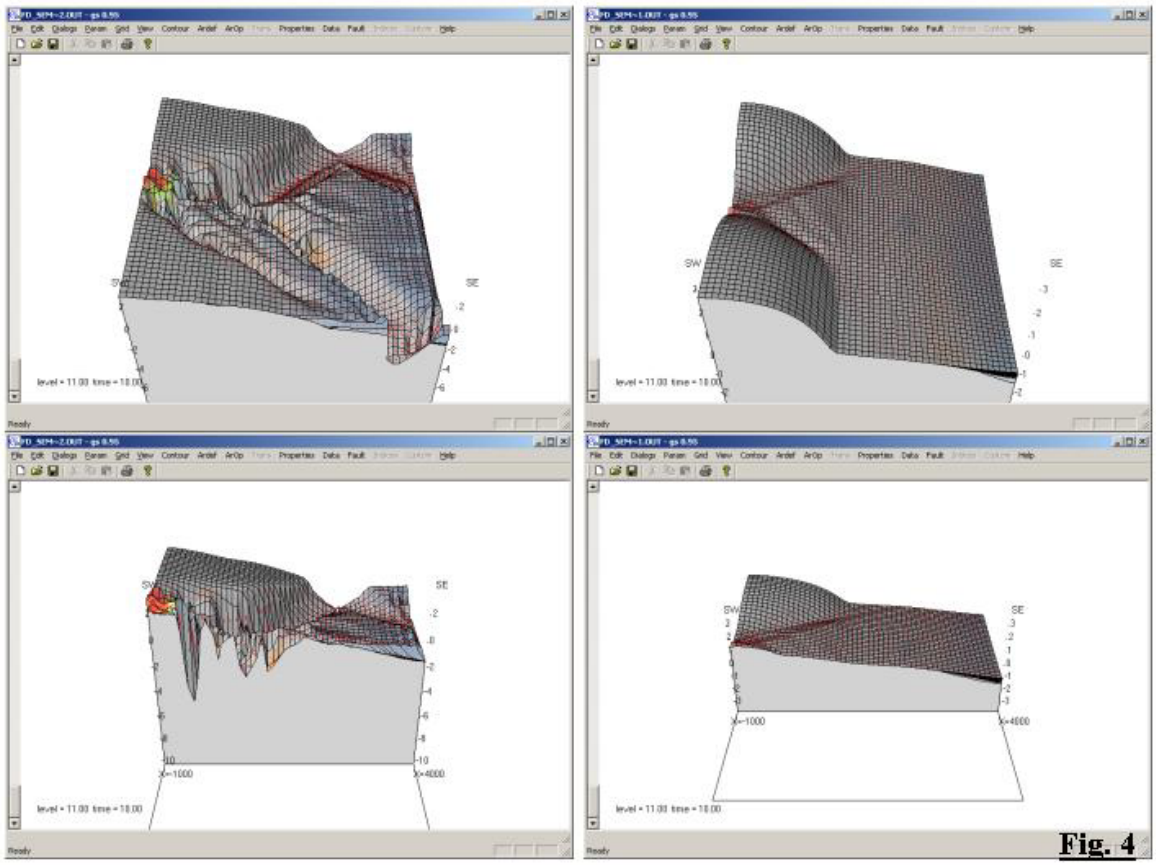

Fig. 4

Fig. 5. Non-normalised prior distribution (black line) with one standard deviation uncertainties (vertical bars) after likelihoods had been elicited for models 1-5 (top-left), 1-6 (top-right), 1-7 (bottom-left) and 1-8 (bottom-right).

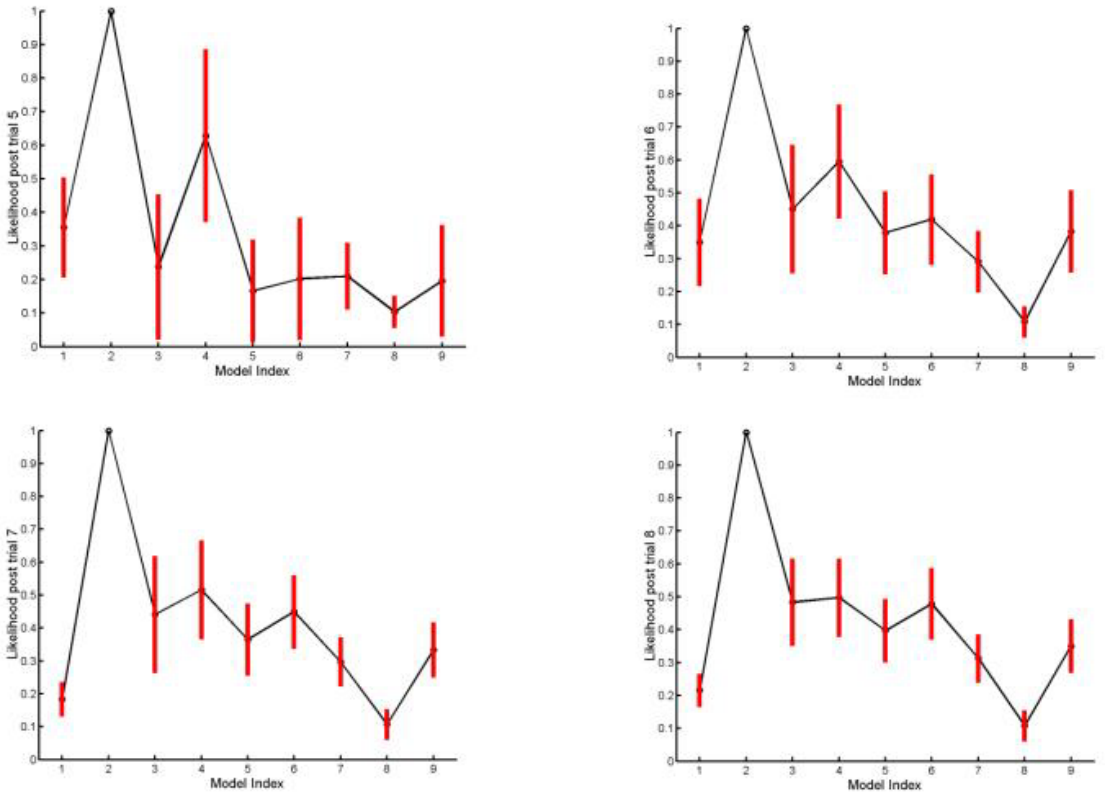

Fig. 5

Figure 6. *Plot of the final non-normalised prior distribution (black line) and its one standard deviation uncertainties (vertical bars) after likelihoods had been elicited for models 1-9 (bottom-right).*

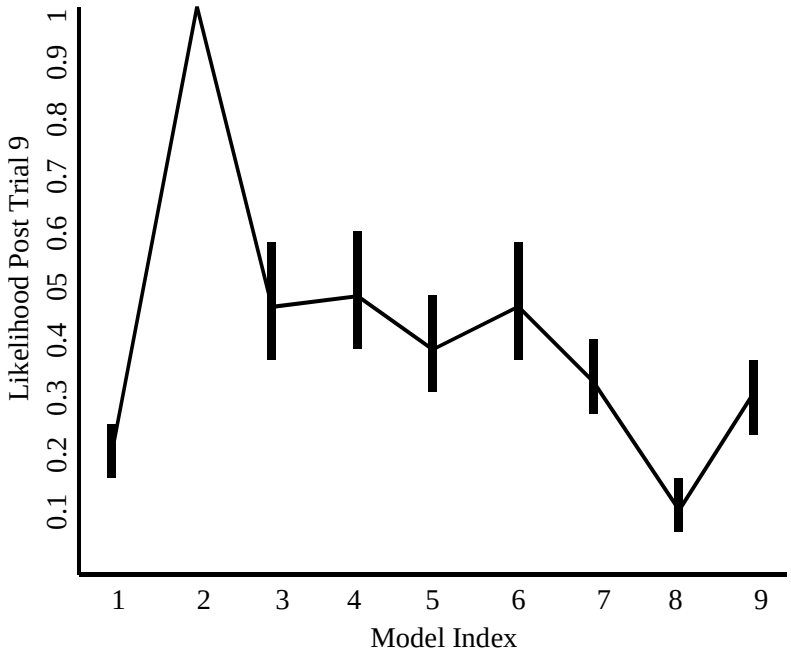

Fig. 6