Research Note

GEOSPHERE

GEOSPHERE; v. 15, no. 1

https://doi.org/10.1130/GES01586.1

4 figures; 6 tables; 1 set of supplemental files

CORRESPONDENCE: clare.bond@abdn.ac.uk

CITATION: Randle, C.H., Bond, C.E., Lark, R.M., and Monaghan, A.A., 2019, Uncertainty in geological interpretations: Effectiveness of expert elicitations: Geosphere, v. 15, no. 1, p. 108–118, https://doi.org/10.1130/GES01586.1.

Science Editor: Shanaka de Silva
Associate Editor: Jose M. Hurtado

Received 18 July 2017
Revision received 1 October 2018
Accepted 6 November 2018
Published online 10 January 2019

Uncertainty in geological interpretations: Effectiveness of expert elicitations

Charles H. Randle[1], Clare E. Bond[1], R. Murray Lark[2,*], and Alison A. Monaghan[3]

[1]Geology and Petroleum Geology, School of Geosciences, University of Aberdeen, Kings College, Aberdeen, AB24 3UE, UK
[2]British Geological Survey, Keyworth, Nottinghamshire, NG12 5GG, UK
[3]British Geological Survey, Edinburgh, EH14 4AP, UK

ABSTRACT

Uncertainty in geological interpretations creates an often unquantified risk for sub-surface industries. The challenge of quantifying interpretation uncertainty has been addressed using various methods. For interpretation of borehole data, empirical quantification of uncertainties can be derived from comparison of interpretations with a withheld set of borehole data not used in the interpretation. This approach requires dense, high-quality borehole data sets. A proposed alternative is to use expert elicitation to extract expert geologists' mental models of uncertainty. We investigated whether expert elicitations are a viable alternative to the direct quantification of uncertainty in three different geological settings by comparing elicited distributions to empirically derived uncertainty distributions. We show that uncertainty distributions derived from expert elicitations are different from those observed in empirical uncertainty quantification. This means that expert elicitations are not as appropriate for estimating uncertainty as these empirical approaches. Expert elicitations, however, offer other benefits to an interpretation workflow, such as providing insight into and challenging different conceptual models of the geology.

INTRODUCTION

A key component of many explicit geological modeling workflows is the use of raw data by expert geologists to build interpretations of the geology along cross sections (e.g., GSI3D; Kessler and Mathers, 2004). These interpretations will always be uncertain to some degree (Mann, 1993). Understanding possible differences between models based on interpretations and the true geology is important because it allows the end user of a geological model to assess a model's uncertainties and how these may impact decisions made using the model.

Differences between the modeled and true geology are considered as errors, the presence of which causes uncertainty that can be described quantitatively by a statistical distribution. The distributions provide an indication of the range of probable differences in modeled geology. As a result, they are important for end users because they indicate the potential range in model outcomes from interpretation.

Methodologies for empirically quantifying uncertainties in cross sections derived from borehole data have been developed (Lark et al., 2013, 2014; Randle et al., 2018). In these methodologies, data are withheld from a set of boreholes that are interpreted by geologists to create cross sections. The withheld data are used to measure the difference between the interpretations and the true geology, as recorded in the withheld borehole. This difference is referred to as the error in the interpretation. The distribution of these errors, and hence the uncertainty, is analyzed statistically to identify factors (such as the local density of boreholes) that determine how the uncertainty behaves. The intention is to see if prediction of uncertainty is possible in future geological settings by using this behavior.

Empirical quantification of uncertainty is time consuming and requires dense high-quality data sets, ideally with multiple geologists, to create a range of modeled geologies. In situations where these dense borehole data sets and geologists are not available, there is no proven methodology to follow.

The most common limiting factor is the availability of data. In built-up areas, quantification of shallow subsurface uncertainty is often viable due to the relatively high sampling of the subsurface (from infrastructure projects and groundwater management, etc.). Deeper geologies or remote regions that have fewer boreholes provide a greater challenge for quantification of uncertainty.

An experienced geologist who has undertaken interpretations of many data sets may have a mental model of the uncertainty in these interpretations. A mental model is a mental representation of the geology (e.g., Libarkin et al., 2003) and is used here to reflect how a geologist may also visualize uncertainty associated with their mental model of the geology. This mental model reflects their awareness of how the particular configuration of geological observations constrains their interpretation. There may be flexibility in the position and continuity of interpreted horizons and other geological features. If it were possible to access the geologists' mental model of uncertainty in a cross section obtained in a particular setting by a particular

*Currently at the University of Nottingham, School of Biosciences, Sutton Bonington Campus, Sutton Bonington, Leicestershire LE12 5RD, UK

workflow and to represent this model in terms of a statistical distribution, then this may prove a more cost-effective way to quantify uncertainty than by empirical methods. While the mental model of uncertainty that a geologist develops may be tacit, the assumptions that lie behind it need not be. The geologist is aware, for example, of whether they consider the possibility that a contact between two lithologies is discontinuous because of faulting. Any process to extract the mental model of uncertainty (for example through expert elicitation) will help to make assumptions associated with the model explicit. Expert elicitation facilitates discussion between geologists about model assumptions and provides the model user with a context for the interpretation. The objective of this study is to assess the feasibility of using expert elicitation for the quantification of uncertainty in cross-section interpretations using one methodology from the set of methods known as expert elicitation.

Expert elicitation is an umbrella term for the extraction of information from either individual or groups of experts in a chosen field when that information is held in tacit form and is not easily written down. Within geoscience, elicitation has taken several forms of both structured and unstructured experiments. For example, Bond et al. (2007) investigated the idea of conceptual uncertainty through elicitation, observing a large range in mental, or conceptual, models for the same seismic image. The work of Polson and Curtis (2010) followed this by eliciting the probability of the existence of various geological structures within a location. Their results highlighted not only the probability of the existence of the geological structures but perhaps more interestingly the effects of group dynamics and the misunderstandings of the experts on the uncertainty ranges elicited.

We are not aware of any studies in the Earth sciences that have evaluated the effectiveness of expert elicitation in uncertainty quantification. Other studies have come close; the work of Bond et al. (2007) found that the majority of experts (79%) did not interpret the correct tectonic setting for a synthetic seismic image, and Lark et al. (2015) aimed to quantify uncertainty in map line work by eliciting uncertainty in a set of hypothetical locations; however, there was no independent validation of the results. So, here we test whether uncertainty distributions derived from expert elicitations are comparable with distributions obtained empirically.

We do this through expert elicitation of uncertainty in an interpretation of specific cross sections where the uncertainty has been quantified empirically through the experiments of Lark et al. (2014) and Randle et al. (2018). The elicited distributions of uncertainty are compared to the empirically derived uncertainty distributions built from those experiments. Should the elicited distributions differ significantly from the empirically derived distributions, then the expert elicitation process may be inappropriate for the problem posed, or it may indicate that the mental models of the geologists are not representative of the true uncertainty present. In contrast, if the elicited uncertainties are consistently similar to the empirically derived uncertainty distributions, then we can conclude that expert elicitation is a suitable alternative to empirical approaches.

METHODOLOGY

Group expert elicitation can take two forms. The first involves the elicitation of distributions from each member of the group, then combining each member's distributions with other members' distributions (e.g., Cooke, 1994). The combination is not carried out equally however. The more experienced experts' distributions are given a greater weighting than those with less experience. This methodology has the advantage of minimizing the risk of issues with intragroup dynamics and aims to ensure that the most experienced experts have the greatest influence.

The second option, employed here, is to use the group to come to a consensus rather than apply a numerical aggregation to each individual's distributions (e.g., Oakley and O'Hagan, 2016). This approach is known as behavioral aggregation and has the disadvantage of reliance on the group to come to a consensus as well as the capability to consider every expert's opinion. The key advantage is that the procedure makes the differences between expert views explicit and encourages discussion. If the process is effectively moderated, we might expect the consensus to be more informative than a numerical weighting of contrasting opinions where the contrasting assumptions are never explicitly addressed or resolved. Reagan-Cirincione (1994) showed how behavioral aggregation can be effective with careful moderation (to facilitate discussion and avoid domination by particular individuals) and visual feedback in the form of the elicitation panel's individual distributions and the group consensus distribution as it evolves.

SELECTING WHERE TO ELICIT

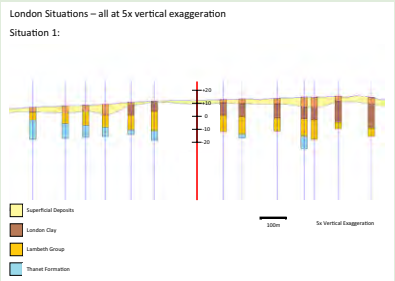
Uncertainty was elicited in three geological settings for which previous experiments had quantified the uncertainty (Lark et al., 2014; Randle et al., 2018). The first elicitation concerned the uncertainty in the position of the base of the London Clay Formation—a marine clay conformably deposited on lower units in a layer-cake stratigraphy (Ellison et al., 2004). The second evaluated uncertainty is the rockhead at the base of the superficial deposits below the city of Glasgow—an erosional surface resulting from glacial and glaciofluvial processes (Hall et al., 1998). The final elicitation was for the rockhead at the base of the superficial deposits below Manchester—an erosional surface similar to that of Glasgow but with less variation in lithology and elevation (Price et al., 2012).

In our experiments, we tested uncertainty through the use of notional boreholes. At each notional borehole, we elicited the range of potential positions of the geological contact of interest by eliciting the expert's prediction of the uncertainty in the position of the contact in the borehole. Multiple notional boreholes were used at each geological setting to capture the predicted uncertainty in different geological contexts along section lines. Each of these notional boreholes is referred to in the rest of the paper as a "situation."

Situations were selected to cover a range of the parameters found to be related to uncertainty in the empirical uncertainty quantification experiments

of Lark et al. (2014) and Randle et al. (2018). In the London setting, Lark et al. (2014) found that the variance of interpretation errors can be modeled as a function of distance to nearest borehole. We therefore identified situations for the London elicitations in which distance to the nearest borehole varied. For Glasgow, Randle et al. (2018) found that the variance of interpretation errors could be modeled by a combination of the distance along the cross section and depth of the rockhead below the present-day land surface; thus, situations were selected to cover a range of these values. For Manchester, Randle et al. (2018) did not identify any parameter within the individual sections that could be used to model the variance in interpretation error; therefore, an alternative situation selection method had to be used. In Randle et al. (2018), for Manchester, experts interpreted sections in which boreholes had been removed; then, the interpretations were tested against those removed 149 boreholes, giving a measure of interpretation error. Elicitation situations were selected from the removed borehole locations in Randle et al. (2018) to cover the range of interpretation errors, representing the uncertainty.

Images of each situation were created using the section viewer in the software GSI3D. GSI3D was used to create the image because this is the software used for similar interpretations at the British Geological Survey (BGS). The aim is to replicate a typical view that a British Geological Survey expert would have while carrying out an interpretation. Each image included a scale centered at the elevation at which the target surface had been interpreted in the corresponding empirical uncertainty quantification experiment. This was done to minimize any bias that could occur if the experts were asked to interpret relative to some other point such as the ground surface or sea level. This also ensured that the elicited distributions were easily comparable to the uncertainty quantification experiments, where uncertainty was presented relative to an interpretation. The images used are available in the Supplemental Material[1].



[1]Supplemental Material. Consists of elicitation briefing material and situations presented to experts for interpretation of location of rockhead. Please visit https://doi.org/10.1130/GES01586.S1 or access the full-text article on www.gsapubs.org to view the Supplemental Material.

EXPERT SELECTION

All experts were from geological modeling teams within the British Geological Survey. This single source of experts was chosen so that all participants would have a similar approach to geological modeling, having been trained on and with experience of using the cross-section modeling package GSI3D, while still having individual experience and prior knowledge. Each expert was also chosen for their experience in carrying out interpretations within the specific geological settings studied or their experience in similar settings (i.e., on sedimentary packages above bedrock). Each participant was asked to self-assess their experience at the start of the elicitation process to confirm our selection was appropriate. No apparently significant differences in experience were noted. In total, we recruited six experts to the London elicitation exercise and four experts each for Manchester and Glasgow. All the experts were invited to take part on a voluntary basis. Note that the methodology can be applied to any group of any composition, and this experiment was not designed to evaluate geological modeling at the British Geological Survey specifically.

ELICITATION PROCESS

The Sheffield Elicitation Framework (SHELF) procedure as described in Oakley and O'Hagan (2010) was used as a guideline for all three of our elicitations. Elements of the procedure were omitted (registration of conflict of interest and post-elicitation meetings) due to logistical constraints, but the overall form of the elicitation mirrored the SHELF procedure. Each started with a briefing document (available in Supplemental Material [footnote 1]) sent out to all participants two weeks before the date of the elicitation. The briefing document contained a description of the process, the reasoning for taking part in the elicitations, and a detailed description of how the experts should consider the elicited uncertainty distributions to work, with demonstrations of the principles of accuracy and precision and how to treat them in the elicitations.

On the day of the elicitation, the session started with a brief run-through of the elicitation process, with the briefing document providing a frame of reference. Care was taken to ensure that the participants fully understood the process and principles and that any misunderstandings were addressed for each of the experts. This was to ensure that all participants had the same understanding of the elicitation process and the thinking behind it.

For each elicitation situation, the same four-step process was followed (Fig. 1). The first step was to present the group with the geological situation both in paper in the form of a handout for each expert and simultaneously on a projector visible to the whole group (Fig. 1A). The experts were permitted to draw on and annotate their handouts as they saw fit, something that the majority of the experts did for most situations.

After a brief period, the experts were asked as a group to suggest minimum and maximum elevations of the target contact (the base of the London Clay for London and rockheads for Glasgow and Manchester; Fig. 1B). The reason for this step is that a group exercise will ensure that the group is considering the same range of possibilities, thus easing the process of coming to a consensus later.

Step 3 (Fig. 1C) is the elicitation of the median and upper and lower quartiles of the distribution. This was done individually at first, with the experts writing their numbers on a form. After all experts had written their values, the forms were retrieved, and the values were presented on a flip board. Care was taken to ensure that no expert felt rushed to determine their values. In all cases, there was no conferring between experts.

For the final step (Fig. 1D), the experts were asked to come to a consensus on the values for the median, lower, and upper quartiles. The order in which the consensus was reached for each quartile varied depending on the exact situation and the difference in experts' values. More often than not, the median was the first consensus reached. Experts were not required to state which of the values were their own, and although all of the separate distributions were shown on the flip board, they were not attributed to individuals. However, in all cases where a consensus was not quickly reached, the experts stated which values were theirs and their thinking behind them. While the experts were coming to a consensus, the distributions arising from their individual quartile values were presented graph-

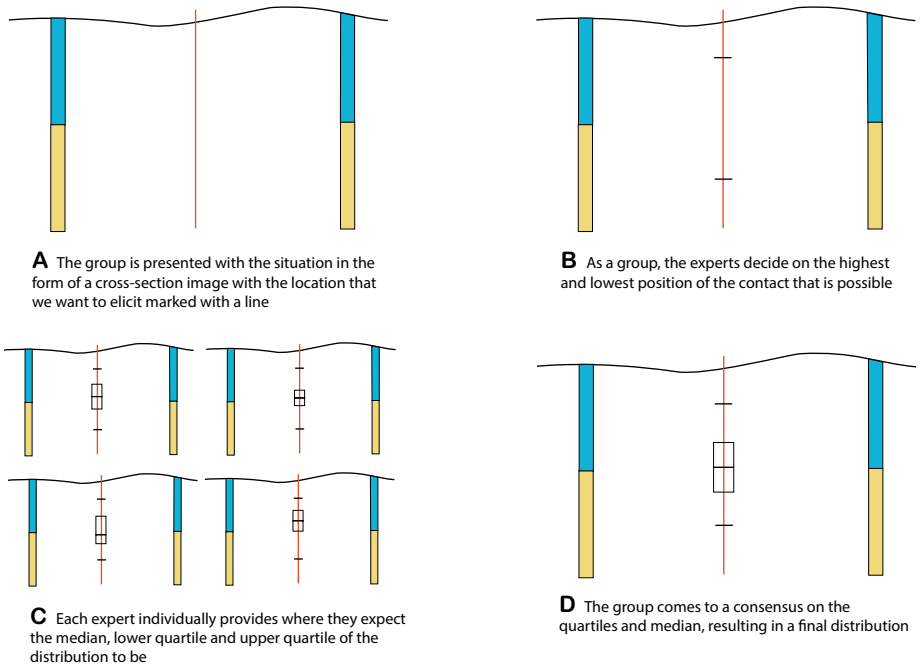


Figure 1. Summary of each step of the Sheffield Elicitation Framework (SHELF) process (A)–(D) used to elicit the uncertainty.

ically on the projector. This allowed them to visualize the differences in each of their distributions when coming to a consensus proved more challenging.

After all distributions had been elicited, there was a brief recap of each of the situations to ensure that each expert was happy with the resulting distributions. This was completed with the intention of allowing the experts to take what they have learned through the elicitation and apply it to all situations, ensuring that the elicited distribution for the first situation was as valid as the last situation. The experts were then asked to give feedback on the elicitation procedure. This included suggestions of procedural improvements and often also included discussion of uncertainty as a whole and how expert elicitations could fit into current workflows.

UNCERTAINTY VALIDATION

After the elicitations, the resulting distributions of uncertainty were compared to statistical uncertainty models built from the results of the relevant uncertainty quantification experiments. Lark et al. (2014) and Randle at al. (2018) created statistical models that were used to determine the causes and indicators of interpretation error. These models also allowed the prediction of the standard deviation of interpretation error. To validate our elicitations, we took the empirically derived standard deviations of error and created uncertainty distributions from them; we refer to these distributions as the empirical uncertainty. Determination of the empirical uncertainty for each geological setting is summarized below.

For London, interpretation error was found to be related to the distance to the nearest borehole, along with a spatial dependence (i.e., uncertainty at two points close to each other was more similar than the uncertainty at two points far from each other). We simulated this error across all of the expert elicitation situations using the rnorm function within the R statistical framework in order to provide a comparable uncertainty distribution. A normal distribution with a mean of zero was simulated because this was the distribution observed in Lark et al. (2014) and Randle at al. (2018). For Glasgow and Manchester, determining the distributions of uncertainty was simpler, because error was not spatially dependent. In Glasgow, the standard deviation of error was found to be related to depth below surface and distance along section, with no spatial dependence. We determined the standard deviation of error for each expert elicitation situation based on these parameters; then we determined the distribution using the qnorm function to determine the quantiles of the distribution. For Manchester, there was no variation in error standard deviation; therefore, the distribution was treated as uniform across all situations and was derived in the same manner as Glasgow. From here on, uncertainty distributions are referred to as empirical uncertainty.

TABLE 1. THE RESULTS OF THE LONDON ELICITATION

(A) Situation	(B) Minimum (m)	(C) Lower quartile (m)	(D) Median (m)	(E) Upper quartile (m)	(F) Maximum (m)
1	–10	–2	0	5	10
2	–10	–3	–1.5	0	5
3	–5	–1	0	1	5
4	–15	–2.5	1	4.5	12
5	–15	–4.5	0.5	5.5	11
6	–18	–6	0	6	18
7	–18	–7	0	5	11
8	–18	–6	0	5	13

Note: Column (A) is the situation number; (B) is the elicited minimum; (C) is the elicited lower quartile; (D) is the elicited median; (E) is the elicited upper quartile; and (F) is the elicited maximum elevation of the base of the London Clay.

RESULTS

London

Results

Table 1 and Figure 2 present the elicited uncertainty distributions for the eight studied situations. We observe a variety of ranges and interquartile ranges, along with both symmetrical and skewed distributions. Discussion was focused on the boreholes adjacent to the situation elicited and the trends in the elevation of the London Clay suggested from them. The variability in the elevation of the London Clay expected by the experts at the scale of the section informed these discussions. In the feedback phase, the participants said that they were happy with the procedure, with the only comment being that the breaks on the scale bar were too widely spaced for the low variability of the London Clay (more closely spaced ticks on the vertical scale bar were used in the later elicitations to aid interpreters). A further point made was that the boreholes presented to the participants showed stratigraphic rather than litho-

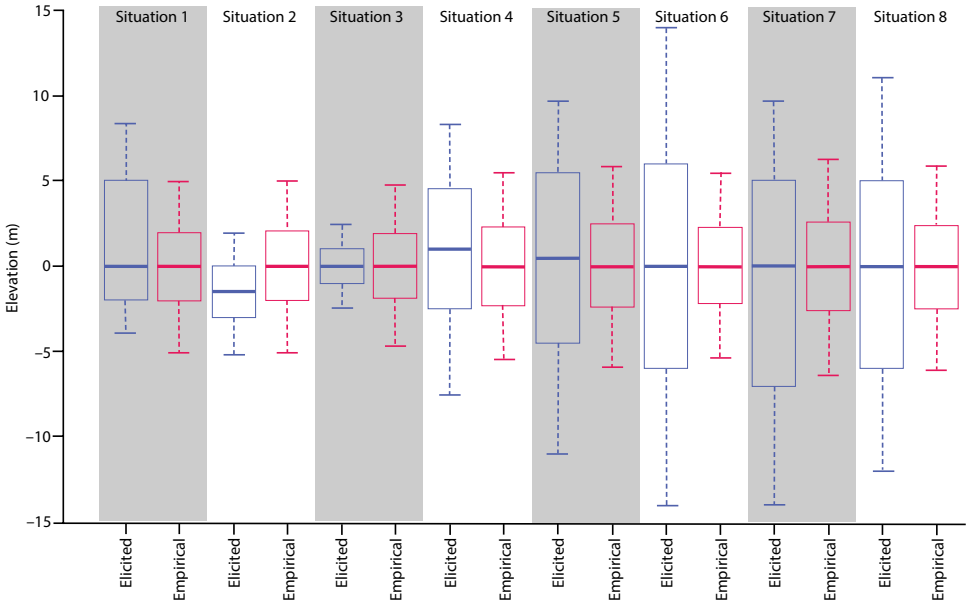


Figure 2. Boxplots comparing the elicited distributions of uncertainty to the empirical distributions for each of the London situations. The ends of the whiskers here represent the 5th and 95th percentiles.

TABLE 2. QUANTILES FOR EMPIRICAL UNCERTAINTY OF THE POSITION OF THE LONDON CLAY DERIVED FROM THE WORK OF LARK ET AL. (2014) AT EACH OF THE ELICITED SITUATIONS

(A) Situation	(B) 5th percentile (m)	(C) Lower quartile (m)	(D) Median (m)	(E) Upper quartile (m)	(F) 95th percentile (m)
1	–5.1	–2	0	2	4.9
2	–5.1	–2	0	2.1	5
3	–4.7	–1.9	0	1.9	4.7
4	–5.5	–2.3	0	2.3	5.5
5	–5.9	–2.4	0	2.5	5.9
6	–5.4	–2.2	0	2.3	5.5
7	–6.4	–2.6	0	2.6	6.3
8	–6.1	–2.5	0	2.4	5.9

Note: Column (A) is the situation number; (B) is the empirical 5th percentile; (C) is the empirical lower quartile; (D) is the empirical median; (E) is the empirical upper quartile; and (F) is the empirical 95th percentile.

logical logs. The participants felt that while it is how the data were presented when the BGS created a 3D model for this region, there may be insights into the potential structures present within lithological logs, and lithological logs would have been available to the modeling geologist. Critical to us is that the same amount of information was presented to the experts in the elicitation experiment as the empirical experiments to which we compare them.

Validation

Table 2 presents the empirical distributions of uncertainty, and Figure 2 shows a comparison of them with our elicited uncertainty distributions. Figure 2 demonstrates that the elicited uncertainty distributions are consistently different from the empirical distributions. The elicited ranges and interquartile ranges are wider than the empirical ranges in six of the eight situations, and the remaining two are narrower. The near-zero median of the empirical distributions is also evident in the elicitations, except for situations 2 and 4, where the distribution is symmetrical but shifted deeper and shallower, respectively. In these two situations, there were boreholes in close proximity that did not prove the base of the London Clay but showed a considerable portion of its thickness—proving where the contact was not present.

Glasgow

Results

For the Glasgow elicitation, we have also observed a wide range of distributions, interquartile ranges and skews (Table 3 and Fig. 3), and the variety is much greater than recorded in London. Discussion was focused on the geo-

TABLE 3. RESULTS OF THE GLASGOW ELICITATION

(A) Situation	(B) Minimum (m)	(C) Lower quartile (m)	(D) Median (m)	(E) Upper quartile (m)	(F) Maximum (m)
1	–10	–3	–0.5	2	9
2	–3.5	–2	–0.5	1	3
3	–9	–4	–2	2	4
4	–13	–5	–2	3	10
5	–4	–2	–1	1	3
6	–5	–1.5	0.5	2	5
7	–10	–6	–2.5	4	7
8	–8	–4	–1	5	10
9	–9	–6	–2	1	7
10	–10	–1	6	10	20
11	–25	–15	–6	0	10

Note: Column (A) is the situation number; (B) is the elicited minimum; (C) is the elicited lower quartile; (D) is the elicited median; (E) is the elicited upper quartile; (F) is the elicited maximum elevation of the rockhead.

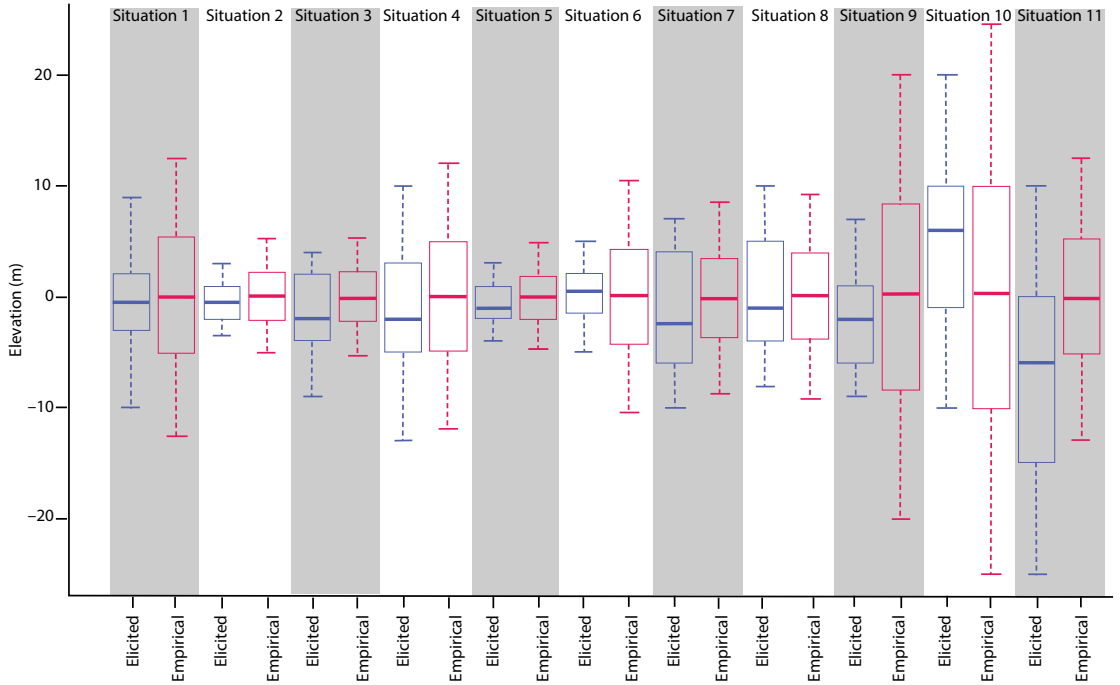


Figure 3. Boxplots comparing the elicited distributions of uncertainty for each of the Manchester situations to the empirical distribution. The ends of the whiskers here represent the 5th and 95th percentiles.

logic structures present within the superficial deposits, such as drumlins, and their effects on rockhead elevation, along with changes in the bedrock geology, such as faulting. In addition, the topography of the ground surface around the elicited situations was repeatedly discussed, with questions posed as to how representative variations in the surface topography are of the underlying geology in built-up areas such as Glasgow where the surface is likely altered by human activity.

Two situations (8 and 10) were identified as locations where a bimodal distribution would be a more appropriate representation of the uncertainty than the options presented by the SHELF process. This is due to there being two conceptual models for the edge of glaciofluvial valleys present in both situations. Situations 8 and 10 were in close proximity to what the experts believed to be a steep-sided glaciofluvial valley. The experts felt that it was unlikely for the slope to be represented within the borehole due to the valley's steep sides; instead, it was more likely for the borehole to intersect with either the base of the valley or the surface outside of the valley structure.

After all situations had been elicited, the experts were presented again with situations 10 and 11 and told that both were situations within the same area, albeit from different boreholes. The results of the elicitations showed two very different interpretations for, effectively, the same cross section. The experts were initially surprised that their elicited distribution was too narrow based on borehole 10 and too broad based on borehole 11. After discussion, the experts felt that their mental models were reflected by the elicited distribution for borehole 10, but they felt that their mental models were incorrect and that they had underestimated the uncertainty present. This example shows the influence of point data in anchoring interpretations (see Bond et al., 2008, and Bond, 2015) for descriptions of anchoring in a geological context).

Validation

Table 4 presents the empirical uncertainty distributions for Glasgow. Comparing them to the elicited distributions (Fig. 3) shows that the elicited uncertainty was close to the empirical uncertainty in five of the 11 situations (2, 4, 5, 7, and 8). In situation 11, the elicited uncertainty was greater than the empirical uncertainty, suggesting underconfidence in this situation. In the remainder of the situations, empirical uncertainty was smaller than elicited uncertainty. This indicates that the experts were overconfident in these situations, underestimating the uncertainty present.

TABLE 4. QUANTILES FOR THE EMPIRICAL UNCERTAINTY FOR GLASGOW DERIVED FROM THE WORK OF RANDLE ET AL. (2018) AT EACH OF THE ELICITED SITUATIONS

(A) Situation	(B) 5th percentile (m)	(C) Lower quartile (m)	(D) Median (m)	(E) Upper quartile (m)	(F) 95th percentile (m)
1	–12.6	–5.1	0	5.1	12.6
2	–5.2	–2.1	0	2.1	5.2
3	–5.3	–2.2	0	2.2	5.3
4	–12	–4.9	0	4.9	12
5	–4.7	–1.9	0	1.9	4.7
6	–10.6	–4.3	0	4.3	10.6
7	–8.7	–3.6	0	3.6	8.7
8	–9.4	–3.9	0	3.9	9.4
9	–20.2	–8.3	0	8.3	20.2
10	–24.5	–10	0	10	24.5
11	–13	–5.3	0	5.3	13

Note: Column (A) is the situation number; (B) is the empirical 5th percentile; (C) is the empirical lower quartile: (D) is the empirical median; (E) is the empirical upper quartile; and (F) is the empirical 95th percentile.

Manchester

Results

The elicited uncertainty distributions are presented in Table 5 and Figure 4. In general, the distributions are narrower than those elicited for London and Glasgow, indicating that the experts believed that uncertainty was lower in Manchester. Discussion centered on the variability of the rockhead at and around each situation and the resulting features of the rockhead that were plausible. Surface topography again proved to be a significant factor in the experts' decisions with repeated mentions of how much the topography had been altered by manmade processes. If the topography were to be significantly altered, it could not be used to determine the subsurface geology. Bimodal distributions were suggested for situations 1 and 4. Situation 1 was the result of the potential presence of either a bench or a channel with an intermediate position of the rockhead not plausible. Situation 4 was affected

TABLE 5. RESULTS OF THE MANCHESTER ELICITATION

(A) Situation	(B) Minimum (m)	(C) Lower quartile (m)	(D) Median (m)	(E) Upper quartile (m)	(F) Maximum (m)
1	–4.5	–2	2 or 0	3 or 1	5
2	–2	–0.75	–0.25	0.375	1
3	–6	–3.5	–1.5	0.5	2
4	–5	–1.5	0.5	2.5	5
5	–3.25	–2	–0.5	1	2.5
6	–12.5	–11.75	–11	–10	–9
7	–3	–1.75	–1	1.75	2.5
8	–12	–6	–4.5	–2.3	1.5
9	–1.5	–0.75	–0.5	–0.2	0.5
10	–2.5	–1.25	–0.5	0.25	1
11	–5	–2	0	2.5	5
	(G) 5th percentile (m)	(H) Lower quartile (m)	(I) Median (m)	(J) Upper quartile (m)	(K) 95th percentile (m)
Empirical	–4.4	–1.8	0	1.8	4.4

Note: Column (A) is the situation number; (B) is the elicited minimum; (C) is the elicited lower quartile; (D) is the elicited median; (E) is the elicited upper quartile; and (F) is the elicited maximum elevation of the rockhead. The values for situation 6 are significantly different due to an error in the positioning of the zero point on the images presented to the experts; hence, their entire distribution appears to be lower in elevation than it should be. The empirical uncertainty derived from Randle et al (2018) is also presented here, showing its (G) 5th percentile, (H) Lower quartile, (I) Median, (J) Upper quartile, and (K) 95th percentile.

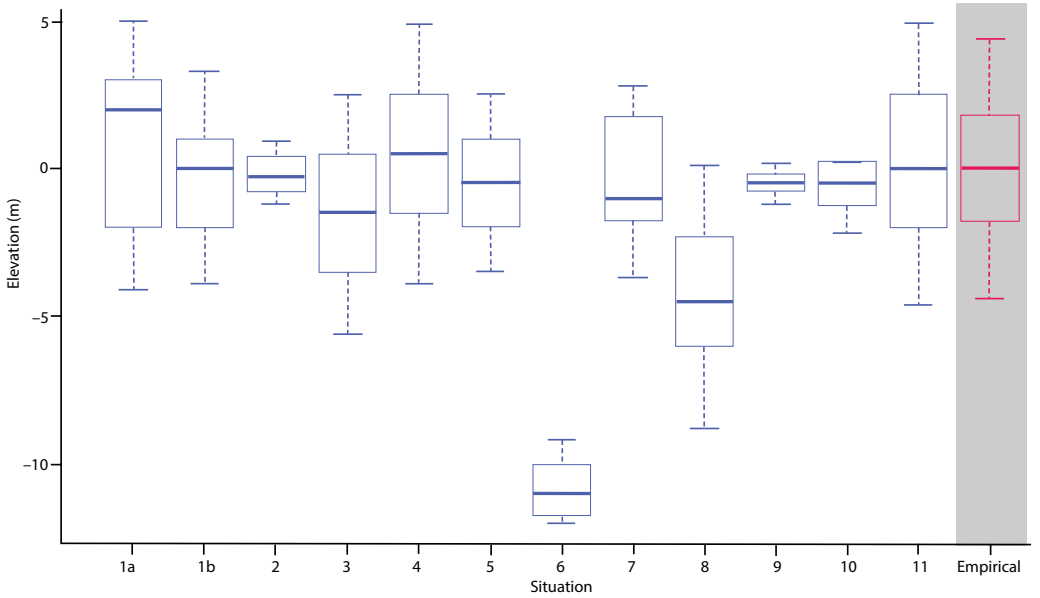


Figure 4. Boxplots comparing the elicited distributions of uncertainty to the empirical distributions for each of the Manchester situations. The ends of the whiskers here represent the 5th and 95th percentiles.

by the potential presence of a backfilled quarry evident from an adjacent borehole with a considerable thickness of artificial ground. Similar to the Glasgow bimodal situations, the experts felt that the sides of the quarry would be steep sided, and, therefore, it would be unlikely for the rockhead defining the quarry wall to be found within the notional borehole, with the borehole likely to sample either the floor of the quarry or the rockhead outside the quarry.

During the discussion phase of the elicitation, after uncertainty had been elicited for all situations, the experts were presented with situations 10 and 11 from the Glasgow experiment as an example of where the elicited uncertainty was significantly different from the empirical uncertainty. The experts felt that the way in which the situations were presented may make a situation seem more certain than it actually is. They also mentioned that they would typically have more data available, in the form of downhole interpretations showing the exact lithology as well as a 3D network of boreholes, which would enable better determination of the processes that could have caused rockhead variation.

Validation

The uncertainty quantification experiment for Manchester in Randle et al. (2018) found no relationship between uncertainty and any parameter relating to the data; hence, the empirical distribution of uncertainty is uniform across all situations (Table 5) and is modeled from the measured distribution of interpretation error.

The empirical uncertainty was similar to the elicited uncertainty for situations 1, 3, 4, 8, and 11; however, the skew and median of the distributions were different in all but situations 11 and 4. The rest of the distributions were narrower than the empirical uncertainty, indicating that the experts were overconfident, resulting in an underestimation of uncertainty in these situations. The medians are also often far from zero, especially situation 6. This indicates that the experts felt that the zero point was set away from where they felt a typical interpretation would be.

SUMMARY

In our three elicitations, we observed both underestimation (Glasgow and Manchester) and overestimation (London) of uncertainty. However, we also observed situations where the elicitations have provided accurate estimations of uncertainty. The overall results are presented in Table 6.

In terms of methodology, the consensus-based approach was successful, with each group capable of coming to a consensus for all but one situation, and with no expert stating that their opinion was being marginalized. The primary drawback found was the existence of bimodal distributions, where a standard single-peaked distribution was not representative of what the experts felt was the actual distribution of error. Combining distributions resulting from two (or more) different geological concepts does not make sense, because the final distribution does not represent expert understanding of the conceptual uncertainty.

TABLE 6. OVERALL RESULTS COMPARING ELICITED AND EMPIRICAL UNCERTAINTY DISTRIBUTIONS

(A) Setting	(B) Number of situations	(C) Number of underestimates of uncertainty	(D) Number of appropriate estimates of uncertainty	(E) Number of overestimates of uncertainty
London	8	2	1	5
Glasgow	11	5	5	1
Manchester	11	6	5	0

Note: Column (A) is the setting; (B) the number of situations; (C) the number of times the elicited distribution underestimated; (D) was approximately equal to; and (E) overestimated the uncertainty relative to the empirical uncertainty distributions.

DISCUSSION

The experiments were designed to test the hypothesis that uncertainty distributions derived from expert elicitations are comparable with distributions obtained empirically. Based on our results, this hypothesis has been rejected. We have observed that elicited distributions often differ from empirically derived uncertainty distributions. In addition, the nature of these differences is inconsistent, with the elicited uncertainty for London being too great and for Glasgow and Manchester too small.

Differences in Uncertainty

There are two potential causes of the difference between the elicited uncertainty and the empirical uncertainty. The first is that the elicitation methodology did not extract the mental model correctly, meaning that some aspect of the methodology was flawed. The evidence to support a flaw in mental model extraction based on feedback from the experts is, however, limited. The only methodological aspect mentioned by the experts was that of data availability, where in a typical interpretation methodology, they would have additional data such as detailed borehole logs, geologic maps, or 3D data. The counterargument to this was that the only additional information available to the interpreters taking part in Lark at al. (2014) and Randle et al. (2018) was a simplified geologic map. Therefore, the comparison of the elicited uncertainty to the empirically derived uncertainty is not biased by a lack of data available to experts in the elicitation experiments described here.

The alternative is that the mental model was extracted correctly but is not representative of the true uncertainty. In many sciences, experts often have their interpretations validated. For example, a doctor will often discover whether their diagnoses are correct; engineering experts will find out whether their designs work appropriately. In some cases within geoscience, this is also true. For example, a geotechnical surveyor will discover if their estimations of slope stability are flawed, and a resource geologist will discover if their estimations of ore grade are correct. However, in many cases, there ceases to be opportunity for validation, especially at the scale of an entire interpretation. Locally, interpretations may be validated, via tunnels and further surveys, etc.; but this feedback is not a part of a normal workflow and would be an exception to what typically occurs. Hence the expert's mental model is often poorly calibrated and not refined through validation feedback. Addressing this in a reasonable manner is a challenge. One solution is to take an iterative approach to geologic modeling in which a portion of the data is initially withheld from the interpreting experts. The resulting model is then compared to the withheld data. This provides the expert with an indication of the uncertainty in their own interpretations. The expert would then correct their model using the additional data to ensure that all data available have been utilized. Such approaches would also result in indications as to how the uncertainty behaves through statistical analyses (e.g., Lark et al., 2014) and may also be used to give a prediction of the uncertainty within a final model. This approach does have the major drawback of essentially requiring the expert to carry out two interpretations, though this could be mitigated by automating some aspect of the initial interpretation or carrying out this process on a small portion of the model and then interpreting the rest of the model using that knowledge. There is also evidence of anchoring bias in geological interpretation (Rankey and Mitchell, 2003), and such a workflow could be influenced by anchoring of experts to the initial data provided, thus resulting in a likely underestimation of uncertainty.

Elicitations as a Qualitative Tool

Aside from testing how appropriate elicitations are for uncertainty quantification, our results also give insight into how elicitations can be used as a qualitative tool to improve interpretations. In post-elicitation discussions, the experts mentioned that they appreciated the opportunity to discuss the geology with their peers, and that they felt that their understanding of the geology had improved as a result. This suggests that discussion of models and assumptions with peers is a valuable component in better understanding the potential uncertainties in models. This conclusion supports the findings of Polson and Curtis (2010), who highlight changes in the understanding of the presence of reservoirs, seals, and faults through the elicitation discussion process. The challenge is how to best use qualitative information from discussions to inform uncertainty. Here we have shown that expert elicitations were not useful for the prediction of uncertainty in the situations tested, but such discussion may reduce uncertainty in future interpretations. How information from discussions is used to inform uncertainty requires testing. With studies such as that of Polson and Curtis (2010) highlighting potential evidence of herding around a single expert, care is needed to understand the dynamics

of group discussion and decision-making based on group consensus. An experiment involving the interpretation of a cross section carried out before and after a discussion could be validated against real data (i.e., Lark et al., 2014; Randle et al., 2018). This would allow the quantification of any change in certainty resulting from the discussion.

CONCLUSION

We have shown through our experiments that expert elicitations do not result in accurate predictions of interpretation error; hence, they are not a suitable alternative to empirical approaches. However, they serve a purpose in that they offer opportunities for discussion that are typically not available in a standard modeling workflow. Our observations suggest that better geologic models and an understanding of the concepts and uncertainties that underpin them would result from the use of expert elicitations.

ACKNOWLEDGMENTS

We would like to thank all those who took part in our elicitations, as well as all those who helped in their facilitation. This work was undertaken while C.H. Randle held a joint University of Aberdeen, College of Physical Science Ph.D. Award and British Geological Survey University Funding Initiative (BUFI) Ph.D. Studentship at Aberdeen University, through Natural Environment Research Council (NERC). The contributions by C.H. Randle, R.M. Lark, and A.A. Monaghan are published with the permission of the Executive Director of BGS (NERC). The authors would like to thank Hazel Gibson and an anonymous reviewer for their comments on the manuscript and confirm that all views expressed are the opinions of the authors.

REFERENCES CITED

Bond, C.E., 2015, Uncertainty in structural interpretation: Lessons to be learnt: Journal of Structural Geology, v. 74, p. 185–200, https://doi.org/10.1016/j.jsg.2015.03.003.

Bond, C.E., Gibbs, A.D., Shipton, Z.K., and Jones, S., 2007, What do you think this is?: "Conceptual uncertainty" in geoscience interpretation: GSA Today, v. 17, p. 4–10, https://doi.org/10.1130/GSAT01711A.1.

Bond, C.E., Shipton, Z.K., Gibbs, A.D., and Jones, S., 2008, Structural models: Optimizing risk analysis by understanding conceptual uncertainty: First Break, v. 26, no. 6, p. 65–71, https://doi.org/10.3997/1365-2397.2008006.

Cooke, N.J., 1994, Varieties of knowledge elicitation techniques: International Journal of Human-Computer Studies, v. 41, no. 6, p. 801–849, https://doi.org/10.1006/ijhc.1994.1083.

Ellison, R.A., Woods, M.A., Allen, D.J., Forster, A., Pharaoh, T.C., and King, C., 2004, Geology of London: Special memoir for 1:50,000 geological sheets 256 (north London), 257 (Romford), 270 (south London), and 271 (Dartford) (England and Wales): Nottingham, UK, British Geological Survey, 114 p.

Hall, I.H.S., Browne, M.A.E., and Forsyth, I.H., 1998, Geology of the Glasgow District: Memoir of the British Geological Survey, Sheet 30E (Scotland).

Kessler, H., and Mathers, S.J., 2004, Maps to models—Finally capturing the geologists' vision: Geoscientist, v. 14, p. 4–6.

Lark, R.M., Mathers, S.J., Thorpe, S., Arkley, S.L.B., Morgan, D.J., and Lawrence, D.J.D., 2013, A statistical assessment of the uncertainty in a 3-D geological framework model: Proceedings of the Geologists' Association, v. 124, no. 6, p. 946–958, https://doi.org/10.1016/j.pgeola.2013.01.005.

Lark, R.M., Thorpe, S., Kessler, H., and Mathers, S.J., 2014, Interpretative modelling of a geological cross section from boreholes: Sources of uncertainty and their quantification: Solid Earth, v. 5, p. 1189–1203, https://doi.org/10.5194/se-5-1189-2014.

Lark, R.M., Lawley, R.S., Barron, A.J.M., Aldiss, D.T., Ambrose, K., Cooper, A.H., Lee, J.R., and Waters, C.N., 2015, Uncertainty in mapped geological boundaries held by a national geological survey: Eliciting the geologists' tacit error model: Solid Earth, v. 6, p. 727–745, https://doi.org/10.5194/se-6-727-2015.

Libarkin, J.C., Beilfuss, M., and Kurdziel, J.P., 2003, Research methodologies in science education: Mental models and cognition in education: Journal of Geoscience Education, v. 51, no. 1, p. 121–126, https://doi.org/10.1080/10899995.2003.12028056.

Mann, C.J., 1993, Uncertainty in geology, *in* Davis, J.C., and Herzfeld, U.C., eds., Computers in Geology—25 Years of Progress: Oxford, Oxford University Press, p. 241–254.

Oakley, J.E., and O'Hagan, A., 2010, SHELF: The Sheffield Elicitation Framework (version 3.0): Sheffield, UK, School of Mathematics and Statistics, University of Sheffield, http://tonyohagan.co.uk/shelf.

Polson, D., and Curtis, A., 2010, Dynamics of uncertainty in geological interpretation: Journal of the Geological Society of London, v. 167, p. 5–10, https://doi.org/10.1144/0016-76492009-055.

Price, S.J., Kessler, H., Burke, H.F., Hough, E., and Reeves, H.J., 2012, Model metadata report for Manchester and Salford, NW England: Open Report OR/12/068: Nottingham, UK, British Geological Survey.

Randle, C.H., Bond, C.E., Lark, R.M., and Monaghan, A.A., 2018, Can uncertainty in geological cross-section interpretations be quantified and predicted?: Geosphere, v. 14, no. 3, p. 1087–1100, https://doi.org/10.1130/GES01510.1.

Rankey, E.C., and Mitchell, J.C., 2003, That's why it's called interpretation: Impact of horizon uncertainty on seismic attribute analysis: The Leading Edge, v. 22, p. 820–828, https://doi.org/10.1190/1.1614152.

Reagan-Cirincione, P., 1994, Improving the Accuracy of Group Judgment: A Process Intervention Combining Group Facilitation, Social Judgment Analysis, and Information Technology: Organizational Behavior and Human Decision Processes, v. 58, no. 2, p. 246–270, https://doi.org/10.1006/obhd.1994.1036.