Chapter 15
Expert Judgement for Geological Hazards in New Zealand

Annemarie Christophersen and Matthew C. Gerstenberger

Abstract Expert judgement is important for the short- and long-term assessments of natural hazards in New Zealand, contributing to their risk analyses and informing decision-making. The problems are complex and usually require input from experts from different sub-disciplines. Expert judgement, like all human cognitive processes, is prone to biases. Therefore, we aim to use methods that are robust, transparent, reproducible and help reduce biases. The Classical Model treats expert opinion as scientific data and its performance-based weighting of experts allows us to measure the uncertainty of a quantifiable problem. We have developed a protocol for risk assessment, including structured expert judgement, which is centred around workshop-style interactions between experts to share knowledge. The protocol borrows heavily from the framework for the risk management process of the International Organization for Standardization. We outline seven recent applications of structured judgement, mostly in seismology and volcanology. Most of them use the Classical Model to aggregate the expert judgement. We discuss challenges and insights, concluding that developing an optimal protocol for expert judgement is a continuing journey.

15.1 Introduction

New Zealand lies in the south-west Pacific Ocean, along the junction between the Pacific and Australian tectonic plates (Fig. 15.1). The collision of the tectonic plates causes rugged mountains, active volcanoes and frequent earthquakes. Secondary geological hazards arise from landslides, tsunamis and flooding. A damaging earthquake can occur anywhere in New Zealand and a volcanic eruption can cause ash fall over most of the North Island. Given the small size of the country and the interdependencies of infrastructure, logistics and business, a major earthquake or volcanic eruption can affect the whole society. Assessing these hazards, either as immediate

A. Christophersen (✉) · M. C. Gerstenberger
GNS Science, 1 Fairway Drive, Avalon, New Zealand
e-mail: A.christophersen@gns.cri.nz

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_15

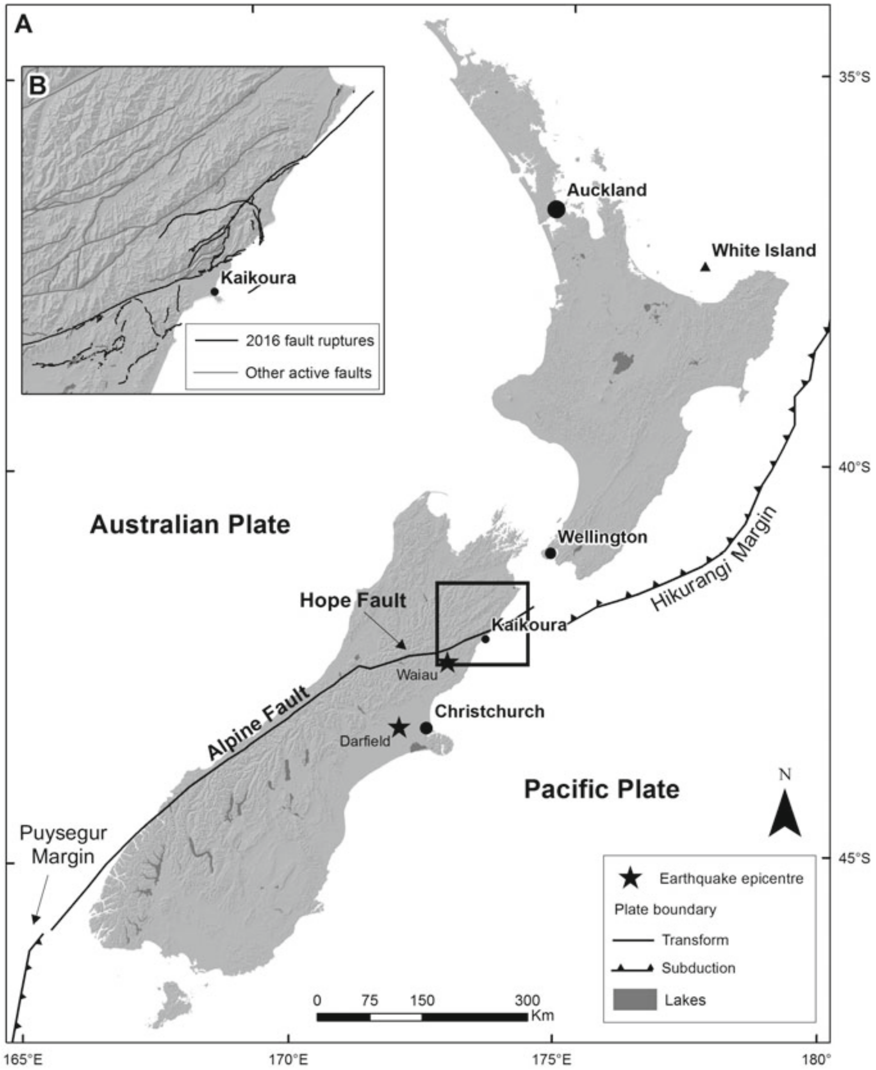

Fig. 15.1 Map of New Zealand; (A) showing the position on the plate boundary, with the Puysegur Margin in the south-west, the Alpine and Hope Faults in the South Island and the Hikurangi Margin in the east of the North Island. The stars indicate the locations of the two major earthquakes that initiated project 2–4 in Table 15.1. Also show is White Island volcano (project 6)

threats or in the long term, typically requires expert judgement; in part, this requirement is due to the low probability of major events and the limited data available for model building.

GNS Science advises the New Zealand government on geological hazards and contributes to the management of public information on geological hazards and associated emergencies (New Zealand Ministry of Civil Defence and Emergency Management 2015). It has similar functions to geological survey institutions in other countries. GNS Science manages the GeoNet system for the detection of earthquakes, land movement, volcanic activity and the potential for local-source tsunamis. GeoNet coordinates responses to natural hazard events.

Whenever the earth rumbles, rolls or fumes, scientists gather at the GeoNet offices to work out what has happened, is happening and might happen next. Scientists from different sub-disciplines share their data and knowledge to interpret what is going on. This informal expert judgement, for example, when complemented by rigorous statistical models for earthquake (Christophersen et al. 2017 for an overview), has been very effective in providing scientific advice to New Zealand government agencies, the media, public and other stakeholders. In contrast to understanding what is going on during an event response, long-term hazard models estimate the probability of occurrence of a specific hazard, in a specific future time period, as well as its intensity and area of impact. These models provide a basis for decision-making aimed at reducing the impacts of geological hazards to society. The development of long-term hazard models also involves elements of expert judgement.

Expert judgement, like most human thinking and judgement processes, is prone to biases that are often hidden from awareness (Bang and Frith 2017). Kahneman (2011), who jointly with Tversky pioneered the study of biases (Tversky and Kahneman 1974), describes the brain as consisting of two systems. System 1 is almost automatic and instinctive, while System 2 deals with rational thought and conscious decision-making. Working with System 2 requires energy and focus; this is mentally draining. The brain aims to preserve energy and preferably uses System 1 that takes many short-cuts, called heuristics, to process information and reach conclusions. Heuristics allow for faster processing of information but can cause biases and flawed decision-making.

For the development of robust geological hazard models and to be able to give the best possible scientific advice, we are interested in structured expert judgement (SEJ). The purpose of SEJ, as defined by Hanea et al. (2018), is to (1) address questions that theoretically could be measured or calculated if there was sufficient time and enough data, (2) follow reproducible and transparent rules, (3) anticipate and aim to mitigate biases, (4) be thoroughly documented and (5) provide opportunities for empirical evaluation and validation. Given the complexity of the problems that we address in geological hazards, we do not expect experts to reach consensus on any given question. Quite the contrary, we are keen to explore the uncertainty of a question of interest. In many cases, we need to estimate the likely occurrence of low-probability events. This makes it challenging to measure the success of any protocol and test for reproducibility. Therefore, we are looking for a method that has robust foundations and has been well scrutinized with evidence of skill in other applications.

The Classical Model treats expert judgement as scientific data and follows scientific principles from probability and statistics (Cooke 1991). It is built on rational consensus, in which experts agree on the method of aggregating individual judgement rather than seeking consensus on any specific problem. The method weights experts' judgement based on the experts' ability to estimate uncertainty for questions

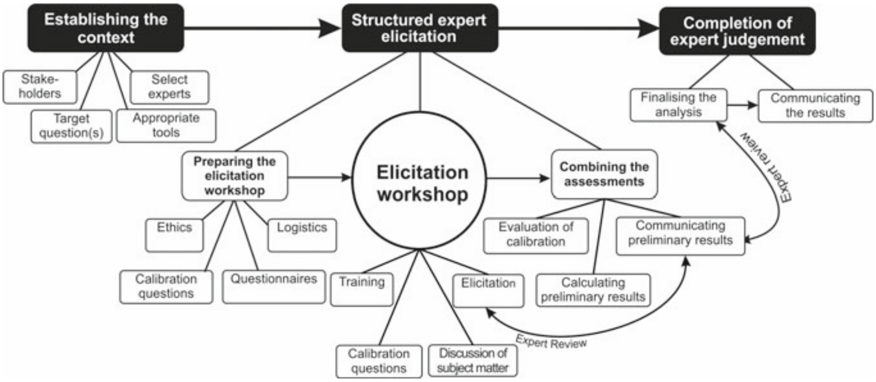
Establishing the context
Stake-holders
Select experts
Target question(s)
Appropriate tools
Structured expert elicitation
Preparing the elicitation workshop
Ethics
Logistics
Calibration questions
Questionnaires
Elicitation workshop
Training
Elicitation
Calibration questions
Discussion of subject matter
Combining the assessments
Evaluation of calibration
Communicating preliminary results
Calculating preliminary results
Expert Review
Completion of expert judgement
Finalising the analysis
Communicating the results
Expert review

Fig. 15.2 Our suggested protocol of a structured expert judgement with an elicitation workshop

with known answers, so-called calibration or seed questions (Cooke 1991; Quigley et al. 2018). The Classical Model suits our requirements well. We have developed a protocol for applying the Classical Model in workshop-style sessions for experts to share their knowledge and understanding of the problem so that they can best estimate the answer, including the uncertainty, to the problem at hand (Fig. 15.2).

In this chapter, we provide an overview of the biases that we try to mitigate. We introduce the protocol that we have used for multiple elicitations in the last few years, in which the Classical Model is ideally applied, and which is centred around workshop-style interactions. The main part of this chapter introduces seven recent examples of expert judgement applied to seismic and volcanological hazards. We discuss some of the challenges encountered as well as the benefits of using SEJ.

15.2 Developing a Protocol for SEJ

We began developing our procedures for SEJ within the context of risk assessment. Between 2010 and 2013, GNS Science led the development of risk assessment methods for CO2CRC (Gerstenberger et al. 2012). CO2CRC is Australia's leading carbon capture and storage research organization (CO2CRC 2011) and operates a study site in the onshore Otway Basin in south-western Victoria, Australia, for injection experiments (Jenkins et al. 2012). As part of the risk method development, we investigated Bayesian networks as tools for modelling complex problems (Gerstenberger et al. 2015) and explored SEJ methods for working with experts when data are unavailable or sparse. In Sect. 15.2.1, we provide an overview of common biases to be avoided, followed by a rational for the workshop-style expert interaction in 15.2.2, and a section on the Classical Model for assessing the risk and quantifying uncertainty in 15.2.3. Our expert judgement protocol is described in Sect. 15.3.

15.2.1 Common Biases

There is a large body of literature investigating biases, their causes and possible ways of mitigating them. Broadly speaking, biases fall into three categories with some overlap between them. Cognitive biases are mistakes in reasoning, evaluating, remembering, or other cognitive process. Motivational biases occur when the judgement is influenced by the expectation of the results and outcomes. Group biases may occur due to group dynamics. Montibeller and von Winterfeldt (2015) provide a recent review on cognitive and motivational biases and their mitigation in decision and risk analysis. More recently they have extended their analysis to include group biases (Montibeller and von Winterfeldt 2018).

The boundaries between different categories of individual biases are not always clear cut. For example, confirmation bias, "the seeking or interpreting of evidence in ways that is partial to existing beliefs, expectations, or a hypothesis at hand" (Nickerson 1998), is classified as motivational bias by Montibeller and von Winterfeldt (2015) while Kunda (1990) and Westen et al. (2006) discuss the cognitive aspects of confirmation bias. Nickerson (1998) discusses how confirmation bias results from not considering alternative hypotheses and that in turn can be associated with overestimating the accuracy of one's judgement. A narrow range of variation on estimated values (over-precision) is associated with overconfidence bias (Montibeller and von Winterfeldt 2015). Overconfidence bias is also used to describe the observation that people overestimate their own skill (overestimation) and that they believe they are better than others (over-placement). Over-precision, i.e. not appreciating the uncertainty of one's knowledge, is more prevailing than either overestimation or over-placement (Moore and Healy 2008), and is referred to as overconfidence in this chapter.

Anchoring is a bias that occurs when the assessment of a numerical value is based on an initial estimate and is not sufficiently adjusted to accommodate other information (Tversky and Kahneman 1974). This bias also applies when assessing confidence intervals and thus links with overconfidence. In short judgement can go wrong in many ways.

Montibeller and von Winterfeldt (2015, 2018) provide extensive lists of biases in the above-mentioned categories, and mitigation options. One bias missing from their compilation of cognitive biases is authority bias (Milgram 1963, 1974), which refers to the inclination to follow the lead of an authority figure. However, once the authority is challenged (by other group members or the facilitator, if in a workshop-style format), it is easier for individuals to disobey the authority (Milgram 1974). Groups can reinforce individual biases; in particular, if all experts view a problem from a similar perspective, flaws can be enhanced (Kerr and Tindale 2011). However, group processes can also have advantages in surmounting biases (Bang and Frith 2017).

Careful facilitation, good elicitation design and training of the experts can help to mitigate some cognitive biases. Motivational biases are challenging to mitigate in an individual. The best approach to achieving an unbiased final judgement is to include a number of experts with different viewpoints, challenge viewpoints in discussions and encourage alternative opinions. It is also useful to let experts provide their judgement confidentially to avoid peer pressure.

It is noteworthy that individuals generally only consider one hypothesis at a time and tend to assume that this hypothesis is true (Nickerson 1998). Consequently they look for evidence to confirm this hypothesis. Nickerson (1998) suggests that this form of confirmation bias can be mitigated by training experts to think of alternative hypotheses early in the elicitation process. This supports workshop-style sessions similar to our response to major earthquakes, where all streams of evidence, be it in the form of data or models, are presented and discussed prior to eliciting judgement.

15.2.2 Workshop-Style Expert Interaction

There are a number of advantages in group processes: they allow for the pooling of relevant information and for error checking, and can enhance individual task motivation (Kerr and Tindale 2011). Recent research confirms that groups tend to perform better than most individuals (Hemming et al. 2018). A recent literature review on common problems of decision-making in individuals and groups found that group processes have advantages in surmounting biases, exploring good models of the world and finding good solutions to problems (Bang and Frith 2017). In particular, discussions in small groups and without time pressure benefit from the knowledge held by individuals (Bang and Frith 2017 and references therein). This is consistent with our observations from the GeoNet-led earthquake responses, where experts from different sub-disciplines come together, unfortunately under time pressure, and share their knowledge to understand a complex problem. In workshop-style sessions, each expert represents the key findings from their sub-discipline. This allows for informed discussion and sharing of all relevant information. In such situations, experts can assess the arguments and form opinions. Research shows that individuals are more likely to change their mind for a well-argued opinion than for one stated with high confidence (Trouche et al. 2014). Other advantages of workshop-type interactions, going beyond accuracy of the final result, include that individual group members can voice their opinions, which helps to foster feelings of fairness/justice and inclusiveness, and increased legitimacy of and willingness to rely on the results.

Disadvantages of group interaction can be the pressure to conform to a majority view, the risk of being led astray by a dominant leader and the inattention to novel and unshared information (Kerr and Tindale 2011). The first two concerns may be mitigated by encouraging open discussion, in which the facilitator challenges dominant experts and thus makes it easier for the experts to disagree with the dominant person. Encouraging different viewpoints and exploring alternative hypotheses may also mitigate confirmation bias.

15.2.3 The Classical Model to Quantify Uncertainty

The Classical Model is a method for SEJ that mathematically aggregates expert judgements, based on the experts' ability to assess uncertainty. Experts provide their uncertainty for two types of questions: target questions and calibration questions. Target questions are the variables that cannot be adequately answered with other methods and thus require expert judgement. Calibration questions are similar in nature to the target questions and have values that are not known to the experts during the elicitation but become known during the analysis or are known to the analyst. Experts provide their uncertainty as percentiles, typically the fifth, fiftieth and ninety-fifth. Thus, they are asked for their best estimate and the 90% credible range for the true value to lie within. We tend to ask for an 80% credible range, i.e. for the tenth, fiftieth and ninetieth percentile, in an attempt to counterbalance the experts' overconfidence.

There are two measures to evaluate the experts' performance: statistical accuracy, also referred to as calibration, and informativeness (Cooke 1991). The statistical accuracy is the probability with which one would falsely reject the hypothesis that the experts answer according to the multinomial theoretical distribution determined by the inter-quantile intervals. Theoretically, calibration can take values between 0 and 1 but in practice they hardly ever get close to one, and most individual experts achieve a calibration below 0.05, see Chap. 10, this volume. Cooke (1991) defines a quantity that is based on how an expert estimates uncertainty over the number of calibration questions in relation to the percentiles of the credible range. For example, with the credible range of 80% mentioned above and ten calibration question, the true answer to the calibration question is expected to fall below the tenth percentile for one question, between the tenth and the fiftieth for four questions, between the fiftieth and the ninetieth for another four questions and above the ninetieth for one question. A transformation of this quantity is distributed like a chi-square random variable with three degrees of freedom. The calibration measures how this quantity diverges from the theoretical distribution. However, Chap. 10, this volume, illustrates that calibration does not clearly distinguish between well-calibrated experts. For example, two experts with nearly identical assessments on ten calibration questions can have a 0.44 difference in calibration score. On the other hand, experts, who are not well calibrated, can have a very low calibration score. Cooke (1991) argues that ten calibration questions and a significance level of 0.05 are sufficient to distinguish whether an expert is well calibrated or not.

The second measure of performance is informativeness. For example, an expert might provide very wide uncertainty intervals and by this potentially achieve good calibration but be not very informative. To calculate informativeness, an intrinsic range is determined for each calibration and target question. This covers the lowest and highest uncertainty estimates of all experts, and the true answer for each individual question plus an overshoot of each interval to capture the possible minimum and maximum of the interval. The informativeness of an expert is measured by comparing the estimated uncertainty widths with the intrinsic range and scaling the divergence using either a uniform or log-uniform distribution that covers the intrinsic range. Details and illustration of the methods are given by (Cooke 1991; Chap. 10, this volume; Quigley et al. 2018). Informativeness is a strictly positive function; the higher the score, the more informative an expert is. Typical values for informativeness can be found in the TU Delft expert judgement data base (Cooke and Goossens 2008). For 322 experts across the pre-2006 study the informativeness ranged from 0.25 to 3.81, with half of the experts scoring above 1.47, Chap. 10, this volume.

The experts' calibration and informativeness can be combined in different ways to derive weights to apply to the target questions (Cooke 1991). The combination of experts' weight is called the decision-maker. Different types of weights are available: global, itemized and optimized. Global weights average each expert's informativeness across all calibration questions. Raw weights are then calculated for each expert. Experts with a calibration score below a selected level of, for example 0.01, may be given a weight of zero, if a cut-off is chosen. The weights are then normalized across all experts with non-zero weights.

Itemized weights take advantage of the fact that informativeness for any expert can vary across questions while calibration is usually calculated over all calibration questions. Itemized weights are calculated for each question and each expert separately as the product of the informativeness on that question and the calibration score over all calibration questions. Again, experts with a very low calibration score may be given a weight of zero and therefore be excluded from the normalization of weights.

Optimized weights are calculated by varying the level of the calibration cut-off to maximize the score of the decision-maker. This may lead to some experts getting zero weights. However, zero weight does not mean zero value because all experts contribute to the intrinsic range. Figure 10.12 in Chap. 10, this volume, gives an example, in which the optimized decision-maker uses only two of ten experts, but the exclusion of one particular zero-weighted expert would lead to a significant reduction in the performance of the decision-maker.

The weighted combination of the experts' judgements is applied to the calibration and the target questions. This way, the Classical Model validates both individual expert assessments and the performance-based combinations against observed data.

As further discussed in Sect. 15.3.2.2, we usually administer the calibration questions in the early stages of the workshop to be able to show the initial results to experts before they finalize their answers to the target questions. This is against the standard recommendations to make the calibration questions as indistinguishable from the target questions as possible to be unbiased performance measures (Cooke 1991; Quigley et al. 2018). However, there are two advantages in showing experts the calibration results. While individual experts tend to be overconfident, i.e. they provide too narrow uncertainty intervals and therefore are not well calibrated, the decision-maker tends to find the true value of the calibration question. Seeing that the decision-maker of the Classical Model finds the answers that the individuals struggled with builds confidence in the method. Secondly, as a consequence of realizing their own overconfidence, we find that experts widen their confidence intervals when answering the target questions. This way we are likely to better measure the uncertainty of the target questions, because the experts have learned to counter-bias their overconfidence. On the down-side, the performance on the calibration questions may not then be a true reflection of the performance on the target question(s).

15.3 A Risk-Based Protocol

The International Organization for Standardization's principles on risk management (ISO 2009) provides a useful framework to adapt to an expert elicitation protocol. The risk management process has three main components: (1) establishing the context, (2) risk assessment and (3) risk treatment. "Communication and consultation" and "monitoring and review" inform each step of the process. The risk assessment is split into the sub-components of risk identification, risk analysis and risk evaluation. We have modified the ISO framework for risk assessment in carbon capture and storage (Gerstenberger and Christophersen, 2016, project 1, Table 15.1) and volcanic eruption forecasting (Christophersen et al. 2018, project 6, Table 15.1). Here we adapt the same framework to a protocol for structured expert judgement (Fig. 15.2). There

Table 15.1 An overview of recent expert elicitations, the methods used and the roles of the authors. MG stands for Matt Gerstenberger and AC for Annemarie Christophersen

	Project	Method(s) used	Roles
1	Risk assessment in carbon, capture and storage	Classical Model in workshop-style setting	Project leader, workshop facilitator, analyst (MG) coordinator of calibration questions, analyst (AC)
2	Time-dependent seismic hazard model for the recovery of Christchurch 2a source model 2b GMPE model	Classical Model in workshop-style setting	Project leader, workshop facilitator, analyst, coordinator of calibration questions (MG) Contributor to calibration questions (AC)
3	Probability of large earthquake following Kaikōura earthquake	Informal elicitation of probabilities and uncertainties in workshop-style setting	Project leader, facilitator, analyst and expert (MG) and expert (AC)
4	Probability of large earthquake following Kaikōura earthquake	Classical Model in workshop-style setting	Project leader, facilitator, analyst (MG), coordinator of calibration questions, analyst (AC)
5	Australian national seismic hazard model 5a source model 5b GMPE model	Classical Model in workshop-style setting	Facilitator, coordinator of calibration questions, analyst (MG); contributor to calibration questions (AC)
6	Development of eruption forecasting tool	Individual probability estimates in workshop-style setting	Project leader, workshop facilitator, analyst (AC)
7	National-level long-term eruption forecasts	Classical Model in workshop-style setting	Control expert (AC)

are three main components: establishing the context is the starting point as in the ISO framework; however, risk assessment is replaced with structured expert elicitation and the risk treatment with the completion of the structured expert judgement. We describe the three different components and their building blocks below.

15.3.1 Establishing the Context

Establishing the context includes four main components: (1) identifying the stakeholders and their roles, (2) defining the target question(s), (3) selecting appropriate tools and (4) selecting the experts.

Stakeholders can include a wide range of people, who may or may not be involved directly with the elicitation. For geological hazards in New Zealand, the public are also stakeholders and are usually informed about the outcome. There are several roles within an SEJ project; the problem owner, the coordinator, the facilitator and the analyst (e.g. Hemming et al. 2017). The problem owner is often the person who initiated the elicitation, or, who has been delegated the task of being responsible for the SEJ. The coordinator manages the process, including time lines and collection of responses. The facilitator handles the interactions between experts and needs to be diplomatic, and in our case, able to facilitate group processes with a wide range of different personalities. The facilitator needs to be aware of biases and how to mitigate them. The role requires a good understanding of the problem to be addressed and neutrality with respect to the outcome. The analyst is responsible for processing and analysing the responses and providing feedback to the experts. Applying the Classical Model further requires someone to coordinate the calibration questions. Depending on the scope of the project, the roles can be undertaken by one person, if no conflict of interest exists, or shared by many.

The target question(s) need(s) to be unambiguous, clear and well defined. For example, when asking for the probability of a large earthquake in central New Zealand, it is important to define the magnitude threshold, the region and the timeframe. Experts might want to know whether the earthquake has to be nucleating within the defined region or whether an earthquake that occurs at the boundary of the region and only partially within the region is seen as occurring within the region. It is helpful to write down the target question(s) early in the process and get feedback from various stakeholders whether the problem is appropriately addressed by the target question(s). We find that in discussions with experts during the elicitation workshop that the target question(s) may be slightly modified for clarity.

Appropriate tools include any material, methods or models that can help the experts evaluate the problem. For the risk assessment in carbon, capture and storage (Sect. 15.4.1), the tool was a Bayesian network model. For the time-dependent seismic hazard model for the recovery of Christchurch (Sect. 15.4.2), the tool was the hazard model, the various earthquake forecast models and the ground motion prediction equations. Appropriate tools can include all the background information that can be useful for the experts to make their assessment. It may take some time to prepare the material for the elicitation process and to decide on the most appropriate method of presenting the material.

Selecting appropriate experts is a key component of any SEJ. Good judgement does not depend only on substantive expertise, i.e. knowledge of the domain in question but also on the ability to adapt one's knowledge to novel events, and the ability to communicate one's knowledge and the limitations of one's knowledge in terms of quantiles and probabilities (Hemming et al. 2018). Traditionally, an expert has been defined by qualification, track record and experiences. More experienced experts have been expected to give better advice (Burgman et al. 2011). However, expert status defined by the citations (Cooke et al. 2008) or ranking on an 11-point scale (0 = 'no expertise', 5 = 'moderate expertise', 10 = 'highly expert') in the areas of training, professional experience and current role (Burgman et al. 2011) are a poor guide to actual performance. For geological hazards, we usually select a combination of experts with local and site-specific knowledge and general experts with subject-related experience from elsewhere. These are usually the primary drivers to illuminate the problem. In addition we include challengers, who are related domain experts, who can bring a different perspective to addressing the target question(s), and overall questioners, who also have specific sub-discipline knowledge but can look at the overall system and ask big picture questions. The use of students or early career scientists, who start the process without already having an answer and therefore have the ability to take in information from all sources and draw informed conclusions, can also help to minimize bias (Gerstenberger and Christophersen 2016 and references therein). We refer to these experts as "assimilators". For workshop-style sessions with the experts, we find that eight to 15 experts is a good number and allows for the different expert types to be included, as well as for free discussion with a manageable facilitation burden.

15.3.2 Structured Expert Elicitation

A well-facilitated elicitation workshop is central to our protocol. The workshop needs to be well prepared, including considering ethics requirements, preparing the questionnaires for the calibration and target questions and testing their utility by having colleagues and/or other stakeholders, who are not involved in the elicitation workshop, to answer them ahead of time. The preparation also includes logistics, such as travel arrangements, arranging a meeting facility, catering and preparing all necessary material for the workshop. The elicitation workshop itself includes several elements such as training, calibration questions, discussion of the subject matter and of course the elicitation itself. We consider the combining of the assessments to be part of the structured expert elicitation, and again there are several components including the processing of the questionnaires, evaluation of the calibration and communication of the preliminary results for the experts to review and provide feedback on. In the following we describe each component in more detail.

15.3.2.1 Preparing the Elicitation Workshop

Human ethics approval is required for all research conducted about people. For geological hazards the subject is usually the earth and working with experts does not necessarily require an ethics procedure. It is still important to follow ethical principles such as respecting people, minimizing harm to participants and researchers, ensuring informed and voluntary consent to participate in the research, respecting privacy and confidentiality, avoiding conflict of interest and being socially and culturally sensitive. Research that asks experts about their personal experiences and their thoughts will require an ethics procedure to ensure the research does not cause harm to the participants. Procedures for human ethics approval vary in different countries and local practices will need to be followed.

Calibration questions are central to the Classical Model to allow for performance-based weighting of the experts' judgement. Ideally the calibration questions are close in nature to the target questions so that the experts use similar thinking processes and so that performance on the calibration questions is relevant for the target question(s). Calibration questions have been classified into predictions, where answers are not known during the time of the elicitation but will become known during the analysis, and retrodictions, where the answers are known already but not to the expert during the elicitation (Cooke and Goossens 1999). Calibration questions are further distinguished by whether they are from within the domain of the target question or from an adjacent domain. Domain predictions are ideal, followed by domain retrodictions or adjacent predictions; less ideal is adjacent retrodictions (Cooke and Goossens 1999; Quigley et al. 2018).

Finding suitable calibration questions is not an aspect of the Classical Model that is widely discussed in the literature, even though it can be challenging, in particular when the target questions are small probabilities or parameters for models. Recent work proposes some strategies to finding suitable calibration questions (Quigley et al. 2018). Among them are using results from future measurements that are performed before the analysis is complete; unpublished measurements and mining data for relevant but unusual features. Given that we elicit the calibration question in a workshop-style setting, where experts do not have access to the Internet or their computers, we can use published data, as well as data sets that the experts are very familiar with but cannot access and query at the time. Questions about the experts' own datasets can be useful to highlight the overconfidence bias; experts tend to think that they know simple summary statistics much better, particularly from their own data, than they can recall. Seeing the results of the calibration questions and realizing that the true values are often outside their confidence ranges, gives experts a whole new appreciation of the limitation of their knowledge and consequently experts tend to increase their uncertainty bounds. It can be useful to work with colleagues of the experts to identify calibration questions under the premise that the questions will be kept confidential until after the elicitation. For our recent applications of SEJ in geological hazards in New Zealand (Sect. 15.4), our target questions were mostly about probabilities, weighting models or conditional probabilities for discrete Bayesian network models (Table 15.2). While we could find calibration questions in the same

Table 15.2 Details on the methods for the recent expert elicitations listed in Table 15.1

Project	Number of experts	Number of calibration questions	Number of target questions	Type of target questions	Workshop duration	Time for experts to review their estimates	Aggregation
1	10	10	335	Conditional probabilities	2 half-days	About 1 month	Classical model
2a	12	14	14	Weights of models	3 days	2 weeks	Classical model
2b	5	11	12	Weights of models	1 day	Only on the day	Classical model
3	11	None	1	Probability	2 h	2 days	Average weights
4	14	16	4	Probabilities	2 days	Extra time available but not taken	Classical model
5a	15	17	84	Weights of models	1 day	Only on the day	Classical model
5b	10	16	77	Weights of models	1 day	Only on the day	Classical Model
6	11	None		Conditional probabilities	2 half-days	1 week	Average weights
7	28	24	133	Probabilities, time to eruption, durations of next eruption, vent location	1 day	A couple of months	Classical model

domain, they were not of the same nature as the target questions. In such cases, there is always an element of doubt about whether the expert performance on the calibration questions is valid for the target questions. We aim to include more calibration questions than the recommended number of eight to ten for eliciting three quantiles (Cooke 1991) to be able to test the sensitivity of the performance weights to including different calibration questions. However, the number of calibration questions needs to be balanced with the time required for the experts to answer them and the mental energy required that takes the focus away from the target question(s).

Preparing the elicitation workshop also includes preparing questionnaires for both the calibration and target questions. One aspect of this is the wording of the questions to remove any ambiguities and make them as clear as possible. Another aspect is what medium to use. Paper and pencil seem to work best in workshop-style settings, so that experts can scribble notes at the sides. If the target questions are many conditional probabilities such as in examples 15.4.1 and 15.4.2, it is useful to collect the data in an electronic format, such as a spreadsheet or an online questionnaire. Having the answers electronically circumvents tedious data entry and possible challenges in deciphering handwritten notes. We aim to make the process as convenient as possible for the experts and sometimes offer different options for providing the answers. If we use the Classical Model to aggregate the expert judgement, we include some basic information about the method on the questionnaire. The information contains a figure of a Gaussian distribution function with the percentiles that we elicit marked and the expected location of the answers to the seed questions with respect to the percentiles.

The logistics of the workshop depend on where the workshop is held and include organizing travel, a suitable venue, and catering, including meeting special dietary requirements, to ensure that the experts feel well taken care of and can concentrate on the elicitation exercise. It is also important to plan breaks and opportunities to refresh, to avoid fatigue and allow System 2 of the brain to be engaged.

15.3.2.2 Elicitation Workshop

The elicitation workshop has four important components: training, administrating the calibration questions, discussion of the subject matter, and the elicitation.

The training aims to make experts aware of biases and encourages them to question their knowledge and to facilitate thorough estimates of uncertainty. It includes an introduction to the Classical Model to explain the method and illustrate the question format. It is useful to discuss one or more calibration questions in detail to demonstrate how to think about the percentiles that are elicited. It is good practice to encourage experts to think about the extreme values first to counterbalance the anchoring bias. Administering the calibration questions within the first part of the workshop, following the initial introduction and training, allows for an analyst to process the results and to show them to the experts during the workshop. We find that, despite training, experts are overconfident in their knowledge. Once they have gone through an hour of answering calibration questions and have been presented with the results, they appreciate their overconfidence and tend to widen their confidence intervals. We have not yet mixed further calibration questions in with the target question to formally test this general observation. We are aware that this might influence and change the way the experts answer the target questions compared to the calibration questions. Therefore, the calibration questions become less relevant for performance weighting but are more important as a training tool for estimating uncertainties, as further discussed in Sect. 15.5.1.

Showing the results of the calibration question during the workshop demonstrates that the combined results (see Sect. 15.3.2.3) usually find the correct answer for the calibration questions, despite most individual experts being overconfident in their knowledge. This observation builds the experts' confidence in the method. Giving experts immediate and definitive experience in answering challenging and complex questions of similar type to the target questions also builds the confidence in their own capability. We find that it is not uncommon for some experts to initially think they are unable to estimate any useful answer (despite their inherent overconfidence in any single answer). The feedback on the calibration questions tends to alley these initial concerns.

While most experts seem to enjoy the learning opportunity provided in the way we administer and discuss calibration questions, some experts feel a bit apprehensive and put on the spot, similar to taking an undesired examination. This apprehension seems to be of particular concern when the facilitator and/or analysts are close colleagues. We aim to process the questionnaires so that individual answers, also for the target question, are not even known to the analyst and to ensure confidentiality of individual estimates. This cannot be fully guaranteed because sometimes handwritten or illogical responses need to be confirmed with the experts.

Most of the time during the workshop is spent on discussing the subject matter. This usually includes presentations by domain experts with plenty of time for questions and discussion. The presentation of material requires careful facilitation to avoid anchoring. We encourage experts to think broadly and to consider what might be missing from the presented material and how they can account for unknowns in their uncertainties.

The elicitation of the target questions begins during the workshop. We generally hand out the questionnaires with the target questions before discussions on the subject matter starts, so that experts have the target questions in front of them and can take relevant notes during the discussion. Experts fill in their questionnaire individually, usually within the room. If the target questions fall into different topic areas, we discuss the particular topic area and ensure everyone has the same understanding of the questions being asked, and then allow time for experts to fill in their estimates without interruption. If, during the discussion there are any dominant views, the facilitators try to challenge them by making counterarguments so that experts do not fall for dominance bias. The facilitators encourage experts throughout to answer question to the best of their own knowledge and understanding, and to consider the limits of the knowledge and how best to reflect that in their uncertainty estimates.

Experts usually have extra time beyond the completion of the workshop to review and finalize their answers to the target questions, as indicated in Table 15.2 for our different example applications.

15.3.2.3 Combining the Assessments

Combining the assessments includes a more thorough evaluation of the calibration questions than during the elicitation workshop, processing of the questionnaires, calculating the preliminary results and communicating these to the experts for review.

For the evaluation of the calibration questions, according to the Classical Model, there is software called Excalibur (Cooke and Solomantine 1992), which is freely available (Lighttwist Software 2008) and runs on a Windows operating system. Notes on expert elicitation with Excalibur and a tutorial are also available online (Aspinall 2008; Colson 2016).

For the initial analysis of the calibration questions we usually use global weights without optimization (see Sect. 15.2.3). There are two parameters that can modify the decision-maker and the weights between experts; these are the calibration power and the intrinsic range. The calibration power allows us to compare the calibration of experts between studies with different numbers of calibration questions, and is defined as the ratio of the number of calibration questions used in two different studies, see Chap. 10, this volume. It can vary between 0.1 and 1, with 1 for the studies having the same number of calibration question and 0.1 for one having ten times as many questions as the other. A calibration power of 0.5 reduces the resolution of the significance test to that of one with half as many questions. In practice, reducing the calibration power distributes the weights more equally between experts. The recommendation is to only use a power of less than 1 if all experts have calibration scores less than 0.05 and to avoid giving all the weight to one badly calibrated expert, see Chap. 10, this volume. We often reduce the calibration power because we want to equalize the weights between experts. In our first application (Sect. 15.4.2), one expert got nearly all the weight. However, there was doubt about the calibration questions perfectly representing the target questions. We have reduced the calibration power in subsequent applications with similar motivation, while ensuring not to reduce the overall performance of the decision-maker.

The intrinsic range defines by how much the support of the variable is extended beyond the minimum and maximum percentile of any experts (see Sect. 15.2.3). In Excalibur this value can vary from 0.01 to 100, where 0.10 is the default and corresponds to 10% extension of the overall range on either side. The intrinsic range is important for determining informativeness. A larger support will result in higher informativeness of experts whose quantiles are more widely spread.

The processing of the questionnaire depends on the extent of the target questions; if only a small number of variables are elicited this can be fast and straight-forward. If model parameters or model weights are elicited this might involve lengthy calculations. For large numbers of target questions having the experts fill in their answers in some electronic form can help to reduce the burden of data processing.

We always communicate the initial results to the experts for them to provide feedback on the outcome (Fig. 15.2). This is particularly important when eliciting model parameters and/or model weights. The overall result can be surprising and counter-intuitive. We want to hear experts' thoughts on the overall results. There may be a possibility that experts use this opportunity to sway results in a way they would like to see them go (motivational bias). However, in our applications, we have not observed any evidence for this.

15.3.3 Completion of the Expert Judgement

Completion of the expert judgement involves finalizing the analysis and communicating the results. The final results take experts' feedback into account. For geological hazards, it is important to communicate the results to a wide range of stakeholders, including the public. GNS Science has a social science team that conducts research into how messages are best understood and communicated to reach the relevant stakeholders.

15.4 Application of SEJ for Natural Hazards in New Zealand

Here we introduce seven recent projects that include elements of expert judgement. Table 15.1 provides an overview of the projects, the methods used and our roles. Two of the projects do not strictly fit within the umbrella of geological hazards in New Zealand. Project 1 is about a Bayesian network model for the detection of injected CO_2 in a saline aquifer and sums up the development of our risk assessment method for carbon, capture and storage that led us to introduce structured expert judgement and Bayesian network modelling to geological hazards projects. Project 5 is about the recent update of the Australian national seismic model, which is exemplary for involving the wider research community in seismic hazard assessment.

15.4.1 Risk Assessment in Carbon, Capture and Storage

The Otway Stage 2C project of the CO2CRC involved a test injection of 15,000 tons of supercritical gas mixture at the CO2CRC Otway site in the Australian state of Victoria. The objective was to examine the limits of detecting the gas plume with seismic surveying on the surface and to conduct detailed pressure monitoring of the injection (Pevzner et al. 2015). The risk register for the Otway injection site identified the risk of not being able to detect the injection plume with seismic surveying and not being able to demonstrate stabilization of the plume. We had the opportunity to apply the risk assessment method that we had developed during our CO2CRC involvement, in particular Bayesian networks and structured expert elicitation with the Classical Model, to address these risks. The development of the Bayesian network model structure was an informal and iterative process through remote interaction between GNS staff and CO2CRC. The conditional probabilities for the Bayesian network were elicited in a workshop over two half-days in March 2013, in an SEJ process including the application of the Classical Model. We had the opportunity to investigate possible calibration questions ahead of time (Christophersen et al. 2011). Since we administered the calibration questions during the face-to-face workshop, we could ask questions from the published and grey literature as well as about specific data from the Otway basin. Asking the experts about their own data was particularly useful to understand overconfidence. Experts were critical about the calibration questions during the workshop. One expert questioned the quality of the work chosen from the grey literature and was encouraged to consider that in the uncertainty estimate. The critique allowed for a solid discussion on the purpose of the calibration questions.

The result of the Bayesian network was a 74% probability of detecting the plume, and a 57% probability that there will be consistency between the model-predicted plume behaviour and the observations. The plume detection has been successful (Pevzner et al. 2017).

15.4.2 A Time-Dependent Seismic Hazard Model for the Recovery of Christchurch

The New Zealand National Seismic Hazard model (NSHM; Stirling et al. 2012) estimates earthquake ground shaking and forms the basis for structural design in New Zealand. The NSHM applies the well-established practice of probabilistic seismic hazard analysis, which has three key components: the fault source model, the distributed source model and ground motion prediction equations. The NSHM is regularly updated to include the latest science.

The Canterbury earthquake sequence increased the rate of seismicity in the Canterbury region well above the long-term rates and the seismicity is expected to stay elevated for years, if not decades (Gerstenberger et al. 2014, 2016). The elevated seismicity warranted the development of a new time-varying seismic hazard model for the Canterbury region because the NSHM was expected to underestimate the seismic hazard due to ongoing aftershocks and the possibility of further triggered earthquakes. The new seismic hazard model has the same components as the NSHM: a fault source model, a distributed source model and ground motion prediction equations. The fault model was extended from the 2010 NSHM update but was not subject to SEJ. The distributed source model is the dominant contributor in this case and is a combination of earthquake-clustering models of three timescales (short-term, medium-term and long-term). Weights for the models were elicited in a two-day workshop including the application of the Classical Model.

The ground motion prediction equation component of the model was extended to include a new Christchurch-specific model (Bradley 2010, 2013). A one-day workshop was held to elicit the necessary parameters and weights for the ground motion prediction equations, again including the application of the Classical Model.

The resulting hazard model represents the seismic hazard for the Canterbury region for the next 50 years. The model has been used to provide earthquake probabilities to a range of end users on timescales from 1 day to 50 years. The 50-year hazard forecast has informed the revision of the New Zealand building design guidelines and other aspects of the rebuilt of Christchurch.

15.4.3 Informal Elicitation of the Probability of Large Earthquakes in Central New Zealand Impacted by Slow Slip and the Kaikōura Earthquake

The 14 November 2016 Kaikōura earthquake with magnitude M = 7.8 triggered wide-spread silent and slow movement along the plate boundary, also called slow slip events (SSE); these events can take weeks to months to occur but are not felt by people. By 25 November, observations from global positioning system (GPS) stations indicated that three regional SSE were occurring. While SSE in these regions have been observed numerous times in the past 20 years, they had never occurred simultaneously before and one of them appeared to have a larger slip rate than previously observed. These observations raised concerns about the impact of the SSE on future earthquake occurrence. On 25 November, the New Zealand Ministry of Civil Defense and Emergency Management (MCDEM) was briefed about the concerns, and consequently expected formal advice from GNS Science on the likelihood of future M ≥ 7.8 events in central New Zealand, including any potential impact of the ongoing SSE on this likelihood.

While GNS Science has provided earthquake forecasts in response to large earthquakes since the September 2010 M = 7.1 Darfield earthquake (Christophersen et al. 2017), no earthquake forecasting model implicitly considers SSE. To fulfil MCDEM's expectation, GNS Science used expert elicitation. We had about a week to pull together different strands of evidence including the forecasts from the statistical model, results from synthetic earthquake data (Robinson et al. 2011) and the NSHM (Stirling et al. 2012). We analysed the effect of SSE on seismicity, calculated Coulomb stresses and consulted with international experts (Gerstenberger et al. 2017). It was not possible to develop subject-appropriate calibration questions within that short time period and with an active response to the mainshock still ongoing. On 1 December 2016, we held a two-hour workshop with 11 New Zealand experts, who were mostly GNS Science staff. We presented and discussed all information available at that time. Experts then individually estimated the probability of an M ≥ 7.8 events in central New Zealand within the next year. Everyone provided their best estimate and a 90% confidence interval. The results were communicated to MCDEM and to the public via the GeoNet website.

15.4.4 SEJ and the Classical Model to Assess the Probability of Large Earthquakes in Central New Zealand Impacted by Slow Slip

In the year following the Kaikōura earthquake, GNS Science conducted further research on the effect of SSE on earthquakes (Kaneko et al. 2018; Wallace et al. 2017) and continued to consult with international colleagues two workshops were held, including an initial one at the annual meeting of the Southern California Earthquake Center, in California to discuss initial model developments. Subsequently, we conducted a second SEJ on the one-year anniversary of the Kaikōura earthquake to estimate the probability of large earthquakes in central New Zealand within the subsequent one and ten years. The second elicitation workshop was held over two days at GNS Science and was attended by fourteen experts from four different countries and nine different organizations. We applied the Classical Model with calibration questions that were again derived from the published literature and relevant publicly available data sets that the experts could not access during the workshop.

The most striking observation when comparing the results of the expert elicitation in December 2016 and November 2017 is an increase of the uncertainty estimates in 2017, even though the 2016 estimates were 90% confidence intervals versus 80% in 2017. Although a direct comparison is difficult, this observation is consistent with our expectation that through training and a much more thorough process the experts increase their uncertainty once they have seen the results from the calibration question. It also seems that the experts' answers were more anchored on the results from the statistical model in the 2016 December when experts had not gone through the SEJ process.

15.4.5 Australian National Seismic Hazard Model

Geoscience Australia is an agency of the Australian government and is responsible for the Australian national seismic hazard model. In the 2018 update of the model, NSHA18, Geoscience Australia undertook a new, and so far unique for seismic hazard, approach: it invited the Australian earthquake hazard community to submit peer-reviewed seismic source and ground motion models for consideration (Allen et al. 2018; Griffin et al. 2018). This resulted in 16 seismic source models and 20 ground motion models being proposed and contributing to NSHA18, demonstrating the range of expert opinions on characterizing seismic hazard for a low seismicity region like Australia. Following similar methods as described above for the Canterbury hazard model, Geoscience Australia held two expert elicitation workshops in March 2017 to weight different seismic source models and ground motion models. The workshop applied the Classical Model and GNS Science assisted with the calibration questions and workshop facilitation. The 17 workshop participants represented the collective expertise of the Australian earthquake hazard community. Feedback from the workshop participants was positive, with experts reporting being challenged by, but enjoying, the calibration and elicitation process.

The NSHA18 yields much lower hazard estimates than previous assessments (Allen et al. 2018). This is due to a number of factors, including the revision of earthquake magnitudes and the use of more modern ground motion models than previously available. Given tight timelines, there was no chance for the experts to review their contribution once the hazard was calculated. For future studies, Geoscience Australia recommends to re-engage with the experts to allow them to review and reassess their choices, despite concerns that experts may be motivated to tweak answers to move results closer to their expectation (Allen et al. 2018). Such a review process would be consistent with our protocol (Fig. 15.2).

15.4.6 Development of an Eruption Forecasting Tool

Volcanic eruptions are usually preceded by a period of unrest, during which small earthquakes occur around the volcano; the volcano can emit increasing amounts of gas, and ground deformation may be observed. GeoNet coordinates the volcano monitoring team that consists of GNS staff based at three sites. The team meets regularly (partly remotely) to review the status of all 12 monitored New Zealand volcanic centres. It sets the Volcano Alert Levels (Potter et al. 2014) and the Colour Codes of the International Civil Aviation Organization and regularly estimates the probability of forthcoming eruptions for internal health and safety policy requirements (Deligne et al. 2018; Jolly et al. 2014). In recent years, there have been small volcanic eruptions, including the fatal December 2019 Whakaari/White Island eruption. New Zealand has the potential for much more disruptive volcanic eruption.

There are limited quantitative tools in eruption forecasting (Sparks et al. 2012) that can help the volcano monitoring team to assess the probability of upcoming eruptions. Given the success of Bayesian networks in the CO2CRC-project to model complex problems, we proposed to trial Bayesian networks as decision-support tool in volcano monitoring (Christophersen et al. 2018). We started with a small team with wide-spread expertise. In an informal process, the team adapted a published Bayesian network model for eruption forecasting (Hincks et al. 2014), which was reviewed by some members of the volcano monitoring team. In a structured process, we elicited the conditional probabilities for the Bayesian network in a workshop over two half-days in early December 2015. The workshop included a presentation on the Classical Model and some example calibration questions to introduce the method. We did not have the time and resources to develop appropriate calibration question for the conditional probabilities of the Bayesian network. Given the previous experience with experts' unease about the calibration questions, we decided against using the Classical Model so as to not distract from the main objective of exploring the potential use of Bayesian networks in volcano monitoring and eruption forecasting. In feedback questionnaires, the experts were supportive of applying the Classical Model in future elicitations. The finding of the project was that Bayesian networks are promising tools for volcano monitoring with many recommendations for future work, mainly focussing on developing Bayesian networks with continuous variables and exploring dynamic Bayesian networks but also including SEJ for parameterising the model.

15.4.7 National-Level Long-Term Eruption Forecasts

Volcanoes cause many different hazards, including ash fall, pyroclastic density flows, lava flow and lahars. These hazards can impact near and far from the volcano, before, during and after an eruption (National Academies of Sciences 2017). Many volcanic hazards depend on the weather conditions like wind direction and rain, the presence of snow and ice, and the local topography (Stirling et al. 2017). Thus, the development of a comprehensive volcano hazard model is a complex task. The first step involves quantifying the frequency, size and location of eruptions for each volcano. A recent project led by Massey University with broad collaboration across other New Zealand organizations including GNS Science, conducted an SEJ to estimate the timing and sizes of the next eruption for 12 volcanoes (Bebbington et al. 2018). A total of 28 experts including volcanologists, statisticians, and hazards scientists, provided estimates that were combined using the Classical Model to arrive at hazard estimates. The same experts contributed to an informal expert elicitation to outline the next steps for developing a national probabilistic volcanic hazard model for New Zealand (Stirling et al. 2017). Given the wealth of material to elicit, the discussion during the workshop was kept relatively short. There was ample opportunity for experts to revise their answers. The results and challenges of the study have been well documented (Bebbington et al. 2018).

15.5 Discussion and Conclusion

There are many applications for expert judgement in geological hazards in New Zealand. We have introduced seven recent applications that we have been involved within different roles. The problems are often complex and require input from multiple sub-disciplines. Being aware of the human brain's preference to take shortcuts, potentially causing biases, we are interested in robust expert elicitation protocols that minimize biases and quantify uncertainty. We have introduced a protocol for expert judgement that is based on risk assessment methods and has workshop-style interaction at its heart, so that experts can share all evidence and reach a good understanding of the problem. The Classical Model is well suited to explore the uncertainty around the complex issues that we are addressing. Here we discuss how our application of the Classical Model differs from the standard recommendations, some of the challenges we have encountered, and the benefits of our protocol.

15.5.1 Tweaks in Applying the Classical Model

The way that we apply the Classical Model differs in two ways from the standard recommendations (e.g. Cooke 1991; Quigley et al. 2018). Firstly, we clearly set apart the calibration questions from the target questions and use them as a training tool to improve uncertainty estimates for the target questions. As a consequence, experts may potentially have a different philosophy when assessing confidence bounds for target and calibration questions. Ideally, experts become better in assessing uncertainty, which may mean they are more likely to increase their uncertainty bounds. Alternatively, because the calibration exercise is separate, it allows experts to reduce their uncertainty bounds on the target questions to obtain desired results. Thus, if expert behaviour is inconsistent between the two question sets, our approach may reduce the value of the calibration questions for absolute performance weighting. However, in our opinion the potential for improved quantification of uncertainty through training outweighs the potential for inconsistent expert behaviour between calibration and target questions.

Secondly, we reduce the power of the calibration when aggregating the expert judgements. Using full power when determining the weights assumes the list of calibration questions is exhaustive and fairly represents the knowledge required for the elicitation; we do not feel this is a reasonable expectation and allow for probable inadequacy of the selected calibration questions by reducing the power. The effect of reducing the power is to distribute the weights more evenly across the experts. Reducing the power can be carefully balanced to not significantly reduce the overall performance of the decision-maker.

15.5.2 Challenges

Over the past few years, we have moved from relatively informal elicitation processes to a more structured protocol. We have encountered various challenges along the way, often associated with stakeholders' unfamiliarity with SEJ. For example, several times stakeholders welcomed the use of the Classical Model because they thought that the performance-based weighting would allow them to know who to ask in the future. We explained that the weights are specific to a particular calibration exercise with questions designed for the target questions, and that the discussion in the workshop-style sessions is essential for the experts to gain a comprehensive understanding of the problem and arrive at their estimates. Also, following ethics protocols, the weights are anonymous and not shared.

Another challenge has been the small size of the team at GNS Science that is involved in SEJ; for some applications, team members were also domain experts. It can be difficult to keep role separation and avoid real or perceived conflicts of interest.

Sections 15.3.2.1 and 15.3.2.2 include a description of the challenges in developing calibrations questions. In particular, when the target questions are weights and probabilities there may be extra difficulty in finding suitable calibration questions. These challenges can be overcome with experience and allowance for extra time, both for the project team and the experts.

15.5.3 Benefits of Our Protocol

Our protocol has solid foundations, being built on the principles of risk management (ISO 2009) and using the Classical Model for combining expert judgement (Cooke 1991). It aims to provide training for experts to recognize their own biases and limitations in their knowledge. The discussions during the workshop ensure the representation of varied opinions and experiences. One key aspect of the protocol is that the uncertainty estimates can be propagated through to the results (Gerstenberger and Christophersen 2016).

Given that it is challenging to quantify the success of any protocol, in particular when estimating low-probability events, we measure the success of our protocol by its acceptance by the stakeholder, its ability to produce results in a timely manner and its solid foundations. We find that experts enjoy the experiences, in particular the thorough discussions of the subject matter in the workshop-style setting. Therefore, they contribute their thoughts and understanding of the problem, which leads to the development of new knowledge and advanced understanding.

15.5.4 Outlook

Developing measures for evaluating different SEJ methods continues to be an important topic for further research. Given that our focus is on applications of SEJ, we do not have much of an opportunity to conduct methodological research into SEJ. Hence it is important for us to stay involved with the international community on SEJ to have the opportunity to present and discuss our work with the experts in the field. As a consequence of these interactions, and further experiences in future applications, our protocol will continue to evolve.

Acknowledgments We would like to acknowledgement Roger Cooke and Willy Aspinall for their support and encouragement in applying the Classical Model. We thank all of the experts who took part in different projects. David Rhoades and Anca Hanea reviewed the manuscript before submission. Annemarie Christophersen thanks Anca Hanea for many lively and educational discussions, and acknowledges a travel grant by the Royal Society of New Zealand to attend the workshop on "Science, uncertainty and decision-making in the mitigation of natural risk" of the European Union funded COST Action "Expert Judgment Network: Bridging the Gap Between Scientific Uncertainty and Evidence-Based Decision Making".

References

Allen, T. I., Griffin, J., Leonard. M., Clark, D., & Ghasemi, H. (2018) The 2018 National Seismic Hazard Assessment for Australia: model overview. *Geoscience Australia, Canberra, Australia.* https://doi.org/10.11636/Record.2018.027.

Aspinall, W. (2008) Expert judgment elicitation using the classical model and EXCALIBUR. Retrieved 8 June 2018, from http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf.

Bang, D., & Frith, C. D. (2017). Making better decisions in groups Royal Society Open Science 4 https://doi.org/10.1098/rsos.170193.

Bebbington, M. S., Stirling, M. W., Cronin, S., Wang, T., & Jolly, G. (2018). National-level long-term eruption forecasts by expert elicitation. *Bulletin of Volcanology, 80,* 56. https://doi.org/10.1007/s00445-018-1230-4.

Bradley, B. A. (2010). NZ-specific pseudo-spectral acceleration ground motion prediction equations based on foreign models. University of Canterbury.

Bradley, B. A. (2013). A New Zealand-specific Pseudospectral acceleration Ground-Motion prediction equation for active shallow crustal earthquakes based on foreign models. *Bulletin of the Seismological Society of America, 103,* 1801–1822. https://doi.org/10.1785/0120120021.

Burgman, M. A., et al. (2011). *Expert Status and Performance PLOS ONE, 6,* e22998. https://doi.org/10.1371/journal.pone.0022998.

Christophersen, A., Deligne, N. I., Hanea, A. M., Chardot, L., Fournier, N., & Aspinall, W. P. (2018). Bayesian Network modeling and expert elicitation for probabilistic eruption forecasting: pilot study for Whakaari/White Island. *New Zealand Frontiers in Earth Science, 6,* 23.

Christophersen, A., Gerstenberger, M., & Nicol, A. (2011) The feasibility of using seed questions for weighting expert opinion in CCS risk assessment. CO2CRC.

Christophersen, A., Rhoades, D. A., Gerstenberger, M. C., Bannister, S., Becker, J., Potter, S. H., & McBride, S. (2017). Progress and challenges in operational earthquake forecasting in New Zealand. *Paper presented at the New Zealand Society for Earthquake Engineering Technical Conference, Michael Fowler Centre, Wellington*, 27–29 April 2017.

CO2CRC. (2011). CRC for Greenhouse Gas Technology. http://www.co2crc.com.au/ (2018).

Colson, A. R. (2016). Excalibur tutorial. Retrieved 8 June 2018, from https://www.expertsinuncertainty.net/Portals/60/ESR%20Warsaw/Excalibur%20tutorial.pdf?ver=2017-11-27-124915-117.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* USA: Oxford University Press.

Cooke, R. M., ElSaadany, S., & Huang, X. (2008). On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering & System Safety, 93,* 745–756.

Cooke, R. M., & Goossens, L. H. J. (1999). Procedures guide for structured expert judgment. Luxemburg.

Cooke, R. M., & Goossens, L. L. H. J. (2008). *TU Delft Expert Judgment Data Base Reliability Engineering & System Safety, 93,* 657–674.

Cooke, R. M., & Solomantine, D. (1992). EXCALIBUR—integrated system for processing expert judgements. Delft, The Netherlands.

Deligne, N. I., Jolly, G. E., & Taig, T. (2018). Evaluating life-safety risk for field work on active volcanoes: VoLiST, a volcano observatory's decision-support tool Journal of Applied Volcanology in review.

Gerstenberger, M., Christophersen, A., Buxton, R., Allinson, G., Hou, W., Leamon, G., et al. (2012). *Integrated risk assessment for CCS Energy Procedia, 37,* 2775–2782.

Gerstenberger, M. C., & Christophersen, A. (2016). A Bayesian network and structured expert elicitation for Otway Stage 2C: Detection of injected CO2 in a saline aquifer *International Journal of Greenhouse Gas Control*, *51*, 317–329. https://doi.org/10.1016/j.ijggc.2016.05.011.

Gerstenberger, M. C., Christophersen, A., Buxton, R., & Nicol, A. (2015). Bi-directional risk assessment in carbon capture and storage with Bayesian Networks. *International Journal of Greenhouse Gas Control, 35,* 150–159. https://doi.org/10.1016/j.ijggc.2015.01.010.

Gerstenberger, M. C., Kaneko, Y., Fry, B., Wallace, L., Rhoades, D., Christophersen, A., & Williams, C. (2017). Probabilities of earthquakes in central New Zealand. Lower Hutt (NZ). https://doi.org/10.21420/g2fp7p.

Gerstenberger, M. C., McVerry, G. H., Rhoades, D. A., & Stirling, M. (2014). Seismic hazard modelling for the recovery of Christchurch. *New Zealand Earthquake Spectra, 30,* 17–29. https://doi.org/10.1193/021913EQS037M.

Gerstenberger, M. C., Rhoades, D. A., & McVerry, G. H. (2016). A Hybrid time-dependent probabilistic seismic-hazard model for canterbury. *New Zealand Seismological Research Letters, 87,* 1311–1318. https://doi.org/10.1785/0220160084.

Griffin, J., et al. (2018). *Expert elicitation of model parameters for the 2018 National Seismic Hazard Assessment*. Canberra, Australia: Geoscience Australia.

Hanea, A. M., McBride, M. F., Burgman, M. A., & Wintle, B. C. (2018). The value of performance weights and discussion in aggregated expert judgments. *Risk Analysis, 38,* 1781–1794. https://doi.org/10.1111/risa.12992.

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2017). *A practical guide to structured expert elicitation using the IDEA protocol Methods in Ecology and Evolution, 9,* 169–180. https://doi.org/10.1111/2041-210X.12857.

Hemming, V., Walshe, T. V., Hanea, A. M., Fidler, F., & Burgman, M. A. (2018). Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management PLoS ONE, *13* https://doi.org/10.1371/journal.pone.0198468.

Hincks, T. K., Komorowski, J. C., Sparks, S. R., & Aspinall, W. P. (2014) Retrospective analysis of uncertain eruption precursors at La Soufrière volcano, Guadeloupe, 1975–77: volcanic hazard assessment using a Bayesian Belief Network approach. *Journal of Applied Volcanology*, *3*, 1–26.

ISO. (2009). *ISO 31000:2009 Risk management—Principles and guidelines*. Geneva, Switzerland: International Organization for Standardization.

Jenkins, C. R. et al. (2012). Safe storage and effective monitoring of CO2 in depleted gas fields Proceedings of the National Academy of Sciences, *109*, E35-E41 https://doi.org/10.1073/pnas.1107255108.

Jolly, G. E., Keys, H. J. R., Procter, J. N., & Deligne, N. I. (2014). Overview of the co-ordinated risk-based approach to science and management response and recovery for the 2012 eruptions of Tongariro volcano. *New Zealand Journal of Volcanology and Geothermal Research*, *286*, 184–207 https://doi.org/10.1016/j.jvolgeores.2014.08.028.

Kahneman, D. (2011). *Thinking, fast and slow*. Straus and Giroux, New York: Farrar.

Kaneko, Y., Wallace Laura, M., Hamling Ian, J., & Gerstenberger Matthew, C. (2018). Simple physical model for the probability of a subduction-zone earthquake following slow slip events and earthquakes: application to the Hikurangi Megathrust. *New Zealand Geophysical Research Letters, 45,* 3932–3941. https://doi.org/10.1029/2018GL077641.

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting?: A social psychological analysis. *International Journal of Forecasting, 27*, 14–40 https://doi.org/10.1016/j.ijforecast.2010.02.001.

Kunda, Z. (1990). *The Case for Motivated Reason Psychological Bulletin, 108,* 480–498. https://doi.org/10.1037/0033-2909.108.3.480.

Lighttwist Software. (2008). Excalibur. Retrieved 8 June 2018, from http://www.lighttwist.net/wp/excalibur.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Taylor & Francis.

Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis risk. *Analysis, 35,* 1230–1251. https://doi.org/10.1111/risa.12360.

Montibeller, G., & von Winterfeldt, D. (2018). Individual and group biases in value and uncertainty judgments. In L. M. A. Dias, J. Quigley (Eds.) Elicitation, vol 261. International Series in Operations Research & Management Science. Springer, Cham. https://doi.org/10.1007/978-3-319-65052-4_15.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 502–517. https://doi.org/10.1037/0033-295X.115.2.502.

National Academies of Sciences, Engineering, and Medicine. (2017). Volcanic Eruptions and Their Repose, Unrest, Precursors, and Timing. The National Academies Press, Washington, DC. https://doi.org/10.17226/24650.

New Zealand Ministry of Civil Defence and Emergency Management. (2015). The Guide to the National Civil Defence Emergency Management Plan 2015.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

Pevzner, R., Caspari, E., Gurevich, B., Dance, T., & Cinar, Y. (2015). Feasibility of CO2 plume detection using 4D seismic: CO2CRC Otway Project case study—Part 2: Detectability analysis GEOPHYSICS 80, B105-B114 https://doi.org/10.1190/geo2014-0460.1.

Pevzner, R. et al. (2017). Stage 2C of the CO2CRC Otway Project: Seismic Monitoring Operations and Preliminary Results Energy Procedia, *114*, 3997–4007 https://doi.org/10.1016/j.egypro.2017.03.1540.

Potter, S. H., Jolly, G. E., Neall, V. E., Johnston, D. M., & Scott, B. J. (2014). Communicating the status of volcanic activity: Revising New Zealand's volcanic alert level system. *Journal of Applied Volcanology*, 3.

Quigley, J., Colson, A., Aspinall, W., Cooke, R. M. (2018) Elicitation in the classical model. In L. C. Dias, A. Morton, J. Quigley (Eds.) *Elicitation: The science and art of structuring judgement*. Springer International Publishing, Cham, pp. 15–36. https://doi.org/10.1007/978-3-319-65052-4_2.

Robinson, R., Van Dissen, R., & Litchfield, N. (2011). Using synthetic seismicity to evaluate seismic hazard in the Wellington region. *New Zealand Geophysical Journal International, 187,* 510–528. https://doi.org/10.1111/j.1365-246X.2011.05161.x.

Sparks, R. S. J., Biggs, J., & Neuberg, J. W. (2012). *Monitoring Volcanoes Science, 335,* 1310–1311. https://doi.org/10.1126/science.1219485.

Stirling, M., et al. (2017). Conceptual development of a national volcanic hazard model for New Zealand frontiers in Earth. *Science, 5,* 51.

Stirling, M., McVerry, G., Gerstenberger, M., Litchfield, N., Van Dissen, R., Berryman, K., Barnes, P., Wallace, L., Villamor, P., Langridge, R., Lamarche, G., Nodder, S., Reyners, M., Bradley, B., Rhoades, D., Smith, W., Nicol, A., Pettinga, J., Clark, K., & Jacobs, K. (2012). National seismic hazard model for New Zealand: 2010 update. *Bulletin of the Seismological Society of America*, *102*, 1514–1542. https://doi.org/10.1785/0120110170.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*, 1958–1971. https://doi.org/10.1037/a0037099.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185,* 1124–1131.

Wallace, L. M. et al. (2017). Large-scale dynamic triggering of shallow slow slip enhanced by overlying sedimentary wedge. *Nature Geoscience*, *10*, 765 https://doi.org/10.1038/ngeo3021.

Westen, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: An FMRI study of emotional constraints on partisan political judgment in the 2004 U.S. *Presidential election Journal of cognitive neuroscience, 18,* 1947–1958. https://doi.org/10.1162/jocn.2006.18.11.1947.