Predicting the Validity of Expert Judgments in Assessing the Impact of Risk Mitigation Through Failure Prevention and Correction

Mario P. Brito[1,*] **and Ian G. J. Dawson**[2]

ABSTRACT: Operational risk management of autonomous vehicles in extreme environments is heavily dependent on expert judgments and, in particular, judgments of the likelihood that a failure mitigation action, via correction and prevention, will annul the consequences of a specific fault. However, extant research has not examined the reliability of experts in estimating the probability of failure mitigation. For systems operations in extreme environments, the probability of failure mitigation is taken as a proxy of the probability of a fault not reoccurring. Using *a priori* expert judgments for an autonomous underwater vehicle mission in the Arctic and *a posteriori* mission field data, we subsequently developed a generalized linear model that enabled us to investigate this relationship. We found that the probability of failure mitigation alone cannot be used as a proxy for the probability of fault not reoccurring. We conclude that it is also essential to include the effort to implement the failure mitigation when estimating the probability of fault not reoccurring. The effort is the time taken by a person (measured in person-months) to execute the task required to implement the fault correction action. We show that once a modicum of operational data is obtained, it is possible to define a generalized linear logistic model to estimate the probability a fault not reoccurring. We discuss how our findings are important to all autonomous vehicle operations and how similar operations can benefit from revising expert judgments of risk mitigation to take account of the effort required to reduce key risks.

KEY WORDS: Autonomous unmanned vehicles; expert judgment; extreme environments; risk mitigation; risk perception

1. INTRODUCTION

In recent decades there has been a rapid growth in the development and application of autonomous intelligent vehicles across a variety of contexts and for a range of purposes (e.g., remote repairs in space, military reconnaissance, driverless cars, material handling systems in factories and laboratories, etc.) (Bishop, 2000; Cox & Wilfong, 2012). This growth has been particularly evident in the use of autonomous technologies to explore remote, uncharted, or extreme underwater territories for scientific research (Bellingham & Rajan, 2007; Jenkins et al., 2010; Singh et al., 2004). For example, autonomous underwater vehicles (AUVs) are now regularly used by oceanographers, marine biologists, and climate scientists to explore underneath large ice shelfs and fast-moving sea ice in the Arctic and Antarctic regions (Banks, Brandon, & Garthwaite, 2006; Dowdeswell et al., 2008).

[1] Associate Professor in Risk Analysis and Risk Management, Centre for Risk Research, University of Southampton, Southampton, UK.

[2] Associate Professor of Risk Management, University of Southampton, Southampton, UK.

[*] Address correspondence to Mario P. Brito is University of Southampton, Southampton Business School, Building 2, SO17 1BJ, Southampton, UK; tel: +44 (0)23 80597583; fax: +44 (0)23 8059 3844; m.p.brito@soton.ac.uk

0272-4332/20/0100-1928$22.00/1

AUVs are mechatronic systems that operate without any physical connection to a ship or human and, once launched, independently navigate underwater following a preplanned mission profile and typically transmit data to an onshore laboratory for real-time analysis (McPhail, 2009; Yoerger, Jakuba, Bradley, & Bingham, 2007). These missions are inherently risky due to the potential for technological failure and the uncertainty surrounding the physical conditions that the AUVs will encounter (Brito & Griffiths, 2016). The loss of an AUV can prove extremely costly due to the expense of developing, building, deploying, and operating the vehicles. Furthermore, AUV loss may lead to substantial delays in, or even the termination of, large-scale, long-term international research projects (Brito, Griffiths, & Challenor, 2010).

State-of-the-art risk analysis and management processes have been instrumental in facilitating successful AUV missions (Brito et al., 2010, 2012). However, due to the novel nature of AUV technology and the unique characteristics of each mission, risk analysis tends to be heavily reliant on expert judgments to compensate for the absence of past data about mission success rates and the factors that might influence the estimated likelihood of failure (Brito et al., 2012). Expert judgments have been used to effectively reduce risk in the field of AUV operations. For example, a study by Brito et al. (2012) showed that the use of expert judgments regarding the effectiveness of mission failure mitigation measures for an AUV mission in the Arctic led to a 24% reduction in the probability of losing the vehicle for a single mission of 336 km (an estimated reduction of this size can make the difference between whether a mission is executed or aborted). Therefore, it is important to assess the reliability of experts in assessing the probability of failure mitigation. Failure mitigation and probability of failure mitigation are terms that are used several times in this article. Failure mitigation is a set actions to correct failures that have occurred in the past. The probability of failure mitigation is the likelihood of these corrective actions annulling the failure. A probability of failure mitigation of 1 means that the experts expect the mitigation action to annul the failure. Whereas, a probability of failure mitigation of 0 means that experts expect that the mitigation actions will not annul the failure.

Previous research in expert judgment elicitation has focused on the reliability of experts to estimate the likelihood of a catastrophic event taking place. To our knowledge, the reliability of experts in estimating the probability of failure mitigation has not been addressed before. Some notable limitations of the extant research in this area are that it is not clear to what extent (i) experts weight each mitigation variable when assessing their composite influence on the overall probability of failure mitigation and (ii) the accuracy of expert judgments about the effectiveness of overall failure mitigation is determined by their assessment of the effectiveness of each mitigation action. This article addresses this knowledge gap. Specifically, we present a generalized linear model (GLM) that enables us to use past AUV mission data to retrospectively assess the accuracy of expert estimates of the probability of failure mitigation. Moreover, we use GLM to identify which actions/variables (knowledge, past successes, effort needed to mitigate fault, etc.) most strongly predict the accuracy of the expert's probability of mitigation estimates. Our findings provide important insights for autonomous vehicle operations and allow us to suggest how such operations can benefit from revising expert judgments to increase the probability of successful missions.

1.1. AUV Failure Mitigation

AUV failures are typically caused by a single faulty component, human error, or a combination/sequence of minor faults that, individually, would not normally thwart the mission (Brito et al., 2010). A "fault" (e.g., engine breakdown, data transmission failure, etc.) in an AUV mission is deemed to be any AUV-specific operational error that results in a failure to achieve the mission plan. Risk mitigation for an AUV deployment under ice is achieved by reducing the probability of a given fault occurring. A fault can be mitigated by redesigning the faulty system or component, by correcting the fault, or by not triggering that system/component during deployment. Failure mitigation is the process of annulling the consequences of fault by redesigning the system to correct the fault or by designing a fail-safe or fault-tolerant system (Leveson & Harvey, 1983). Failure mitigation is generally regarded as the most important strategy for managing catastrophic risks in complex technology failures (Subramanian, Elliott, Vishnuvajjala, Tsai, & Mojdehbakhsh, 1996).

Unlike a rover deployment on planet Mars, when an AUV is deployed underneath an ice shelf it is not possible to communicate with the vehicle. Consequently, risk reduction in AUV missions cannot be achieved by implementing traditional risk management strategies such as contingency planning or by designing flexibilities into the system. The implementation of a contingency plan would imply that an adverse outcome had occurred and a predetermined course of action could be taken to continue to the mission. For example, if an AUV was lost under ice a contingency might be to deploy a second AUV. However, in an unexplored and inaccessible environment, it is generally impossible to determine with any certainty the cause of AUV loss because the most important sources of evidence for the accident investigation are stored in the vehicle itself. Hence, in such a scenario, the deployment of another vehicle could lead to another AUV loss caused by the same systematic fault. Similarly, adding flexibility into the mission (e.g., varying the objectives, exploration area, measures, etc.) would also be ineffective because this would probably result in a failure to obtain the required data sample and, thus, failure to meet the mission objectives.

In practice, the risk of AUV loss is typically managed by running monitoring distance trials before the actual mission or via failure mitigation processes. Monitoring distance is a concept analogous to "burn in time," whereby the AUV is monitored under benign conditions for a given period before it is committed to the mission. In this article, we focus on failure mitigation, via fault correction and prevention, because this approach to managing operational risk management has previously demonstrated the most effective outcomes in practice (Brito et al., 2010, 2012).

Most high reliability organizations (HROs) have developed systems for taking into account fault mitigation (Feather & Cornford, 2003). For example, Feather and Cornford present a hazard management framework, developed by the national aeronautical and space administration (NASA) to monitor and update the likelihood of design failure modes occurring. The system, denoted as defect detection and prevention (DDP), is a probabilistic model. The key assumption is that each fault may have a number of prevention, detection, and alleviation (PACTs) methods (i.e., fault mitigations). The expected efficiency of each PACT in mitigating the fault is assessed by a group of specialized field experts. The DDP system considers that multiple PACTs may have an adverse or positive effect on a failure mode because PACTs are not independent and, therefore, may introduce a fault in the system. The probability model aggregates all these effects to quantify the probability of fault mitigation. As described below, similar methods have been used in AUV fault mitigation.

In 2010, the international submarine engineering (ISE) Explorer AUV conducted a record breaking mission, travelling a total of 10,000 km underneath fast sea ice (Kaminski et al., 2010). The risk analysis for the ISE Explorer missions included the quantification of the impact of risk mitigation actions (Brito et al., 2012). Specifically, the probability that a given mitigation action would eliminate a fault was estimated by experts during a risk assessment workshop. The AUV risk profile used for estimating the probability of AUV loss for a given distance was first created using the expert's subjective judgments. It was then revised based on the probability of failure mitigation estimated by experts. The subsequent risk analysis for ISE Explorer considered the risk profile with and without mitigation. Six missions were later conducted underneath ice in 2011: missions 51 (30.51 km), 52 (55.8 km), 53 (131.22 km), 54 (336.24 km), 55 (325.98 km), and 56 (324.45 km), with failure mitigation leading to increased risk reductions (cf. without mitigation) of 11, 13, 16, 24, 16, and 16%, respectively, for each of the six missions.

Failure mitigation is an integral part of engineering risk management. Previous research has found a positive relationship between the extent to which the causes of a failure are understood and the subsequent probability of mitigating that failure in future AUV missions (Subramanian et al., 1996). The probability of failure mitigation is assumed to be a proxy for the probability of fault not reoccurring. Yet, there is no evidence to relate the probability of failure mitigation with the probability of fault not reoccurring. Moreover, the extent to which each of these factors influences the probability of failure mitigation has not been quantitatively assessed. This is important for two reasons. First current risk models can be updated using the probability of mitigation agreed by the experts. If other variables are deemed also relevant then these should also be included in the model (Brito et al., 2012; Hill, Thomas, & Allen, 2000). Second, it is important to have a means to assess expert judgment elicitations regarding the effectiveness of the fault correction and prevention because this can be the main contributor to risk reduction.

1.2. Expert Judgment Elicitation

Expert judgment elicitation is a discipline in risk analysis which typically seeks to obtain estimates of risk and uncertainty from experts when such information is needed to augment historical/statistical data or when historical/statistical data is unavailable (Merkhofer, 1987; Otway & von Winterfeldt, 1992).

Reviews of expert judgment elicitations indicate that although previous elicitations have been applied with good intentions, they have often had methodological short comings that have led to judgmental biases and overconfidence (Kastenberg, 1987; Kouts, 1987; Wright, Bolger, & Rowe, 2002). For example, Keeney and Winterfeldt (1991) identified that (i) elicitation processes often lacked input from a heterogeneous range of experts, (ii) the experts were rarely trained in assessing probabilities, (iii) experts were not given sufficient time to assimilate the relevant information, and (iv) elicitors neglected to use state-of-the-art methods for eliciting the judgments (Keeney & Winterfeldt, 1991). Relatedly, Bolger and Wright (1994) highlighted that care is needed when determining whether someone is or is not "an expert." Specifically, they stated that an expert should not be selected based on his/her social, professional, or political status, but based on the extent to which he/she can provide a judgment that is both ecologically valid (i.e., the judgment task is one that the individual regularly performs in his/her professional role) and learnable (i.e., the judgment task typically leads to some form of feedback on the correctness or reliability of the judgment and that feedback can be utilized in similar future tasks). Hence, care is needed in both the selection of experts and in the selection and application of methods to elicit the judgments.

Formal processes can address many of the potential shortcomings associated with expert judgment elicitations. Such processes have been advocated by academic researchers (Kynn, 2008; Tversky & Kahneman, 1974) and employed by several HROs and government institutions (e.g., see Bonano, Hora, Keeney, & Winterfeldt, 1990; Goossens, Cooke, Hale, & Rodić-Wiersma, 2008). These formal processes typically involve the introduction of elicitation techniques that aim to minimize biases and use structured methods to combine and then aggregate the judgments mathematically or behaviorally (mathematical aggregation implies that expert judgments are elicited individually and then combined using analytical functions) (Morris, 1977; O'Hagan et al., 2006; Otway & von Winterfeldt, 1992; Phillips & Wisbey, 1993).

1.3. A Static Risk Model for AUVs Using Expert Judgments

Risk models based on expert judgments have already been successfully developed for AUV deployments. For example, Brito et al. (2010) conducted a formal expert judgment elicitation to build the risk profile for the Autosub3 AUV missions under the Pine Island glacier in Antarctica. On this occasion, eight experts, who included senior AUV engineers and AUV users based in the United States, took part in a judgment elicitation process that was closely based on the elicitation methodology proposed by Otway and Winterfeldt (1992). Specifically, the experts were individually asked to estimate the probability that each fault in the Autosub3 fault history would lead to the loss of Autosub3 in four different operating environments: open water, coastal water, sea ice, and ice shelf. In addition to the likelihood that a fault would lead to loss, each expert was also asked to state his/her confidence in the assessment in the form of a weight which could vary from 1 (not confident) to 5 (very confident). The judgments were aggregated using both the linear and the log mathematical aggregation methods. A similar approach was used successfully by Griffiths, Brito, Robbins, and Moline (2009) to build a risk model for two Remus 100 AUVs. On this occasion, instead of providing a single-point estimate, each expert provided the parameters of a distribution for the likelihood of fault leading to loss. The parameters provided were the lower bound, upper bound, lower quartile, upper quartile, and median. Here the analysis did not take into account the probability of failure mitigation. However, a risk model that takes into account the probability of failure mitigation is presented in Brito et al. (2012). The risk model considered in previous research comprised the duplet $<F_i, L_i>$, where F_i stands for fault i and L_i is the likelihood of fault i leading to AUV loss. Details on how to incorporate the probability of failure mitigation in the risk model are presented in the following section.

1.4. Accounting for the Probability of Failure Mitigation

In 2012, the model proposed by Brito et al. (2010) was further developed to include the probability of failure mitigation (Brito et al., 2012). Analytically, the probability of fault leading to loss given a mitigation action is quantified using Equation 1. Brito et al. (2010) define the probability of a failure i being mitigated as P_{M_i}. This probability is estimated by experts. A P_{M_i} of 0 is assigned if there is no confidence that the mitigation action will eliminate the failure, and a value of 1 is assigned if there is certainty that the mitigation strategy will remove the failure. For each failure, the experts were asked to agree on a single figure for the probability of failure being mitigated given the mitigation actions defined by the project team (P_{M_i}). The probability of AUV loss given a fault i, L_i, in environment E and the fault mitigation M_i is calculated using Equation 1. F_i stands for fault i.

$$P\ (L_i\,|F_i, E, M_i)) = \ P\,(L_i\,|F_i, E))\,(1 - P_{M_i}) \quad (1)$$

This model holds no knowledge of subsequent missions.

For each mission carried out by an AUV, the data obtained can be used for assessing the observed (i.e., objective) probability of fault reoccurring. If a large number of missions have been conducted, expert judgments of risk and risk mitigation effectiveness could then be replaced by the observed data. However, in reality, it is generally the case that data from only a few missions, at most, are available for each AUV. Hence, the ideal scenario would be to devise a method of ascertaining if expert judgments are representative of the observed probability of fault reoccurring when only a small sample of past AUV mission data is available. Such a method would not only be highly beneficial for AUV operations, but also for any new technology where something is known about both the operational risk and risk mitigation effectiveness. Such a method would overcome situations where there is a lack of observed data to ascertain whether or not expert judgments are representative of future observations. How we addressed the need for such a method is discussed in the following section.

2. A NOVEL METHOD TO ASSESS THE IMPACT OF PROBABILITY OF MITIGATION

GLMs (McCullagh & Nelder, 1989) neatly synthesize likelihood-based approaches to regression analysis. There are several extensions of this theory involving models with random terms in the linear predictor. In this article, a GLM is employed to estimate how the dependent variable, (i.e., probability of fault not reoccurring) is influenced by several independent (predictor) variables.

For a given AUV mission, the fault reoccurrence can be seen as a binomial trial whereby the mission either succeeds or fails. Therefore, a causal model can be defined where the probability of fault reoccurring would depend on both (i) an *a priori* assessments of the probability of fault being mitigated and (ii) experts perception of the failure mitigation process during the subsequent observed missions. Therefore, the probability of a fault not reoccurring can be synthesized in the GLM presented in Equation 2:

$$\text{logit}\,(p_i) = \alpha_0\ + \sum_{j=1}^{m} \alpha_j X_{j,i} + b_i \qquad (2)$$

$$r_i = \ Binomial(p_i, n_i)$$

$$b_i \sim Normal(0, \tau)$$

where logit is:

$$logit(p_i)\ =\ \log\left(\frac{p_i}{1 - p_i}\right) \qquad (3)$$

The p_i is the dependent variable and X_{ij} are independent variables. In order to test the effectiveness of fault mitigation, p_i is the probability of a fault not reoccurring. Each mission is seen as a trial for the fault mitigation action. The noise in the observations is measured by b_i. This variable is assumed to be normally distributed with standard deviation τ. The posterior is modeled with a binomial distribution, where n_i is the total number of possible outcomes and r_i is the number of favorable outcomes (i.e., fault did not emerge). The maximum likelihood expression for the GLM is automatically generated by a Bayesian inference statistical software tools, such as OpenBugs. The inference is conducted using the Markov Chain Monte Carlo (MCMC) method (Breslow & Clayton, 1993). The MCMC inference is a stochastic Bayesian inference which estimates the properties of the marginal probabilities based on samples of the conditional probability function. The stopping criterion is defined by the MCMC error. Best practice is to stop the simulation when the MCMC error is 5% of the standard deviation (Breslow & Clayton, 1993).

2.1. ISE Explorer Case Study

We started by using the behavioral risk model created by Brito et al. (2012) for the ISE Explorer AUV. We used this model because the probability of failure mitigation is provided for each fault. The information missing from Brito et al. (2012) is with respect to the variables that may influence the probability of fault not reoccurring. This risk model was developed using a behavioral risk elicitation process denoted as SHELF (Oakley & O'Hagan, 2010). Five experts participated in the expert risk judgment elicitation, which took place in two parts. The first part took place in Halifax, Canada, on January 26–29, 2010 and the second part took place in Vancouver, Canada, in 2011. Five experts were selected because ISE Explorer is an AUV developed by ISE Ltd. The number of individuals with knowledge on this vehicle is extremely small. The experts selected comprised the most experienced users and developers of the vehicle. It was important to have a balance between knowledge of the vehicle and knowledge of Arctic operations. All selected experts had conducted several missions in the Arctic and had previously provided probability judgments for AUV missions risk assessments.

For the elicitation process that took place in Halifax, the experts received training in probabilistic inference and statistics and then provided judgments via the SHELF expert judgment elicitation approach (O'Hagan et al., 2006). The SHELF (Oakley & O'Hagan, 2010) process was developed by Sheffield University as a behavioral approach to eliciting expert judgments. Distinct features of SHELF are that experts are specifically encouraged to provide probability distributions, instead of a single probability assessment, and are expected to agree on the final assessment. In the first round of the elicitation, experts agree on the lower and upper bound of the probability assessment. Then individually, experts define their uncertainty about the assessment by specifying a distribution. This distribution is specified using the median, lower quantile, and upper quartile. Each distribution is then plotted for all experts to examine and the reasons underpinning a given distribution are discussed. Finally, the experts agree on the values for the median, lower quartile, and upper quartile for the distribution that represents the groups view. This process was adopted for the ISE Explorer to assess the probability of each fault leading to loss in the target environment. Following the completion of the risk assessment for each fault, the experts were asked to assess the probability that each mitigation action would annul the respective fault. Following a series of test trials of the ISE Explorer in benign environments, 54 faults were identified by the engineers and then assessed by an expert panel.

In June 2011, ISE explorer conducted three missions under fast moving ice in Greenland. The total travel distance was 10,000 km. The second part of the expert judgment elicitation that took place in Vancouver followed this second AUV campaign in the Arctic and had two key aims. The first was to update the risk model in light of the 32 new faults that had emerged on the ISE Explorer (named B05), and the second aim was to validate the approach adopted for risk assessment.

2.2. Preliminary Data

Our aim was to model mitigation at a greater level of granularity. Specifically, we extended the previous work in this field by Brito et al. (2010, 2012) by assessing the extent to which expert risk mitigation judgments could reduce the estimated risk of mission failure for an AUV. Moreover, we also examined the extent to which expert judgments about the effectiveness of *overall* failure mitigation is determined by the expert's assessment of the effectiveness of *each* mitigation action.

The data for our study was collected at the workshop held in Vancouver, British Columbia, from July 21–23, 2011. At this workshop, the expert panel was asked to visit the assessments provided by the panel in Halifax in 2010 for the probability of failure mitigation. The expert group used in the first elicitation was the same as that used in the second elicitation, with one difference being that one expert from the first workshop was unable to attend the second workshop and, therefore, was replaced by another expert from defense research & development Canada. This expert had been an observer at the first workshop and received training at both workshops. For each of the 43 mitigation actions implemented following the 2010 workshop, the experts were asked to answer questions in relation to the faults that occurred during six subsequent operational deployments:

- Builder's sea trials: 8 September to 12 October 2009
- Sea acceptance trails: 29 September to 30 September 2009
- Development trials: homing and positioning: 16 November to 4 December 2009; 11 December 2008 to 22 January 2010: 14 February 2010 to 28 February 2010
- Mission Testing: 2 February 2010 to 12 March 2010

The experts answered the following eight questions (question codes shown in brackets after each question):

(1) Did the fault turn out to be understood? (UND)
(2) Was the mitigation method implemented as described in the workshop? (IMP)
(3) Was a different mitigation strategy implemented? (OIT)
(4) Was the implementation mitigation strategy tested? (MTE)

(5) Did the mitigation prove robust in the field? (ROB)
(6) What was the effort taken to mitigate the fault? (EFF)
(7) Was the effort as expected? (EEP)
(8) Was the cost of implementing the mitigation as expected? (CEP)

Subramanian et al.'s (1996) analysis of patterns of fault mitigation in safety critical software systems identified effort and knowledge as key factors in the implementation of the fault mitigation. Questions 1, 5, 6, 7, and 8 attempted to capture the influence of these factors. Questions 2–5 attempted to establish whether or not the mitigation action discussed at the workshop was implement and whether or not the mitigation action proved robust in the field. The experts' agreed (mean) answers to these questions are presented in Table I.

The failure data after the workshops was collected from the operations of two identical vehicles B05 and B06. These vehicles were operated by the same team. Table II presents a summary of the missions conducted after the expert judgment elicitation conducted in Halifax, January 26–29, 2010.

3. ANALYSIS AND RESULTS

We conducted the analysis in two stages. First, we attempted to test if it was possible to fit a GLM to all variables considered by the experts to determine if they are significant in the probability of fault not reoccurring; this is presented in section 3.1. Our collinearity analysis and our proposed reduced model are presented in section 3.2.

3.1. Preliminary Analysis

Subramanian et al.'s (Subramanian et al., 1996) analysis of patterns of fault mitigation in safety critical software systems identified knowledge and effort as key factors in the implementation of the fault mitigation. Hence, in the first stage of our analysis, we tested the proposition that the probability of fault not reoccurring would depend on knowledge–related factors (i.e., probability of mitigation [PMIT] and understanding of the fault mitigation action [UND]), and effort-related factors (i.e., whether the mitigation strategy was tested [TES] and effort needed to implement the mitigation action [EFF]). The variable CEP had strong positive association with EEP. Consequently, we used only EEP in the analysis. The logit model, capturing these variables, is presented in Equation 4.

$$\begin{aligned}\mathrm{logit}\,(p_i) &= \alpha_0 + \alpha_1 PMIT + \alpha_2 UND + \alpha_3 IMP \\ &\quad + \alpha_4 OIT + \alpha_5 MTE + \alpha_6 ROB + \alpha_7 EFF \\ &\quad + \alpha_8 EEP + b_i \end{aligned} \tag{4}$$

This model was implemented in OpenBugs software and used MCMC inference to estimate the proportion coefficients as depicted in Equation 4. The results for the mean, standard deviation (*SD*), 5% quantile, median, 95% quantile, and the MCMC error are presented in the Table III.

The results showed that UND, OIT, MTE, and ROB were not significant in the estimation of the probability of failure not reoccurring. The results also showed that it was not possible to obtain a significant intercept. Collinearity analysis was required to assess whether or not some variables in this model were linearly related and, therefore, whether their effect was not significant in the estimation of the probability of a fault not reoccurring. The correlations between variables are presented in Table IV. There was a correlation between UND and ROB, with a Pearson's coefficient of 0.632 ($p < 0.005$). Other notable Pearson's coefficients are presented. For example, the coefficients between UND and PMIT and between UND and IMP were 0.4545 and 0.605, respectively ($p < 0.005$). Also, there was a correlation between ROB and UND, with a coefficient of 0.635 ($p < 0.005$) and between ROB and MTE with a coefficient of 0.565 ($p < 0.005$). Notably, there was a correlation between OIT and IMP, with a coefficient of 0.419 ($p < 0.005$).

This correlation was of interest because it showed that even amongst variables deemed significant (PMIT, IMP, EFF, and EEP) for the logistic GLM presented in Table III there was potential for multicollinearity. Nevertheless, we elected to explore the results for a reduced model comprising these four variables, which is presented in Equation 5:

$$\begin{aligned}\mathrm{logit}\,(p_i) &= \alpha_1\ PMIT + \alpha_3 IMP \\ &\quad + \alpha_7 EFF + \alpha_8 EEP + b_i \end{aligned} \tag{5}$$

Note that $p_i$and b_i are defined as presented in Equations 2 and 3. The results obtained using this model are presented in Table V.

The results obtained for the reduced model presented in Equation 5 showed that both IMP and EEP

Table I. ISE Explorer Failure Data and Expert Assessments

Fault Number	Reoccurrence	Probability of Effective Mitigation; 0 - Not Fixed; 1- Fixed (PMIT)	Did the Fault Turn Out to be Understood? 0 - Not Understood; 1- Understood (UND)	Was the Mitigation Method Implemented as Described in the Workshop? 0 - Not Implemented; 1- Implemented (IMP)	Was a Different Mitigation Strategy Implemented? (OIT)	Was the Implementation Mitigation Strategy Tested? 0 - Tested Understood; 1- Not Tested (MTE)	Did the Mitigation Prove Robust in the Field? From 0 to 1. 0 – Not Robust, Understood; 1- Robust (ROB)	What was the Effort Taken to Mitigate the Fault? (EFF)	Was the Effort as Expected? (EEP)
1	1	0	0	0	0	0	0	0	0
2	–	0.9	1	1	0	1	1	3	1
3	–	0.8	1	1	0	0.8	0.8	80	1
4	–	0.9	0	0	0	0	0	4	0
5	–	0.95	1	1	0	1	1	2	0
6	–	1	1	1	0	0	1	6	0
8	–	0.9	1	1	0	1	0.9	4	0
10	–	0.95	1	0	1	1	1	32	1
11	–	0.95	0	1	0	1	0	40	1
12	–	1	0	0	1	1	1	8	1
13	–	0.8	1	1	0	1	1	24	1
14	–	0.75	1	1	0	1	1	16	0
15	–	0.4	0	0	1	0	1	40	1
16	–	0.95	1	1	0	1	1	8	0
17	–	0	0	0	0	0	0	320	1
18	–	0.9	0	0	1	1	0.9	120	1
19	–	0.8	1	1	1	1	0	40	1
21	–	0.8	1	1	0	1	1	24	1
22	–	0	0	0	1	1	1	24	1
23	–	0.9	1	1	0	1	1	36	0
26	–	0.95	1	1	1	1	1	32	0
28	–	0.95	1	1	0	1	1	1	0

(*Continued*)

Table I (Continued)

Fault Number	Reoccurrence	Probability of Effective Mitigation; 0 - Not Fixed; 1- Fixed (PMIT)	Did the Fault Turn Out to be Understood? 0 - Not Understood; 1- Understood (UND)	Was the Mitigation Method Implemented as Described in the Workshop? 0 - Not Implemented; 1- Implemented (IMP)	Was a Different Mitigation Strategy Implemented? (OIT)	Was the Implementation Mitigation Strategy Tested? 0 - Tested Understood; 1- Not Tested (MTE)	Did the Mitigation Prove Robust in the Field? From 0 to 1. 0 – Not Robust, Understood; 1- Robust (ROB)	What was the Effort Taken to Mitigate the Fault? (EFF)	Was the Effort as Expected? (EEP)
29	1	0.9	0	1	1	1	1	80	1
30	1	0.8	1	1	0	1	1	4	0
31	1	0.95	1	1	1	1	1	32	0
32	1	1	1	0	1	1	1	16	1
33	1	1	1	0	1	1	1	8	0
34	1	0.9	1	1	0	1	1	0.5	0
35a	1	1	1	1	0	1	1	1	0
35b	1	0.6	1	1	0	1	1	28	0
36	1	0.9	1	1	0	1	1	8	0
38	1	1	1	1	1	1	1	16	1
39	1	0.75	1	1	0	1	1	1	0
41	1	0.5	1	1	0	0	0.8	16	1
42	3	0.5	0	0	0	0	0	1	0
43	1	0.5	0.8	1	0	1	1	40	1
44	3	0.5	0	0	0	0	0	1	0
46	2	0.5	1	1	0	1	1	2	1
47	1	0.5	1	1	0	1	1	24	0
48	1	0.5	1	0	1	1	1	4	0
50	1	0.5	0.8	1	0	1	1	40	1
51	1	0.95	1	1	1	1	1	32	0
52	1	0.5	1	1	0	1	1	24	0
53	1	0.5	1	1	0	1	1	24	0
54	2	0.1	0	1	0	1	0	4	0

PMIT was obtained in the Halifax expert judgment elicitation, held on January 26–28, 2010. The estimates for UND, IMP, OIT, MTE, ROB, EFF, and EEP were elicited at the workshop held in Vancouver, July 19–21, 2011. Failure reoccurrence, second column were obtained during missions conducted by B05 and B06 from the April 4–5, 2010 to March 22, 2011.

Table II. Mission Summary Data Since the Expert Judgment Elicitation Workshop Conducted in Halifax, on January 26—29, 2010 Where the Probability of Failure Mitigation was Elicited for 54 Faults

Mission	Vehicle	Location	Date	Duration (Hours [h] and Minutes [min])	Faults
1	B05	Arctic survey	4–5th April 2010	5 h 39 min	42 (7), 44 (7), 46 (5)
2	B05	Arctic survey	7th April 2010	10 h 20 min	
3	B05	Arctic survey	8–9th April 2010	24 h 18 min	54 (8)
4	B05	Arctic survey	12–14th April 2010	62 h 16 min	
5	B05	Arctic survey	16–18th April 2010	60 h 22 min	
6	B05	Arctic survey	19– 22nd April 2010	60 h 5min	
7	B05	Vancouver trials	17th February 2010	52 min	1 (17), 42 (20), 44 (20), 46 (22)
8	B05	Vancouver trials	18th February 2011	35 min	
9	B05	Vancouver trials	22nd February 2011	2 h 38 min	54 (24)
10	B05	Bedford Basin trials	14th June 2011	3 h 8 m	
11	B05	Bedford Basin trials	15th June 2011	3 h	
12	B06	Vancouver trials	28th of February 2011	1 h 46 min	
13	B06	Vancouver trials	1st of March 2011	2 h 43 min	46 (24), 42 (35), 44 (35)
14	B06	Vancouver trials	4th of March 2011	1 h 17 min	
15	B06	Bedford Basin trials	17th of March 2011	51 min	
16	B06	Bedford Basin trials	18th of March 2011	20 h 18 min	
17	B06	Bedford Basin trials	19th of March 2011	14 h 53 min	
18	B06	Bedford Basin trials	20th of March 2011	6 h 35 min	
19	B06	Bedford Basin trials	21st of March 2011	2 h 33 min	
20	B06	Bedford Basin trials	22nd of March 2011	1 h 9 min	

The last column presents the fault reference that has reoccurred in from the data set considered by the experts as in presented in Brito et al. (2012) outside the brackets. The fault number of the fault reoccurrence is presented in the brackets.

Table III. Generic Linear Model Fitted to Explain the Probability of Fault Not Reoccurring based on the Probability of Fault Mitigation (PMIT), Understanding (UND), Whether or Not the Fault Mitigation Agreed Was Implemented (IMP), Whether or Not a Different Mitigation Strategy Was Implemented (OIT), Whether or Not the Mitigation Action Was Tested (MTE), if the Mitigation Proved Robust In the Field (ROB), the Effort Taken to Mitigate the Fault (EFF), and Whether or Not the Effort Was As Expected (EEP)

Variable	Mean	*SD*	5%	Median	95%	MCMC Error	MCMC Error/*SD*
α_0	2.871	2.502	−1.752	2.997	6.908	0.1194	0.0477
α_1 (PMIT)	−267	199.3	−652.2	−232.6	−5.708	9.643	0.0484
α_2 (UND)	511.7	699	−557.7	460.5	1731	32.84	0.0470
α_3 (IMP)	−502.7	375.6	−1229	−437.7	−10.53	18.16	0.0483
α_4(OIT)	447.9	754.9	−783.3	432.2	1682	35.28	0.0467
α_5(MTE)	0.8072	4.293	−4.03	0.1709	8.416	0.2256	0.0526
α_6(ROB)	551.2	739.9	−658.7	548.3	1779	34.43	0.0465
α_7 (EFF)	132.1	98.83	2.735	115	323.2	4.777	0.0483
α_8 (EEP)	−690.8	440.4	−1491	−650.8	−11.67	21.34	0.0485

The simulation was run for 100,000 samples to give a mcmc error of approximately 5% of the standard deviation (*SD*).

are not significant in the calculation of the probability of failure not reoccurring.

Based on the results presented by the GLM, the collinearity analysis, and analysis of the proportional coefficients, we rejected the hypothesis that for AUV missions under ice, the probability of fault not reoccurring could be estimated based on all the variables proposed in section 2.2.

An important finding was that the probability of failure not reoccurring was positively correlated with the probability of failure mitigation provided by the experts. Brito and Griffiths (2018) showed that experts overestimate the probability of failure mitigation. To our knowledge, there is no existing literature exploring this issue, which is particularly important for expert judgment elicitation. The results indicate that a potential reason for this is because the experts did not take into account the effort required to implement the failure mitigation in their assessments for the probability of failure mitigation. The reduced

Table IV. Correlation Analysis Between All Variables Presented In Equation 4

Correlations		PMIT	UND	IMP	OIT	MTE	ROB	EFF	EEP
PMIT	Pearson correlation	1	0.454^{**}	0.291	0.203	0.394^{**}	0.408^{**}	−0.290	−0.075
	Sig. (2-tailed)		0.002	0.052	0.180	0.007	0.005	0.053	0.624
UND	Pearson correlation	0.454^{**}	1	0.605^{**}	−0.125	0.497^{**}	0.632^{**}	-0.312^{*}	−0.222
	Sig. (2-tailed)	0.002		0.000	0.413	0.001	0.000	0.037	0.143
IMP	Pearson correlation	0.291	0.605^{**}	1	-0.419^{**}	0.467^{**}	0.339^{*}	−0.207	−0.150
	Sig. (2-tailed)	0.052	0.000		0.004	0.001	0.023	0.173	0.325
OIT	Pearson correlation	0.203	−0.125	-0.419^{**}	1	0.195	0.200	0.085	0.300^{*}
	Sig. (2-tailed)	0.180	0.413	0.004		0.199	0.189	0.578	0.045
MTE	Pearson correlation	0.394^{**}	0.497^{**}	0.467^{**}	0.195	1	0.565^{**}	−0.201	0.031
	Sig. (2-tailed)	0.007	0.001	0.001	0.199		0.000	0.186	0.840
ROB	Pearson correlation	0.408^{**}	0.632^{**}	0.339^{*}	0.200	0.565^{**}	1	−0.233	0.016
	Sig. (2-tailed)	0.005	0.000	0.023	0.189	0.000		0.124	0.919
EFF	Pearson correlation	−0.290	-0.312^{*}	−0.207	0.085	−0.201	−0.233	1	0.385^{**}
	Sig. (2-tailed)	0.053	0.037	0.173	0.578	0.186	0.124		0.009
EEP	Pearson correlation	−0.075	−0.222	−0.150	0.300^{*}	0.031	0.016	0.385^{**}	1
	Sig. (2-tailed)	0.624	0.143	0.325	0.045	0.840	0.919	0.009	

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).
The Pearson coefficient for each combination of variables, followed by the p value. the number of data points, n, is 45.

Table V. Generic Linear Model Fitted to Explain the Probability of Fault Not Reoccurring Based on the Probability of Fault Mitigation (PMIT), Whether or Not the Fault Mitigation Agreed Was Implemented (IMP), the Effort Taken to Mitigate the Fault (EFF), and Whether or Not the Effort Was As Expected (EEP)

Variable	Mean	*SD*	5%	Median	95%	MCMC Error	MCMC Error/*SD*
α_1 (PMIT)	11.2	8.42	2.93	8.42	31.82	0.457	0.0543
α_3 (IMP)	−1.10	5.64	−13.2	0.0808	5.64	0.295	0.0522
α_7 (EFF)	1.30	1.17	0.291	0.944	3.52	0.0609	0.0520
α_8 (EEP)	−1.91	6.07	−10.57	−2.01	9.36	0.316	0.0520

The simulation was run for 100,000 samples to give a MCMC error of approximately 5% of the standard deviation (*SD*).

model that attempts to capture the effect of both the probability of failure mitigation and of the effort for implementing it is presented in the following section.

3.2. Probability of Mitigation Model

Based on the analysis presented in the previous section, we were able to reduce the model proposed in Equation 4. This was achieved using a step-wise reversed regression. The reduced model is presented below in Equation 6:

$$\text{logit}\,(p_i) = \alpha_1 PMIT + \alpha_7 EFF + b_i \qquad (6)$$

Results presented in Table VI showed that it was possible to define a regression model for estimating the probability of fault not reoccurring, which took into account the probability of failure mitigation agreed by the experts and the effort in implementing the mitigation action.

The Pearson's correlation coefficient between PMIT and EFF was −0.290 ($p < 0.053$). There was no linear relationship between these two variables. The scatter plot showing proportional coefficients values of α_1 and α_7 is presented in Fig. 1. This supported the argument that there was not a relationship between PMIT and EFF because the correlation between the two proportional coefficients was -0.03266.

4. DISCUSSION

Risk analysis of autonomous systems deployment in extreme environments is highly dependent on the assessment provided by experts. Our results showed that expert judgments alone were not

Table VI. Generic Linear Model Fitted to ISE's Explorer Failure Mitigation Data to Explain the Probability of Fault Reoccurring Based On the Probability Of Fault Mitigation Agreed By the Experts (PMIT) and Effort (EFF)

Variable	Mean	*SD*	5%	Median	95%	MCMC Error	MCMC Error/*SD*
α_1 (PMIT)	6.67	3.27	3.23	5.98	13.1	0.174	0.0532
α_2 (EFF)	0.865	0.687	0.241	0.686	2.13	0.0370	0.0538

The simulation was run for 60,000 samples to give an MCMC error of approximately 5% of the standard deviation (*SD*).

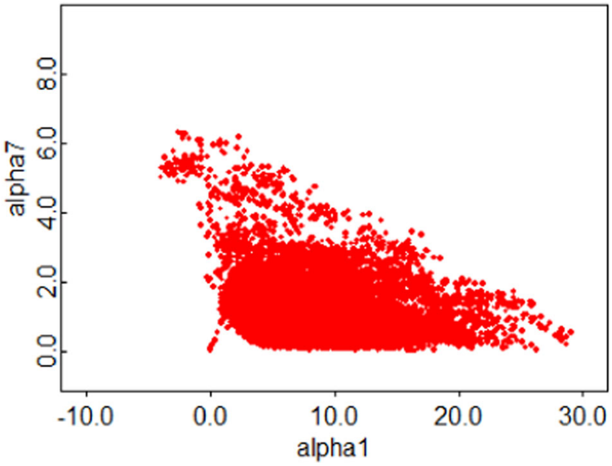

Fig 1. Scatter plot of the proportional coefficients of the generalized linear model presented in Equation 6.

effective at estimating the probability of a fault not reoccurring. The probability of fault not reoccurring was dependent on both the probability of failure mitigation agreed by the experts and the effort taken to implement the fault mitigation action. This insight is important because in such risk analyses it is often assumed that the probability of mitigation agreed by the experts is a static estimate and a proxy for the probability of failure not reoccurring (Brito et al., 2012; Brito, Griffiths, & Mowlem, 2012). Our results showed that these expert estimates can be updated using subsequent operational data. Specifically, for the two dependent variables (PMIT and EFF), we found that the probability of failure mitigation was more significant than the effort alone. Hence, the probability of mitigation estimated by experts can be used as a proxy for calculating the probability of a fault not reoccurring based on an assessment of the intended mitigation action. Expert judgment elicitation is applied in many domains but, to our knowledge, this is the first comprehensive study that has attempted to validate expert judgments in this operational context and using this novel modeling technique.

The successful deployment of critical systems in extreme environments depends on effective risk mitigation. However, the novel nature of these systems often means that there is a lack of past data regarding the factors that determine whether a mission is successful or unsuccessful. Consequently, managers, engineers, and scientists are often forced to rely on expert subjective judgements about specific risks and effective risk management actions. This practice is now adopted in many contexts. Given the potential for variation in the extent to which expert judgments on the probability of risk mitigation are reliable, it is important to assess the effectiveness of such judgments, even when only a small set of operational data is available. We achieved this by using a GLM technique that estimates the probability of fault not reoccurrence in light of the subsequent operation and the prior expert estimate on the probability of failure mitigation. Previous research has found that the probability of failure mitigation agreed by experts can contribute to reducing the risk of AUV loss by up to 24%; a reduction high enough to determine whether a mission is accepted or aborted. However, our analysis showed that the probability of fault reoccurrence is not only dependent on the probability of failure mitigation agreed by the experts, but also on the effort required to implement failure mitigation actions. This finding suggested that current approaches that are used for updating systems or missions risk (Brito & Griffiths, 2018) should better take into account the effort to implement the failure mitigation action in the estimation of the fault reoccurrence instead of only using the probability of failure mitigation agreed by the experts. Since effort is a significant variable in the estimation of the probability of fault not reoccurring, ignoring this variable can lead to an underestimation of the probability of fault not reoccurring.

There are number of ways in which effort can have such an impact on systemic and operational risk. For example, tight project delivery schedules often lead to pressures on time and resources that can substantially influence the amount of effort that is invested into risk mitigation actions. This can lead to “quick fixes” which can decrease reliability by not correcting the root cause for failure. From this perspective, the failure of experts to take into account the effort taken to implement the mitigation presents a variation of the “planning fallacy” observed in project duration estimation (Jørgensen, 2004). The planning fallacy is the tendency for individuals to continue to overestimate their own future performance in spite of evidence that they were overly optimistic in past estimates (Kahneman & Tversky, 1979). This fallacy may have been evident in our study because less effort than required was put into the failure mitigation actions. This also illustrates the benefits of “unpacking” an elicitation to help experts better understand and estimate the relevant components (Hill et al., 2000). Therefore, it may have been beneficial to change the elicitation so that it decomposed the mitigation implementation task into its related subtasks.

A reduction in the effort would lead to a reduction in the probability of fault not reoccurring. In other words, effort is linked with the quality of the task, but with fault correction this is not the case. Hence, an increase in effort can lead to an increase in the quality of the task carried out by those responsible for risk mitigation and, consequently, to an increase in the probability of a failure not reoccurring.

Our findings are important for several reasons. First, research has shown that, following a series of subsequent successes, organizations tend to “fine tune” risk management processes in order to exploit the opportunity to increase operational and financial efficiencies. However, this can reduce the probability of success to such an extent that it can have catastrophic consequences. For example, such fine tuning was reported to be one of the key factors that led to the NASA Challenger disaster in 1986 (Starbuck & Milliken, 1988). By contrast, the approach we have demonstrated here allows engineers to refine the risk model in order to develop a more accurate model of risk and, therefore, to avoid authorizing a risky mission or to cancel a mission with an acceptable risk level.

Second, the technique directly and pragmatically addresses the question of “how good are expert judgments?” This is a question faced by many researchers who facilitate and use expert judgment elicitations. One could argue that if experts were perfect, the number of variables that could determine success or failure could be reduced to an absolute minimum. The probability of fault reoccurring would then, for example, only depend on the probability of failure mitigation as agreed by the experts. Yet, our results show that to estimate the probability of a fault reoccurring, several variables must be considered, as must the effort to mitigate the risk. For some risk assessments, experts are not asked to assess both the probability of a hazard leading to catastrophic event *and* the probability of the mitigation actions being effective in reducing the likelihood of the hazard occurring (Brito et al., 2010). Here, in order to estimate the probability of a catastrophic event, our experts had to estimate the *a priori* risk and then build a mental model of the effectiveness of any potential hazard mitigation action. Our analysis shows that when doing this, experts should pay greater attention to the effort required to implement the mitigation action. This is particularly important in scenarios where there are several faults to mitigate (which, as in present case study, is a common occurrence) because the effort to mitigate these faults can become substantial. This can result in a failure to implement the mitigation action and, consequently, decrease the reliability of autonomous systems (Feather & Cornford, 2003). Our results suggest that “effort” is underweighted in expert probability estimates of risk mitigation in autonomous vehicle missions. This is potentially the case for other novel technologies with high levels of operational and systemic risk and, therefore, future similar projects should explore this problem in detail.

There has been a significant amount of research on expert judgments validation (Colson & Cooke, 2017). In risk assessment, the term cross-validation is used to assess experts’ judgment reliability with respect to a number of seed variables, where in a group of seed variables some seed variables are used to form a training set and the remaining seed variables are used to form a test set (Eggstaff, Mazzuchi, & Sarkani, 2014). Validation of expert judgments have also been conducted to estimate asset residual life (Wang & Zhang, 2008), project task duration (Hill et al., 2000), and to forecast product demand (Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2017). To our knowledge, there is currently no formalized mechanism for individuals and organizations to revisit and evaluate the probability of mitigation once this has been assessed. The method proposed in this paper allows experts to validate expert judgments with respect to mitigation actions. Specifically, the technique verifies the accuracy of the probability of failure mitigation and then uses that data to further refine the accuracy of the risk model. Hence, our technique for analyzing the accuracy of expert probability judgments could be utilized by project designers, engineers, scientists, and operators across a range of domains to better manage the risk of failure.

The central importance of risk mitigation and monitoring is commonly highlighted in risk management standards like that published by the institute of risk management (IRM, 2002). Our methodology provides a reliable means for retrospectively assessing expert estimates of the probability of mitigation. Moreover, our analysis shows that the experts in our case study would have needed to retune their assessments to take into account the required risk mitigation effort because their subjective assessments alone were not sufficient to accurately estimate the probability of specific faults reoccurring.

Researchers in other subject areas have concluded that the ideal number of experts for forecasting should be between four and five (Libby & Blashfield, 1978). In our study we did not attempt to quantify if the number of experts had an influence on the agreed judgment. We selected five experts because this has typically been accepted as good practice and, at the same time, it allowed us to cover all areas of expertise while reducing intercorrelation between experts (Jørgensen, 2004). Furthermore, although it is widely accepted that a group of experts cannot out-perform the best expert in the group (Clemen, 1989), the challenge can then become how to identify who is the best expert. In this study, we did not attempt to identify who was the best expert because we elicited the *agreed* judgment for the probability of failure mitigation and not individual judgments.

The SHELF expert judgment elicitation was conducted to obtain the probability of fault leading to AUV loss. Other group elicitation processes (e.g. DELPHI) could have been used to obtain the data required for this type of study. In this study, we did not attempt to quantify the impact of the expert judgment elicitation method on the reliability of experts' judgment for the probability of failure mitigation. Hence, future research could explore if group size and the type of expert judgment elicitation method have an impact on the reliability of probability of failure mitigation judgment.

We found that our sample of experts tended to underestimate the risks. This tendency might be explained by over confidence that stems from, optimism bias and confirmation bias. Optimism bias is a belief commonly held by individuals that they or something they control is less likely to experience a negative event (Weinstein, 1980).Confirmation bias is tendency for individuals to interpret, recall, or favor information that confirms their existing beliefs, while neglecting or ignoring information that contradicts these beliefs (Nickerson, 1998).

Future studies could specifically set out to test the potential for such overconfidence to lead to underestimations of risk in expert assessments of risk mitigation actions. If such bias were evident in expert judgments, researchers could evaluate the effectiveness of various methods (e.g., decomposition, fault trees) that have previously been found to reduce such bias (Fischhoff, Slovic, & Lichtenstein, 1978; Hora, Dodd, & Hora, 1993).

5. CONCLUSION

Autonomous underwater vehicles provide a means to explore uncharted and extreme environments, and the risk assessment for these missions is inevitably subject to epistemic uncertainty. The quantification of risk of a catastrophic event is, therefore, only possible by resorting to expert judgments and, in some fortunate cases, the analysis of small amounts of data. Nonetheless, such technology gives researchers a unique opportunity to test the feasibility of using expert subjective judgments to better assess and mitigate risk. In other domains, such as the nuclear industry, expert judgments have often been used to evaluate risk, but it has not always been possible to assess the effectiveness of these judgments (see Bonano et al., 1990).

In this article, we have focused on the effectiveness of expert judgments regarding the probability of failure mitigation. While we have identified that expert risk judgments are relevant to understanding the probability of failure reoccurrence, we also found that the effort needed to mitigate faults is an important consideration. Furthermore, we have identified that multiple linear regression models can be defined to estimate the probability of failure reoccurrence and that this does not appear to have been considered in existing literature. It is possible that these multiple linear regression models could be applied to any other novel technologies and domains in which there is only a modicum of historical data and where there is limited technical knowledge about the operational performance of autonomous devices in real-world conditions.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their very insightful comments. We also thank the experts who took part in the expert judgment elicitation conducted in Halifax and in Vancouver. This work was partly funded by the Natural and Environment Research Council Grant NE/I015647/1, “Tracking effectiveness of risk mitigation for autonomous underwater vehicles.”

REFERENCES

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting*, *33*(1), 298–313.

Banks, C. J., Brandon, M. A., & Garthwaite, P. H. (2006). Measurement of sea ice draft using upward-looking ADCP on an autonomous underwater vehicle. *Annals of Glaciology*, *44*(1), 211–216.

Bellingham, J. G., & Rajan, K. (2007). Robotics in remote and hostile environments. *Science*, *318*(5853), 5.

Bishop, R. (2000). Intelligent vehicle applications worldwide. *IEEE Intelligent Systems and Their Applications*, *15*(1), 78–81.

Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems*, *11*, 1–24.

Bonano, E. J., Hora, S. B., Keeney, R. L., & Winterfeldt, D. v. (1990). *Elicitation and use of expert judgment in performance assessment for high-level radioactive waste repositories*. Washington, DC: U.S. Nuclear Regulatory Commission.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Brito, M., & Griffiths, G. (2016). A Bayesian approach to predicting risk of loss during autonomous underwater vehicle missions. *Reliability Engineering & System Safety*, *146*(2), 55–67.

Brito, M. P., & Griffiths, G. (2018). Updating autonomous underwater vehicle risk based on the effectiveness of failure prevention and correction. *Journal of Atmospheric and Oceanic Technology*, *35*, 797–808.

Brito, M. P., Griffiths, G., & Challenor, P. (2010). Risk analysis for autonomous underwater vehicle operations in extreme environments. *Risk Analysis*, *30*(12), 1771–1788.

Brito, M., Griffiths, G., Ferguson, J., Hopkin, D., Mills, R., Pederson, R., & MacNeil, E. (2012). A behavioral probabilistic risk assessment framework for managing autonomous underwater vehicle deployments. *Journal of Atmospheric and Oceanic Technology*, *29*(11), 1689–1703.

Brito, M. P., Griffiths, G., & Mowlem, M. (2012). Exploring Antarctic subglacial lakes with scientific probes: A formal probabilistic approach for operational risk management. *Journal of Glaciology*, *58*(212), 1085–1097.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety*, *163*, 109–120.

Cox, I. J., & Wilfong, G. T. (2012). *Autonomous robot vehicles*. New York: Springer Science.

Dowdeswell, J. A., Evans, J., Mugford, R., Griffiths, G., McPhail, S. D., Millard, N., & Ackley, S. (2008). Autonomous underwater vehicles (AUVs) and investigations of the ice-ocean interface: Deploying the Autosub AUV in Antarctic and Arctic waters. *Journal of Glaciology*, *54*(187), 661–672.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke′s classical model. *Reliability Engineering & System Safety*, *121*, 72–82.

Feather, M. S., & Cornford, S. L. (2003). Quantitative risk-based requirements reasoning. *Requirements Engineering*, *8*(4), 248–265.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 330.

Goossens, L. H. J., Cooke, R. M., Hale, A. R., & Rodić-Wiersma, L. (2008). Fifteen years of expert judgement at TUDelft. *Safety Science*, *46*(2), 234–244.

Griffiths, G., Brito, M., Robbins, I., & Moline, M. (2009). Reliability of two REMUS-100 AUVs based on fault log analysis and elicited expert judgment. *Proceedings of the International Symposium on Unmanned Untethered Submersible Technology* (pp. 1–12), Durham, NH.

Hill, J., Thomas, L. C., & Allen, D. E. (2000). Experts’ estimates of task durations in software development projects. *International Journal of Project Management*, *18*(1), 13–21.

Hora, S. C., Dodd, N. G., & Hora, J. A. (1993). The use of decomposition in probability assessments of continuous variables. *Journal of Behavioral Decision Making*, *6*(2), 133–147.

IRM. (2002). A risk management standard. Retrieved from www.theirm.org/media/886059/ARMS_2002_IRM.pdf

Jenkins, A., Dutrieux, P., Jacobs, S. S., McPhail, S. D., Perrett, J. R., Webb, A. T., & White, D. (2010). Observations beneath Pine island glacier in west Antarctica and implications for its retreat. *Nature Geoscience Letters*, *3*, 468–472.

Jørgensen, M. (2004). A review of studies on expert estimation of software development effort. *Journal of Systems and Software*, *70*(1), 37–60.

Kahneman, D., & Tversky, A. (1979). Intuitive predictions: Biases and corrective procedures. *TIMS Studies in Management Science*, *12*, 313–327.

Kaminski, C., Crees, T., Ferguson, J., Forrest, A., Williams, J., Hopkin, D., & Heard, G. (2010). 12 days under ice – an historic AUV deployment in the Canadian High Arctic. *Paper presented at the Autonomous Underwater Vehicles (AUV), 2010 IEEE/OES Monterey, CA*.

Kastenberg, W. E. (1987). *Kastenberg committee report on draft NUREG 1150*. Livermore, CA: Lawrence Livermore National Laboratory.

Keeney, R. L., & Winterfeldt, D. v. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*, *38*(3), 191–201.

Kouts, H. J. C. (1987). *Methodology for uncertainty estimation in NUREG 1150*, Washington, DC: US Nuclear Regulatory Commission.

Kynn, M. (2008). The ‘heuristics and biases’ bias in expert elicitation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(1), 239–264.

Leveson, N. G., & Harvey, P. R. (1983). Analyzing software safety. *IEEE Transaction on Software Engineering*, *9*(5), 569–579.

Libby, R., & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Performance*, *21*(2), 121–129.

McCullagh, P., & Nelder, J. (1989). *Generalised linear models* (2nd ed). London, UK: Chapman & Hall.

McPhail, S. (2009). Autosub6000: A deep diving long range AUV. *Journal of Bionic Engineering*, *6*(1), 55–62.

Merkhofer, M. W. (1987). Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*(5), 741–752.

Morris, P. A. (1977). Combining expert judgments: A Bayesian approach. *Management Science*, *23*, 679–693.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester, UK: Wiley.

Oakley, J. E., & O'Hagan, A. (2010). SHELF: The Sheffield elicitation framework. Version 2.0. School of Mathematics and Statistics, University of Sheffield, Sheffield, United Kingdom. Retrieved from http://tonyohagan.co.uk/shelf

Otway, H., & von Winterfeldt, D. (1992). Expert judgment in risk analysis and management: Process, context, and pitfalls. *Risk Analysis*, *12*(1), 83–93.

Phillips, L. D., & Wisbey, S. J. (1993). *The elicitation of judgmental probability distributions from groups of experts: A description of the methodology and records of seven formal elicitation sessions held in 1991 and 1992*. Nirex Report NSS/R282, pp. 158, Oxfordshire, UK: Nirex.

Singh, H., Can, A., Eustice, R., Lerner, S., McPhee, N., Pizarro, O., & Roman, C. (2004). Seabed AUV offers new platform for high-resolution imaging. *Transactions of the AGU*, *85*(31), 294–295.

Starbuck, W. H., & Milliken, F. J. (1988). Challenger: Fine-tuning the odds until something breaks. *Journal of Management Studies*, *25*(4), 319–340.

Subramanian, S., Elliott, L., Vishnuvajjala, R. V., Tsai, W. T., & Mojdehbakhsh, R. (1996). Fault mitigation in safety-critical software systems. *Proceedings of the Ninth IEEE Symposium on Computer-Based Medical Systems*, (pp. 12–17), Michigan.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases *Science*, *185*(4157), 1124.

Wang, W., & Zhang, W. (2008). An asset residual life prediction model based on expert judgments. *European Journal of Operational Research*, *188*(2), 496–505.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*(5), 806–820.

Wright, G., Bolger, F., & Rowe, G. (2002). An empirical test of the relative validity of expert and lay judgments of risk. *Risk Analysis*, *22*(6), 1107–1122.

Yoerger, D. R., Jakuba, M., Bradley, A. M., & Bingham, B. (2007). Techniques for deep sea near bottom survey using an autonomous underwater vehicle. *International Journal Robotic Research*, *26*(1), 41–54.