A Behavioral Probabilistic Risk Assessment Framework for Managing Autonomous Underwater Vehicle Deployments

MARIO BRITO AND GWYN GRIFFITHS

National Oceanography Centre, Southampton, Hampshire, United Kingdom

JAMES FERGUSON

International Submarine Engineering, Port Coquitlam, British Columbia, Canada

DAVID HOPKIN

Defence R&D Canada–Atlantic, Dartmouth, Nova Scotia, Canada

RICHARD MILLS

International Submarine Engineering, Port Coquitlam, British Columbia, Canada

RICHARD PEDERSON AND ERIN MACNEIL

Defence R&D Canada–Atlantic, Dartmouth, Nova Scotia, Canada

(Manuscript received 21 December 2011, in final form 21 May 2012)

ABSTRACT

The deployment of a deep-diving long-range autonomous underwater vehicle (AUV) is a complex operation that requires the use of a risk-informed decision-making process. Operational risk assessment is heavily dependent on expert subjective judgment. Expert judgments can be elicited either mathematically or behaviorally. During mathematical elicitation experts are kept separate and provide their assessment individually. These are then mathematically combined to create a judgment that represents the group view. The limitation with this approach is that experts do not have the opportunity to discuss different views and thus remove bias from their assessment. In this paper, a Bayesian behavioral approach to estimate and manage AUV operational risk is proposed. At an initial workshop, behavioral aggregation, that is, reaching agreement on the distributions of risks for faults or incidents, is followed by an agreed upon initial estimate of the likelihood of success of the proposed risk mitigation methods. Postexpedition, a second workshop assesses the new data and compares observed to predicted risk, thus updating the prior estimate using Bayes' rule. This feedback further educates the experts and assesses the actual effectiveness of the mitigation measures. Applying this approach to an AUV campaign in ice-covered waters in the Arctic showed that the maximum error between the predicted and the actual risk was 9% and that the experts' assessments of the effectiveness of risk mitigation led to a maximum of 24% in risk reduction.

1. Introduction

An important category of missions for deep-diving long-range autonomous underwater vehicles (AUVs) is to carry out observations that cannot be conducted by ships or any other instruments. Unlike smaller AUVs, these large AUVs typically operate in uncertain environments, well beyond acoustic range, for example, under sea ice (Ferguson et al. 1999) or in complex seabed terrain (McPhail 2009). In the Arctic, gathering information from beneath sea ice has long had scientific (Dowdeswell et al. 2008; Nicholls et al. 2006; Jenkins et al. 2010) and military (Rothrock and Wensnahan 2007) drivers. AUVs are being increasingly used for polar missions, including for commercial (Kleiner et al.

Corresponding author address: Mario Brito, National Oceanography Centre, University of Southampton, European Way, Southampton SO14 3ZH, United Kingdom.
E-mail: mario.brito@noc.ac.uk

DOI: 10.1175/JTECH-D-12-00005.1

2011) and geopolitical purposes. The science community has recognized that there is a substantial risk of not completing missions and a risk of losing the vehicle when operating in polar seas (Griffiths et al. 2003), and it has developed mathematical methodologies to assess and manage those risks (Brito et al. 2010). However, when AUVs are to be used for data gathering in support of geopolitical or commercial missions, the risks of not completing missions or campaigns are likely to have greater impact and repercussions than for science missions. Consequently, the risk analysis and management processes need to be more robust, transparent, and defensible.

This paper proposes extending the framework devised previously for risk management of scientific AUV missions (Griffiths and Trembanis 2007). The motivation came from the decision by the Canadian Government, acting through Natural Resources Canada (NRCan), to use AUVs as part of its data-gathering program in the Arctic in support of its United Nations Convention on the Law of the Sea (UNCLOS) submission (Crees et al. 2010). Article 76 allows countries to extend their jurisdiction beyond 200 nmi from the baselines if it is demonstrated that the seabed and subsoil of the submarine areas are ''a natural prolongation of its land territory to the outer edge of the continental margin'' (United Nations Oceans and Law of the Sea 2011). Countries that aim to extend the territorial state beyond 200 nmi must provide evidence, based on detailed mapping of the seabed, that the seafloor is an extension of their continental shelf. Given the importance of the dataset, the cost of operations in the high Arctic, and the limited season, credible AUV risk management was required.

In the Griffiths and Trembanis (2007) framework risk management is informed by assessments provided by a group of experts, where the final assessment is one that represents the group judgment. Individual expert judgments can be aggregated mathematically or behaviorally to reach this group judgment. Previously, expert judgments have been mathematically aggregated using the linear opinion pool, where experts have been kept separate during the elicitation (Clemen and Winkler 1999; Griffiths et al. 2009; Brito et al. 2010). There are different schools of thought as to what is the best aggregation method; while some postulate the use of mathematical methods (e.g., Mosleh et al. 1987) others postulate the use of behavioral methods (e.g., Phillips 1999). Mathematical methods should be applied together with substantial calibration and sensitivity studies; the latter are needed to assess the impact of different methods and biases (Keeney and von Winterfeldt 1989; Winkler 1967).

In previous AUV risk assessments, for particular faults or incidents, experts provided judgments of a probability of fault leading to loss that spanned three orders of magnitude (Brito et al. 2010). Despite capturing each expert's estimate of uncertainty using quartiles, the large spread for some assessments remained a concern. It was not feasible to establish whether this spread reflected true uncertainty, or whether the experts' own specific experience or knowledge that was drawn upon to justify their assessments could withstand peer review. To deal with this potential bias issue, experts' judgments were aggregated mathematically into two separate groups, termed optimists and pessimists. It was left to the decision maker to decide which risk assessment to use.

For the last 60 yr there has been a flood of psychological experiments dealing with various aspects of man as an intuitive statistician (e.g., Tversky and Kahneman 1974; Kynn 2008). In probability assessment, bias may take many forms. Sources of biases are classified as motivational or cognitive; both are a result of conscious or subconscious adjustments in the experts' assessments. Motivational biases are introduced because the expert is motivated by his or her perceived personal rewards or concerns for various assessments. Cognitive bias may result from the use of particular modes of judgment, also denoted as mental shortcuts or heuristics. A number of heuristics have been identified, and the most well known are availability, adjustment or anchoring, and representativeness. In availability heuristics the probability assessment is based on the information that the expert either recalls or visualizes, with little attention to the specific details of the present circumstances. With adjustment and anchoring heuristics the most readily available information forms the initial basis for the assessment, and subsequent assessments result in adjustments from this basis. In representativeness heuristics the probability of an event is evaluated according to the degree to which it is considered representative of a population from which it has originated.

One way to reduce bias in expert assessments is by conducting a behavioral judgment elicitation. Here experts are kept in the same room, they debate their views on a particular fault or incident prior to the assessment of risk, and at the end they agree on a single risk distribution that represents the group's consensual assessment (O'Hagan et al. 2006). In most cases, experts may have specialized knowledge in particular fields. The interaction between experts has been shown to provide a better synthesis through their collective expertise (Spetzler and Sael von Holstein 1975).

In this paper we propose a behavioral risk assessment approach for creating a risk profile of an autonomous underwater vehicle in extreme environments. The framework, influenced by Bayesian concepts, allows for the risk assessment to be revisited postcampaign to (a)

assess the outcomes against predictions and (b) provide feedback to improve the collective knowledge of the experts. Bayesian concepts are used by the experts for computing a posteriori probability of loss based on the experts' assessments of the actual mission results. Here, the a posteriori is not calculated as a function of the product of the likelihood and the a priori, which can be implemented in a mathematical algorithm. The Bayesian concept is implicit in the experts' assessments as they estimate the a posterior in light of the campaign results and their previous assessments.

For autonomous underwater vehicles operational risk can be mitigated in two different ways; one way is by the introduction of a monitoring distance. A monitoring distance is particularly effective for long-range vehicles, where the vehicle is monitored for a certain amount of time while within acoustic range before committing to the mission (Brito et al. 2010). A second way to mitigate the operational risk is by removing the most significant risks out of the operational profile. These can be either hardware or software faults or human operator error. While the design phase of an AUV will have sought to reduce significant risks, faults inevitably emerge during testing and use. Previous AUV risk assessment exercises considered that a fault or risk was either mitigated or not mitigated; it was assumed that the design team was certain over each outcome (Griffiths et al. 2003). However, in most cases there is uncertainty over this assessment. For example, the design team may not have sufficient time to test a new component, system, or modification, or perhaps the failure's root cause has not been fully understood. Consequently, this paper presents an approach for updating the AUV risk profile based on the experts' probability assessments on the effectiveness of risk mitigation procedures.

The paper is organized as follows: section 2 outlines the behavioral risk assessment process followed by a summary of the fault history considered by the experts for the motivating example in section 3. Section 4 presents the results of this risk assessment exercise; derivation of the survival estimator used for creating the AUV risk profile forms section 5, which is then applied in section 6 to calculate the AUV operational risk for a set of under–sea ice missions in the Arctic. Original and observed mitigation effectiveness is compared in section 7, and section 8 presents our conclusions.

2. Behavioral expert judgment elicitation

The interpretation of risk can vary from application to application (Kaplan and Garrick 1981). For example, within the subject of oceanic and atmospheric technology, when Dance et al. (2010) attempt to quantify the probability of thunderstorm strike, risk is defined by the following two quantities: the scenario (thunderstorm) and the probability of occurrence. The consequence of a thunderstorm is not considered. On the other hand, Changnon (1999), attempting to quantify the risk of hail in the United States, considered the following three properties: the scenario (hail), the probability of hail in different areas, and the financial costs caused by hail. To avoid any ambiguity we shall now present the definition of risk used in this paper. Two different interpretations are considered—one for a static risk model and a second for a dynamic risk model. For the static risk model, formed by the faults and the experts assessments, risk is defined by the duplet (S, Li), where S is the scenario, in this case AUV loss, and Li is the subjective probability of a fault (F) leading to loss in a given environment (E), that is, $P(L|F, E)$. For the dynamic risk model, risk remains defined by the duplet (S, Li), while S remains the same; Li is now the probability of loss given a specific distance (d), $P(L|F, E, d)$ (Brito et al. 2010). The following subsections provide details of the risk assessment conducted for creating the static risk model.

a. Fault risk assessment

A number of expert judgment behavioral aggregation methods have been published in academic journals and research reports. The Delphi method (Dalkey 1969) is arguably the most referenced, but other approaches, such as the nominal group technique (Delbecq et al. 1975), the aggregation method of Kaplan (Kaplan 1990), and the Sheffield Elicitation Framework (SHELF) method (O'Hagan et al. 2006), are also widely referenced and applied. These methods vary in the form of interaction between experts and facilitators (face to face, mediated by a computer, or anonymously). Time is always at a premium in expert judgment elicitations. Because the aim in this instance was to assess the risk posed by a number of faults, a method that can be applied within an acceptable time limit while maximizing the amount of discussion between experts was adopted by following the SHELF proposed by O'Hagan et al. (2006).

The aim of the elicitation was to estimate the probability of a fault leading to loss in the given target environment. It is well accepted that experts will not always agree on a particular risk assessment, and this may result in differences in the experts' assessments. The SHELF elicitation process defines means for capturing these in a probability distribution. This rationale was pioneered by Phillips and Wisbey (1993) and further developed by O'Hagan (1998). Experts were asked to define the distribution using five parameters: the lower and upper bounds (denoted as L and U, respectively), median M, and lower and upper quartiles (denoted as LQ and UQ,

respectively). Experts were given training on probability assessment and on the concept of subjective probability and Bayesian theory. The steps of the behavioral expert judgment elicitation exercise are set out in the appendix.

b. *Effectiveness of the mitigation assessment*

Risk mitigation of operational faults can be achieved by introducing design changes or new operational procedures. Over the years organizations concerned about risk have developed processes, mostly simulation based, for assessing the impact of risk mitigations. For example, the National Aeronautics and Space Administration (NASA) developed the Defect Detection and Prevention (DDP) system for aiding decision making during the early phase of technology and system development (Feather and Cornford 2003). The DDP uses the assessment of the effectiveness of a risk mitigation strategy for updating the risk outcome of a defect. Here effectiveness of risk mitigation is used for addressing the reliability of the design. To our knowledge, no framework has been proposed for updating operational risk estimation based on the estimated effectiveness of risk mitigation strategies.

Risk mitigation procedures are not guaranteed to be successful, for example, the implementation might not be as simple as initially envisaged or time for operational testing may be too short. Expert assessment presents the formalism for quantifying the effectiveness of these procedures. The expert panel was asked to agree on the probability that the mitigation plan proposed by the design team would completely remove the fault. This assessment followed the risk elicitation described above, and it lasted approximately 5 h. The experts' assessments for the mitigation strategies were based on the information available. This was provided by two design and operational engineers who also attended the workshop.

Following a debate, experts were asked to agree on the assessment, assigning $p_{M_i} = 0$ if the consensus was that the proposed risk mitigation would not mitigate the fault, $p_{M_i} = 1$ if the consensus was that the proposed mitigation would completely mitigate the fault, or an intermediate value. Having estimated the effectiveness of the mitigation strategy the static risk model can be updated using the following:

$$P(\mathrm{Li} \mid F_i, E, M_i) = P(\mathrm{Li} \mid F_i, E)(1 - P_{M_i}), \qquad (1)$$

where $P(\mathrm{Li}|F_i, E, M_i)$ is the probability of loss given fault F_i, environment E, and mitigation strategy M_i. From this equation it is possible to see that if $p_{M_i} = 1$, then $P(\mathrm{Li}|F_i, E, M_i) = 0$ for fault F_i. Experts' assessments for p_{M_i} for all faults in the motivating example are presented in Table 1.

3. Motivating example: Explorer AUV fault history

Two Explorer class AUVs were delivered to NRCan for the 2010 Arctic campaign, both of which had been built by International Submarine Engineering (ISE; Crees et al. 2010). Only vehicle 5's data are used in this paper, vehicle 6 had too few runs for the statistics to be meaningful. Careful records had been kept of the 51 faults and incidents that emerged during the following five testing phases:

1) fabrication and assembly (May–September 2009), in which five faults emerged;
2) sea trials (8 September–12 October 2009), with 19 faults;
3) first homing and positioning trials (16 November–4 December 2009), with 14 faults;
4) second homing and positioning trials (14 January–28 February 2010), with 8 faults; and
5) mission testing (22 February–12 March 2010), with 5 faults.

The record for each fault or incident contained a description of the depth, if at sea; the time into the mission; the action taken by the AUV; a note on the possible impact, if the fault had happened under ice; the corrective action taken, or to be taken; and further comments where applicable. In addition, the phase of the mission deployment when the fault occurred was included to allow for state transition analysis, following the methods described in Brito and Griffiths (2011).

Taking broad categories, the top five types of failures were software (25%), maneuvering (16%), navigation (13%), control electronics (9%), and communications (7%). Payload, ballast, electrical, and homing each had approximately 5%; energy, electrical cabling, propulsion, and collision avoidance each had approximately 2%.

4. Outcomes of behavioral aggregation: Precampaign

This section illustrates and discusses the wide range of behavioral aggregation outcomes, from quickly reached unanimity to judgments with wide spans indicative of agreements to disagree. The aggregated outcomes are listed in Table 1, as the five parameters elicited, the 95% quantile obtained from fitting a beta distribution to the parameters, and the time into the mission the fault occurred. In practical terms, a risk-informed decision should not be based on the mean or median. The reason for this is because of the level of uncertainty associated with these estimates, which in the case of the median is 50%. As described in section 6 of this paper, in order to calculate

TABLE 1. Aggregated expert judgments for the 51 faults that emerged prior to the Arctic campaign.

Fault	Mission	L	LQ	M	UQ	U	95% quantile	Time (h)	Probability of mitigation
1	Assembly_1	1.0×10^{-5}	2.73×10^{-5}	4.0×10^{-5}	0.000 208	0.001	0.000 413	0	0
2	Assembly_2	0	1.68×10^{-8}	6.2×10^{-8}	2.3×10^{-7}	1.0×10^{-6}	6.43×10^{-7}	0	0.9
3	Assembly_3	0.0002	0.001 77	0.005	0.0234	0.1	5.36×10^{-2}	0	0.8
4	Assembly_4	0	1.68×10^{-8}	6.2×10^{-8}	2.3×10^{-7}	1.0×10^{-6}	6.43×10^{-7}	0	0.9
5	Assembly_5	0	1.68×10^{-8}	6.2×10^{-8}	2.3×10^{-7}	1.0×10^{-6}	6.43×10^{-7}	0	0.95
6	2_1	0	1.68×10^{-8}	6.2×10^{-8}	2.3×10^{-7}	1.0×10^{-6}	6.43×10^{-7}	6 h, 52 min	1
7	4_1	0	0	0	0	0	0	5 h, 29 min	1
8	4_2	0.1	0.176	0.216	0.4	0.5	0.48	5 h, 29 min	0.9
9	6_1	0.0001	0.000 87	0.001 46	0.0032	0.005	0.004 64	4 h, 10 min	0.1
10	6_2	1	1	1	1	1	1	4 h, 10 min	0.95
11	6_3	0	1.68×10^{-8}	6.2×10^{-8}	2.3×10^{-7}	1.0×10^{-6}	6.43×10^{-7}	4 h, 10 min	0.95
12	4_3	0.01	0.0487	0.134	0.28	0.75	0.504	5 h, 29 min	1
13	5_1	0.001	0.0046	0.013	0.0207	0.04	0.0362	3 h, 56 min	0.8
14	6_4	0.05	0.3	0.63	0.805	0.95	0.921	4 h, 10 min	0.75
15	6_5	0.1	0.363	0.546	0.73	1	0.901	4 h, 10 min	0.4
16	7_1	0.01	0.059	0.14	0.218	0.5	0.397	5 h, 42 min	0.95
17	8_1	0	0	0	0	0	0	4 h, 55 min	0
18	9_1	0.001	0.001 35	0.003	0.0093	0.05	0.0166	4 h, 56 min	0.9
19	10_1	0	0	0	0	0	0	5 h, 13 min	0.8
20	10_2	0.5	0.6	0.68	0.75	0.8	0.79	5 h, 13 min	0.4
21	18_1	0.001	0.0046	0.013	0.0207	0.04	0.0361	5 h, 51 min	0.8
22	18_2	1	1	1	1	1	1	5 h, 51 min	0
23	21_1	0	0	0	0	0	0	13 h, 43 min	0.9
24	17_1	0.001	0.0046	0.013	0.0207	0.04	0.0361	5 h, 16 min	0.95
25	23_1	0.02	0.052	0.08	0.111	0.2	0.167	4 h, 47 min	0.95
26	24_1	0	0	0	0	0	0	10 h, 39 min	0.9
27	24_2	0.001	0.0122	0.0458	0.077	0.2	0.176	10 h, 39 min	0.8
28	25_1	0.001	0.0046	0.013	0.0207	0.04	0.0361	8 h, 5 min	0.95
29	25_2	0	0	0	0	0	0	8 h, 5 min	1
30	25_3	0	0	0	0	0	0	8 h, 5 min	1
31	27_1	0.001	0.0018	0.0035	0.006 22	0.01	0.009 83	4 h, 17 min	0.9
32	27_2	0	0	0	0	0	0	4 h, 17 min	1
33	27_3	1	1	1	1	1	1	4 h, 17 min	0.6
34	28_1	0.4	0.49	0.61	0.7	0.8	0.798	3 h, 55 min	0.9
35	29_1	0.0001	0.000 87	0.001 46	0.0032	0.005	0.004 64	7 h, 28 min	0.1
36	31_1	0	0	0	0	0	0	7 h, 36 min	1
37	32_1	0.0004	0.004	0.012	0.048	0.2	0.109	7 h, 40 min	0.75
38	33_1	0.0004	0.0049	0.03	0.078	0.3	0.202	8 h, 2 min	0.5
39	35_1	0.001	0.004	0.01	0.0147	0.02	0.0189	5 h, 42 min	0.5
40	36_1	0.001	0.004	0.007 67	0.0117	0.02	0.0197	5 h, 21 min	0.5
41	36_2	1	1	1	1	1	1	5 h, 21 min	0.5
42	37_1	0.001	0.004	0.007 67	0.0117	0.02	0.0197	6 h, 7 min	0.5
43	38_1	0.001	0.005 67	0.0107	0.0157	0.02	0.0191	9 h, 29 min	0
44	40_1	1.0×10^{-5}	2.33×10^{-5}	4.67×10^{-5}	7.33×10^{-5}	0.0001	0.0009	5 h, 16 min	0.5
45	45_1	0.1	0.333	0.533	0.7	0.8	0.78	8 h	0.5
46	45_2	0.001	0.0393	0.163	0.257	0.5	0.451	8 h	0.5
47	46_1	1	1	1	1	1	1	5 h, 10 min	0.5
48	48_1	0.001	0.0046	0.013	0.0207	0.04	0.0361	62 h	0.95
49	49_1	0.1	0.333	0.533	0.7	0.8	0.78	4 h	0.5
50	50_1	0.1	0.333	0.533	0.7	0.8	0.78	60 h	0.5
51	50_2	0.001	0.003 17	0.005 33	0.007 33	0.01	0.009 47	60 h	0.1

AUV survival with distance, the agreed expert judgment must be integrated with a statistical survival estimator. We decided to use the 95% quantile of the agreed distribution because it reduces the amount of uncertainty that we must take into account in the final survival calculation.

One intrinsic part of the Bayesian elicitation process is that experts update their assessments based on arguments presented by other experts. In some cases the most conservative expert changes his assessments to agree with the most optimistic expert. For some other

experts' assessments we may see the opposite. This depends on the robustness of the arguments presented by each expert. Therefore, the 95% quantile of the agreed probability of loss distribution is not the result of the assessment provided by the most conservative expert.

a. Unanimity on faults that would inevitably lead to loss under ice

For five faults (10, 22, 33, 41, and 47) the experts concluded that there would be certain loss under ice, and set all parameters of the distribution to 1. Fault 10 was a time delay relay (TDR) malfunction. Based on the evidence that the vehicle had to be recovered as a ''dead vehicle,'' the experts agreed that if this fault occurred during an under-ice mission, more than 300 m away from an ice hole [the recovery remotely operated underwater vehicle (ROV) operating radius], then it would lead to certain vehicle loss. Faults that lead to the vehicle coming to a complete stop were considered to lead to certain loss. Fault 22 was a vehicle control computer (VCC) configuration fault that caused the vehicle to exit AUV mode and enter the stop mode for no apparent reason while the vehicle was in the water. No cause could be identified, but one of the experts (an Explorer user) had experienced a similar fault previously. A mission-planning error (fault 33) caused the vehicle to time out and stop. An unexpected release of the drop weight, for no apparent reason (faults 41 and 47), was also considered as leading to certain loss. This conclusion was reached quickly, but was followed by a long discussion on the possible causes as an aid to finding a mitigation strategy.

b. Unanimity on faults that would have no impact at all on survivability

For nine faults (7, 17, 19, 23, 26, 29, 30, 32, and 36) the collective conclusion was that there was zero probability of the faults leading to loss. In some cases this was because the component or subsystem that suffered the fault would not be present for the Arctic missions. This was the case for faults 7 (GPS antenna) and 17 (a VCC reboot when a radio modem connection is established). In another case it was because the fault occurred when the operating environment was so dissimilar to the ocean or the Arctic. Fault 19, a drop weight release failure on deck resulting from salt encrustation, would not have happened when in the water.

c. Faults where the phase of the mission may affect the consequence

A class of faults, typified by fault 14, a failure by the VCC to mount the hard drive at boot time, resulted in vigorous discussion by the experts on the probability of loss, because some experts considered the outcome to be strongly dependent on the phase of the mission during which the fault occurred. In this example, there were different views of the combination of this failure to mount the hard drive with the VCC's ability to restart, regain the mission, or return home, compounded with the unknown technical issues that caused the reboot in the first place. Consequently, the aggregated judgment arrived at through discussion showed a large span between a lower limit of 0.05 and an upper limit of 0.95.

d. Faults where individual experts shared particular insights affecting the aggregated outcome

A mistake in setting the VCC configuration (fault 12) gave the wrong sign to the attitude sensor's pitch rate, resulting in the vehicle porpoising on the surface. Individual experts gave different weighting in their considerations to possible extra power consumption from porpoising, thus depleting the battery supply, to possible impact with the bottom in shallow water or at low altitudes, and considered the unknown amplitude of the pitch oscillations from the fault. The resulting distribution with a lower limit of 0.01 and an upper limit of 0.75 reflected the experts' views on the range of behaviors that this fault could engender.

e. Agreement that the fault leads to a wide range of probability of loss

Fault 46 (a problem with a forward plane) was given agreed assessments that spanned three orders of magnitude. This wide range was not due to the need to encompass experts' disagreements but to their uncertainty concerning the outcome of the failure scenario, even after extensive discussion. For fault 46 they found it difficult to assess the risk without more information on vehicle performance following a single plane failure. Experts concluded that the risk distribution had to encompass a low risk tail, for the case where the failed plane would feather and control could be maintained using the functional planes, and a higher risk for when the plane stuck at an extreme angle, causing much higher drag and affecting control severely.

This was in contrast to those faults (13, 21, 24, 28, and 48) where it was clear that the planes failed into a feather configuration, for which the experts agreed quickly on a narrower, lower risk distribution.

f. Insights into instances of where a fault implied a consequential vulnerability

Discussion between the experts on consequential vulnerabilities arising from some faults proved valuable. Through discussion they were able to see beyond the immediate fault. For example, in fault 20 the Photonic Inertial Navigation System (PHINS) serial input failed

to accept inputs from the Doppler velocity log (DVL), GPS, and depth sensor. Nevertheless, the PHINS continued to provide position information. While internally the PHINS reported degraded performance there was no alarm to indicate this to the VCC. As a result, this was a nondetectable fault by the vehicle. The experts' discussions focused on the impact that the distance between the AUV when the fault occurred and the recovery point would have on the fault consequence. Experts argued whether the distance between AUV and the recovery point was sufficient for the induced error to exceed the range of the homing system, before concluding that this was a critical fault, with a lower limit of loss of 0.5 and an upper limit of 0.8.

g. Agreement to use heuristic shortcuts

Working as a group, the experts agreed collectively that they would spend little time on those faults that had a very low, but nonzero, consequence for the risk of loss. For these, they agreed on a standard distribution with a lower limit of 0, a median of 6.2×10^{-8}, and an upper limit of 10^{-6}. Examples are the two Network Time Protocol server failures (faults 2 and 11) because they only affected data logging, not command, control, or navigation, and internal problems with the acoustic survey instruments, again not influencing control (faults 4, 5, and 6). The small probabilities were considered to account for unimagined consequences.

5. Outcomes of mitigation assessments: Precampaign

The experts' collectively agreed probabilities of successful mitigation for the 51 faults presented as a histogram (Fig. 1) shows three separate distributions. One, with a mode at zero, represents those faults for which the experts agreed that the cause of the fault was unknown or unproven (such as for fault 22 described earlier), where there was no disagreement with the Explorer engineers. This distribution also covered faults where the experts were unconvinced that the proposed mitigation strategy would prove effective. This was the case for fault 24, where the database of controller parameters and configuration settings became corrupted. Despite changes to data management protocols, the fault recurred, and a further fix had been conceived but not tested.

The second distribution, with a sharp mode at 0.5, represents those faults the experts considered for which, although the proposed solution was appropriate, the mitigation strategy *had not been sufficiently tested or proven in field trials, or where a recurrence of a similar* fault could not be ruled out. As an example of the former, fault 20 with the PHINS navigation unit resulted in the unit being sent back to the manufacturer, a new motherboard and upgraded software was installed. However, because the unit was not tested, confidence in a true fix was not high. Instances of a fault where the particular occurrence was fixed, but there could be no guarantee that the fault, or a very similar fault, would not recur, were typified by human error or oversight. For example, fault 46, caused by a loose washer within a forward-plane controller. Software configuration faults can also fall into this class, such as fault 41, where the unintended consequence of a software change made prior to the mission changed a fault priority that resulted in the AUV stopping rather than continuing. Experts considered human error could occur again with configuration settings.

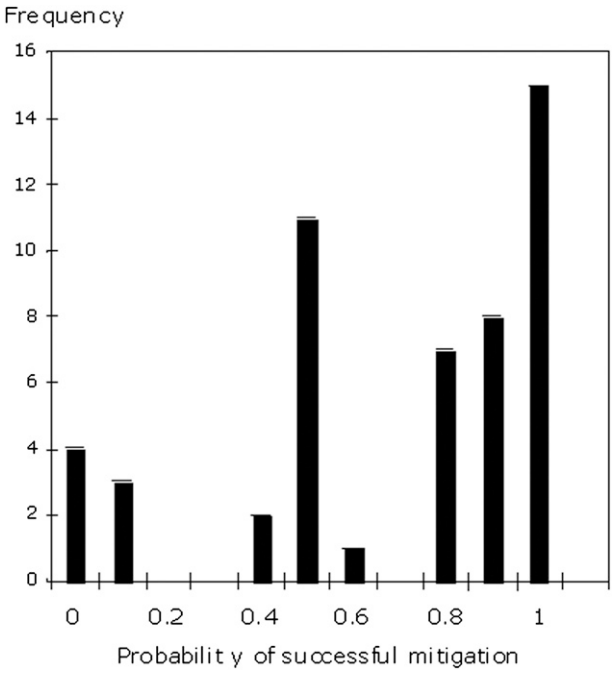

FIG. 1. Histogram of the assignments of probability of successful mitigation for the 51 faults, showing three distinct distributions.

The third distribution of probability of successful mitigation has a mode at over 0.9, indicating a high to very high level of confidence by the experts that the causes of faults were well understood, and that the solutions were known and tested. How well the solutions were tested affected the judgment; for example, for fault 13, the hydroplane not rotating correctly, resulting from an incorrect gauge in the locking washer, was assigned a mitigation probability of 0.8, which would increase with subsequent in-water trials with a washer of heavier gauge material. Certain success for mitigation, retiring the faults completely, was the

agreed outcome for those cases, such as faults 7 and 33, where the components or subsystems concerned would not be present for the Arctic missions. Also certain of mitigation were faults such as 12 (the incorrect sign of the PHINS pitch rate), where the cause was fully understood, straightforward, and a solution was implemented and tested. Some room for doubt, with a probability of success of 0.95, was assigned to those faults, while well understood and corrected could, with a small probability, recur. For example, fault 11, a missing battery in the network time protocol server was a human error, easily corrected, but with a nonzero chance of being repeated.

Combining the assessments on $P(L|F, E)$ with P_M identifies those faults where $P(L|F, E)$ is high but P_M is low. These form an important subset for the engineers to address. Most critical was fault 22, the VCC configuration problem, where $P(L|F, E)$ was 1 and P_M was 0. For all other faults where P_M was less than 0.1, $P(L|F, E)$ was less than 0.01 (Table 1); consequently, the need for effort into improving the understanding of the mitigation required was far less important. Of the 14 faults where $0.4 < P_M < 0.6$ eight were assessed with $P(L|F, E) > 0.5$ (Table 1). This was the most important set of faults for further investigation and improvement in P_M.

6. Survival estimator and confidence limits

The survival distribution for the ISE Explorer AUV was created using the extended version proposed by Brito et al. (2010) of the well-established Kaplan–Meier nonparametric model for estimating and displaying survival functions for small to medium samples of data. The estimator was first introduced by Kaplan and Meier (1958). Since then it has been applied in a wide variety of fields from medical statistics to systems failure analysis (Prentice and Kalbfleisch 2002). In failure analysis, the method uses historical data to compute systems survival as function of time. The historical data consist of failure time data and times of failure-free operations. The survival function is computed from the product of these two sources of data. The estimator uses a censor flag to specify whether at a given time it is considering failure data or survival data (nonfailure). In previous work we have shown that if the censor flag is replaced by probability of loss given that failure F_i emerges in environment E, allow us to calculate system survival for a given environment, this formulation is presented in Eq. (2).

For autonomous underwater vehicles we use distance instead of time as distance is proportional to time. Given a set of data comprising ordered mission ranges and whether each mission ended with a fault that has a probability of leading to loss $[P(L \mid F_i, E)]$, the survivor function $S(r)$ with range r is defined as

$$S(r) = \prod_{r_i<r}\left[1 - \frac{1}{n_i}P(L \mid F_i, E)\right], \tag{2}$$

where n_i is the number (of missions) at risk immediately prior to range r_i and is the number of losses at range r. Thus, the survival dataset will include all fault assessments, the distances at which they took place, and also all fault-free missions (Table 2).

TABLE 2. Fault-free missions prior to the Arctic campaign. Vehicle is assumed to travel at a speed of 1.5 m s^{-1}.

Mission	Time	Distance (km)
3	5 h, 17 min	28.5
11	13 h, 56 min	75.2
12	3 h, 29 min	18.8
13	4 h, 35 min	24.8
14	5 h, 53 min	31.8
15	5 h, 6 min	27.5
16	26 min	2.3
19	6 h, 4 min	32.8
20	6 h, 15 min	33.8
22	5 h, 28 min	29.5
26	11 h, 50 min	63.9
30	6 h, 32 min	35.3
34	5 h, 30 min	29.7
39	6 h, 42 min	36.2
41	9 h, 13 min	49.8
42	3 h, 43 min	20.1
43	6 h, 14 min	33.7
44	5 h, 12 min	28.1
47	5 h, 56 min	37.4

In the general case, the confidence limits for the extended Kaplan–Meier estimator are deduced from the variance in the dataset and the variance in the expert judgments (Brito et al. 2010). However, for this risk assessment, which considers the 95% quantile of the expert judgments obtained from fitting a beta distribution to the five parameters elicited from the experts (Table 1) and not the mean, the variance in the experts' assessments can be ignored. The 95% confidence limits for the estimator then becomes

$$\exp\{-\exp[c_+(r)]\} < \hat{S}(r) < \exp\{-\exp[c_-(r)]\}, \tag{3}$$

where

$$c_\pm(r) = \log[-\log\hat{S}(r)] \pm z_{\alpha/2}\sqrt{\hat{V}}, \tag{4}$$

and where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution; the 5% point was used, which is 1.96.

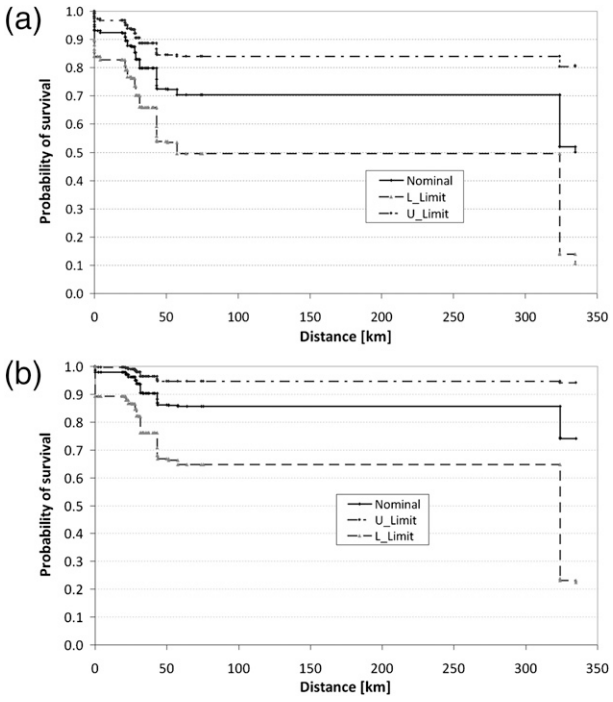

FIG. 2. (a) Kaplan–Meier probability of survival against distance for the dataset as considered by the experts with no mitigation of the individual faults and (b) with the assessment of individual fault mitigation included.

The variance is calculated using

$$V_{\mathrm{KM}} = \frac{1}{\log[S(r)^2]} \sum_{r_i < r} \frac{p_i}{n_i(n_i - p_i)}. \tag{5}$$

The survival distributions for the Explorer AUV (Fig. 2) show the probability of survival against distance without, and with, mitigation of faults for a single mission. Thus, for a 200-km mission without considering the effectiveness of mitigation strategies the probability of survival would be 0.7, increasing to 0.85 when the mitigation measures and their assessed effectiveness were taken into account. That is, the risk of loss is halved on this single mission. These estimates are without considering the use of a monitoring distance for further risk mitigation (Brito et al. 2010).

7. A priori campaign risk prediction

There were four types of missions planned for in the Arctic in 2010 as follows:

1) proving missions around the main camp ice hole (missions 51, 52, and 53);
   (a) a round-trip dive from the main camp to the southwest with the entire payload active (mission 51),
   (b) a transit out to approximately 56 km before returning back to the main camp using the short-range localization (SRL) system with no sensors active (mission 52), and
   (c) a 131-km round-trip mission out to a beacon camp to test the long-range homing system (mission 53);
2) one-way transit mission from the main camp to the remote camp (mission 54);
3) a survey of identified features of interest along the continental shelf (mission 55); and
4) one-way transit mission from the remote camp to the main camp (mission 56).

TABLE 3. All missions during the Arctic campaign, missions 53 and 55, were fault free. Vehicle is assumed to travel at a speed of 1.5 m s^{-1}.

Mission	Time	Distance (km)
51	5 h, 39 min	30.51
52	10 h, 20 min	55.8
53	24 h, 18 min	131.22
54	62 h, 16 min	336.24
55	60 h, 22 min	325.98
56	60 h, 5 min	324.45

The missions' distances covered by the vehicle during the Arctic mission were different from those specified prior to the campaign. Consequently, in order to compare similar quantities, we use the prior risk model to estimate the mission risk of the actual mission distances instead of the planned mission distances. See Table 3 for details.

Brito et al. (2010) showed how monitoring an AUV for a distance d, over which the vehicle could be recovered if a fault developed, could increase the probability of survival over the actual mission distance R. Let the probability of loss for the monitoring distance be $P(d)$ and the probability of loss for the mission range be $P(R)$, then the conditional probability expression leads to

$$P(X < R \mid X > d) = \frac{P(X = R) - P(X = d)}{1 - P(X = d)}, \tag{6}$$

where $P(X < R|X > d)$ is the probability of loss for target mission range R given that the vehicle has survived monitoring distance d. Expression (6) can be derived by manipulating the joint probability function of two statistically dependent events in which $P(X < R, X > d) = P(X < R|X > d) \times P(X > d)$, where the term on the left-hand side equals $P(X = R) - P(X = d)$. The probability of loss for distance greater than d is $P(X > d) = 1 - P(X = d)$. The probability of survival is the complement of the probability of loss. If we denote

the probability of survival as S, then $S(X = d) = 1 - P(X = d)$.

The choice of the distance d is informed by the Kaplan–Meier plots in Fig. 2, balancing a sufficient distance to enable those faults that have historically emerged at short distances against the increasing overhead of choosing a larger distance.

a. Proving missions around the main camp ice hole

In the initial analysis, we considered a monitoring distance of 20 km. However, in the actual deployment the monitoring distance was 31 km (for which the probability of survival is 0.9042); this was, in fact, mission 51.
For mission 52 the probability of survival becomes

unmitigated: $P(\text{surv}) = 0.722$,
mitigated: $P(\text{surv}) = 0.860$, and
mitigated + monitor 31 km: $P(\text{surv}) = 0.96$;

and for mission 53

unmitigated: $P(\text{surv}) = 0.704$,
mitigated: $P(\text{surv}) = 0.856$, and
mitigated + monitor 31 km: $P(\text{surv}) = 0.947$.

b. The survey missions

The product rule can be used to get at the overall probability of surviving three such missions. The overall $P(\text{survival})$ is simply the product of the individual probabilities,

$$P(\text{surv}) = \prod_{i=1}^{n} P(\text{surv})_i. \tag{7}$$

Hence, for this example set of three missions, overall,

unmitigated: $P(\text{surv}) = p_1(d = 324\ \text{km}) \times p_2(d = 326\ \text{km}) \times p_3(d = 336\ \text{km})$
$0.704 \times 0.518 \times 0.50 = 0.182$,
mitigated: $P(\text{surv}) = p_1(d = 324\ \text{km}) \times p_2(d = 326\ \text{km}) \times p_3(d = 336\ \text{km})$
$0.856 \times 0.741 \times 0.74 = 0.469$, and
mitigated + monitor 31 km:
$P(\text{surv}) = p_1(d = 324\ \text{km}) \times p_2(d = 326\ \text{km}) \times p_3(d = 336\ \text{km})$
$0.946 \times 0.82 \times 0.819 = 0.635$.

8. Postcampaign risk assessment

In the previous section we used the extended version of the Kaplan–Meier estimator to quantify the probability of survival with distance predicted from the prior history of faults and incidents. In this section we compare this prediction with the probability of survival derived using exactly the same methodology, but based only on the faults and incidents observed during the deployment. At a second risk assessment workshop the criticality of 17 faults that emerged during the 2010 Arctic campaign were assessed. These faults fell into the categories of control and electronics (17%); navigation (17%); payload (17%); software (12%); and ballast, communications, electrical and cabling, and mission planning (6%) each.

a. Expert judgments on emerged faults

During the expert judgment elicitation a similar pattern to that observed in the first workshop was seen. For some faults, experts quickly reached an agreement, for others a longer debate was required.

For five faults, experts assigned zero to all parameters of the probability of loss distribution. These consisted mainly of payload faults, where experts considered that while these would have an impact on science data gathering, they would pose no risk to the vehicle's safety. Also in this category was fault 3, a VCC configuration fault that meant that engineers could not put the vehicle into mission mode acoustically. This was fixed at the time of the deployment by increasing the telemetry item size in the configuration file. Experts considered this real-time fix to be an important factor, certain of correcting the problem.

In the previous workshop, experts identified a class of faults that had low impact. For these faults they agreed on a probability of loss distribution that was low but nonzero. In this assessment experts did not explicitly reach such an agreement and distribution for this type of fault. Nevertheless, the assessments for faults 12 and 15 showed that such distributions were used in this workshop unconsciously rather than through deliberate agreement. Fault 1, a ground fault on the variable ballasting system, was also considered low impact, but here experts were not so quick to reach agreement.

The agreed distributions for other faults fell into one of three shapes. First are the distributions that were skewed to the left (low probability), with a long tail toward high probability. Experts agreed on such distribution for faults, 2, 8, 9, 10, and 16. These are in general faults that would not result in immediate vehicle loss but would degrade the vehicle's safety with time. For example, fault 2, a CTD sensor failure, resulted from a crack on the sensor that could lead to degraded sonar data with a possible consequential degradation of navigation accuracy. The same rationale was adopted for fault 8, a bottom-avoidance altimeter ground fault, and fault 9, a Mimosa (mission planning software) distance estimation inaccuracy. Fault 10 was a VCC configuration failure that resulted in an under-depth fault response during the mission. Fault 16

TABLE 4. Aggregated risk assessments for the 17 faults that emerged during the arctic campaign. No faults emerged during missions 53 and 55. Faults and their assessments are presented in the order that they were assessed by the expert panel. The two successful fault free missions were added at the end of the table.

Fault	Mission	*L*	LQ	*M*	UQ	*U*	Distance (km)	95% quantile
1	—	0	1.0×10^{-7}	1.00×10^{-6}	1.00×10^{-5}	0.0001	0	4.48×10^{-5}
2	—	0.000 01	5.00×10^{-5}	0.0001	0.0005	0.005	0	0.00106
3	—	0	0	0	0	0	0	0
4	—	0.8	0.85	0.9	0.95	1	0	0.977
5	51	0	0	0	0	0	1.35	0
6	51	0.001	0.005	0.01	0.05	0.1	0.09	0.0933
7	52	0.0001	0.0009	0.003	0.009	0.05	32.4	0.0209
8	52	3.00×10^{-6}	6.00×10^{-5}	0.000 16	0.0008	0.01	48.6	0.001 98
9	—	0.0001	0.0001	0.0003	0.0009	0.01	0	0.001 70
10	—	5.00×10^{-6}	1.00×10^{-5}	0.0005	0.001	0.008	0	0.004 22
11	—	0	0	0	0	0	0	0
12	—	1.00×10^{-7}	2.00×10^{-7}	7.00×10^{-7}	2.00×10^{-6}	1.00×10^{-5}	0	4.13×10^{-6}
13	—	0	0	0	0	0	0	0
14	—	0	0	0	0	0	0	0
15	—	1.00×10^{-7}	2.00×10^{-7}	7.00×10^{-7}	2.00×10^{-6}	1.00×10^{-5}	0	4.13×10^{-6}
16	56	2.50×10^{-5}	0.000 25	0.0025	0.0125	0.025	324.45	0.0225
17	54	0.001	0.01	0.04	0.082	0.1	334.8	0.0964
	53	0	0	0	0	0	131.22	0
	55	0	0	0	0	0	325.98	0

was a lower transducer failure at the end of the Arctic mission at the main camp, just before recovery. The upper bound for all of these distributions was assigned to capture the most critical consequence.

Second, the distribution for fault 4 is skewed to the right, but here experts agreed that the whole distribution should be defined by high probability values. Fault 4 was a VCC configuration fault, where the VCC was configured so that the PHINS was power cycled when the ACE restarts. This would cause loss of PHINS alignment with every VCC reboot. Given the high likelihood of a reboot, experts considered that the probability of loss should be high, because PHINS alignment is a lengthy process at high latitudes.

Finally, there are the normally distributed expert judgments. This is the case for faults 6, 7, and 17. Of these, two were ground faults in critical components. Fault 6 was an acoustic modem ground fault, and fault 7 was a main bus 48-V ground fault. Fault 17 was a communications fault, with no acoustic command at the remote camp. The lower bound may be justified by the fact that these faults did not result in vehicle loss during the deployment; the vehicle was still capable of finding its way to the recovery point. The upper bound reflects the fact that these faults, if compounded with other failures, for example, failure to detect them, would very likely result in vehicle loss.

b. Actual risk during the Arctic campaign

The simplest way to compare two groups of survival data is to plot the corresponding survival distributions on the same axes. However, in this case, it would not support our analysis because we must take into account the effect of the monitoring distance, and this is not explicit in the survival plot. Thus, here, our comparisons are based on single-point observations.

Similarly to our previous analysis the data consist of fault assessments and distances of missions with no faults at all (missions 53 and 55). The experts' judgments (Table 4) form the basis for the extended Kaplan–Meier survival plot (Fig. 3). Having derived this distribution, the probability of survival based on the actual Arctic missions can be calculated and compared to the a priori–estimated risk based on trials data. The differences

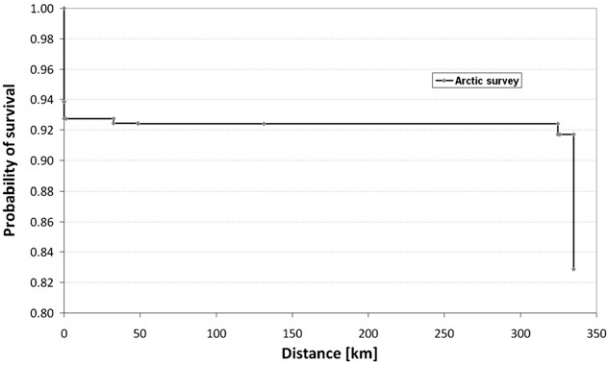

FIG. 3. Kaplan–Meier probability of survival for the AUV based on actual faults and incidents that occurred during the April 2010 Arctic campaign. This should be compared with the distribution for the a priori dataset, including the effect of mitigation effectiveness in Fig. 2b.

TABLE 5. Estimated operational risk for six missions. Numbers are approximated to the third most significant figure.

No.	Objective	Distance (km)	Probability of loss				
			Unmitigated	Mitigated	Mitigated + 30.51 km monitoring as mitigated	Observed risk	Difference
51	Monitoring distance	3	0.173	0.090	—	0.072	
52	Test mission 1	55	0.278	0.140	0.040	0.077	0.037
53	Test mission 2	131	0.296	0.144	0.005 33	0.077	0.0717
54	Survey mission 1	336	0.500	0.260	0.181	0.172	0.009
55	Survey mission 2	326	0.481	0.259	0.180	0.083	0.097
56	Survey mission 3	324	0.296	0.144	0.054	0.077	0.023

between the a priori–estimated and the campaign risk are presented in Table 5. Results show that the maximum error between the estimated a priori risk after mitigation effectiveness was taken into account and the campaign risk was 10%. Both of the risk profiles against mission distance, a priori umitigated (Fig. 2b) and observed (Fig. 3), show an approximately 10% reduction in the probability of survival near the maximum distance. It is important to note that these are independent datasets, based on the trials and the Arctic mission datasets, respectively. Consequently, the mission distances are not the same, and as steps in the Kaplan–Meier plots take place at the mission distances, it is only to be expected that there can be instances where the difference between the two risk profiles may be sensitive to small differences in mission distance. This becomes more pronounced at the longer distances, where there have been fewer missions and the denominator n_i in Eq. (2) is therefore small. For the a priori, mitigated profile, the 10% risk reduction takes place at 324 km, whereas for the observed risk the 10% step takes place at 334.5 km, resulting in the 10% difference. The minimum error was 0.9%. The differences between the a priori unmitigated risk of loss and the campaign risk were far greater. This emphasizes the importance of taking mitigation into account when projecting forward risk based on trials results when a period of addressing and correcting faults occurs prior to the actual deployment, otherwise the campaign risk may be overestimated.

9. Discussion and conclusions

This paper addresses a key issue in any critical system deployment, which is how to quantify the operational risk when deploying a system in hazardous environments. Marine and atmospheric scientists attempt to address these questions prior to the deployment of any critical system. This assessment is generally based on engineering judgment and in some cases may be informed by failure statistics. Good practice is to use formal methods to elicit the necessary judgments, allowing the process to be transparent, capable of replication, and minimize bias that can be consciously or subconsciously introduced by experts. Previous use of formal judgment elicitation for AUV risk assessment applied mathematical methods to aggregate the expert judgments into a single judgment. However, as discussed in this paper and in the referenced work, mathematical methods allow bias to be introduced. Behavioral expert judgment elicitation encourages experts to discuss and then agree on the risk assessment resulting in a more informed assessment from the group. The method presented in this paper captures the experts' judgment in a probability distribution. The agreed expert judgment distributions are unimodal, but can take a number of shapes reflecting the arguments that underpin them. The benefit of obtaining a distribution rather than a single-point assessment is that we can measure the confidence in the risk estimation. Subsequence assessments can be carried out based on the 95% quantile rather than the mean.

After finding a fault one cannot assume that the mitigation action will be completely effective and that the fault will be completely removed from the system. This will depend on a number of factors, such as fault understanding and the intensity and efficacy of testing. Using expert judgment to capture the confidence that the mitigation plan would completely remove the fault has been shown to provide realistic updates of the initial risk estimate.

Validating risk models has always been a concern for any system developer and user. This is particular true for systems where a catastrophic event does not occur during its lifetime. In conventional systems engineering, validation comprises comparing the model estimates against results obtained from field testing. Although risk models are based on expert judgments, a similar rationale applies. Having conducted two workshops, the first prior to an Arctic deployment (covering 51 faults) and the second after the deployment (covering 17 faults), the maximum difference in risk estimates was 10%, indicating an acceptable level of repeatability and demonstrating that the a priori estimate was a good predictor of near-term risk during actual operations after accounting for the effectiveness of mitigation.

APPENDIX

The Behavioral Elicitation Process

The judgment elicitation took place in two separate workshops, the first held in Halifax, Nova Scotia, Canada, from 8 to 10 December 2009, and the second was held in Vancouver, British Columbia, Canada, from 21 to 23 July 2011. The assessments provided on the second workshop were used to validate the assessments provided in the first.

a. Expert selection

A preelicitation meeting was organized by the project stakeholders and the facilitators. The aim was to define the scope of the elicitation exercise, to select experts, and to set out the seed questions that would be used to familiarize the experts with subjective probability. Experts were selected based on their experience in AUV operations, with an emphasis on the ISE Explorer vehicle, but also with a wider representation to provide another perspective. The experts for the first workshop were as follows:

1) Chris Kaminski, from the International Submarine Engineering in Canada: With 18 yr of experience in AUVs, he was part of the Theseus AUV team and had spent three seasons in the Arctic.
2) Jean-Marc la Framboise, from the International Submarine Engineering in Canada. Currently an AUV program manager, he had been involved in AUV development since 1982, working as project manager for 10 AUV development programs. His background is in electrical engineering.
3) Jan Opderbecke, from the Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) in France: He serves as the program manager for a fleet of Explorer AUVs.
4) Jeff Williams, from the University of Southern Mississippi: He is an operations engineer with a background in mechanical, electrical, and software engineering, and has operational experience of an Explorer AUV.
5) Steve McPhail, from the National Oceanography Centre in the United Kingdom. He has 22 yr of AUV experience, designing, building, and operating the Autosub family of vehicles.

For the second workshop Jeff Williams was unavailable, and was substituted by Richard Pedersen from the Defense Research and Development of Canada. He was familiar with the elicitation process, because he was one of the observers in the first risk assessment exercise and had experience as a project manager of AUV campaigns.

b. Preelicitation briefing

The preelicitation briefing note was distributed to experts. The document contained five papers providing an introduction into the elicitation method and details of the SHELF package and supporting literature; the background information for the elicitation; a list of attendees, experts, and all other stakeholders; the schedule for the workshop; extracts of bathymetric charts and example ice coverage charts relevant to the study.

c. Training the experts

In previous studies seed variables have been used to

1) measure expert performance,
2) enable performance-based weighted combination of experts' distributions, and
3) evaluate, and to an extent validate, the aggregated output.

Points 2 and 3 remain controversial; the best way to combine expert judgment based on how well they perform in seed questions is not settled (Cooke and Goossens 2004). In this study two seed questions were used to train experts and make them aware of the fallacies of expert judgment assessments. The first question was, "What is the probability of losing an AUV in an under ice shelf mission of >10 km?"; the second question was, "What is the probability of an AUV abort during missions under sea ice?" The facilitators knew the answers to the seed questions from frequentist statistics, not from judgment. These real facts allowed the facilitators to check that the experts' judgments were realistic. Figure A1 presents the assessment for the two seed questions. The agreed median for question 1 at 0.064 was lower than the actual frequency of loss of 0.22, because two AUVs have been lost in nine missions under shelf ice. The 95% quantile of the agreed distribution was 0.35 above the actual frequency of loss; thus, at the 95% level of confidence that the actual probability of loss would not be above the estimate from the experts, their view agreed with actual loss statistics to date.

d. Eliciting the five parameters of a probability distribution

First, the plausible range was established by reaching agreement collectively through discussion on the lower and upper bounds such that it was extremely unlikely, but not necessarily impossible, that the probability of the fault leading to a loss in the described environment lay outside these bounds. Second, each expert working alone, without discussion, estimated the median and then the lower and upper quartiles. After discussion of the distributions arrived at individually by the experts,

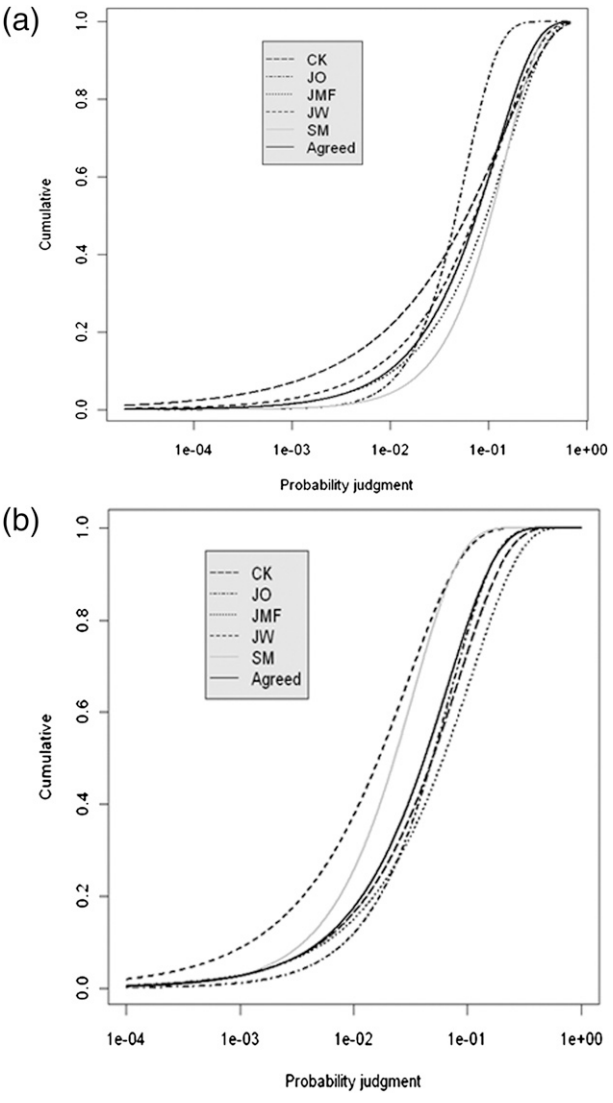

FIG. A1. Assessments for seed questions (a) 1 and (b) 2.

and sharing of knowledge and reasoning about the differences, the experts reached consensus values for the median and quartiles.

REFERENCES

Brito, M. P., and G. Griffiths, 2011: A Markov chain state transition approach to establishing critical phases for AUV reliability. *IEEE J. Oceanic Eng.,* **36,** 139–149.

——, ——, and P. G. Challenor, 2010: Risk analysis for autonomous underwater vehicle operations in extreme environments. *Risk Anal.,* **30,** 1771–1788.

Changnon, S. A., 1999: Data and approaches for determining hail risk in the contiguous United States. *J. Appl. Meteor.,* **38,** 1730–1739.

Clemen, R. T., and R. L. Winkler, 1999: Combining probability distributions from experts in risk analysis. *Risk Anal.,* **19,** 187–203.

Cooke, R. M., and L. H. J. Goossens, 2004: Expert judgment elicitation for risk assessments of critical infrastructures. *J. Risk Res.,* **7,** 643–656.

Crees, T., and Coauthors, 2010: UNCLOS under ice survey—An historic AUV deployment in the Canadian high arctic. *Proc. OCEANS 2010,* Seattle, WA, IEEE/MTS, 1–8.

Dalkey, N. C., 1969: The Delphi method: An experimental study of a group of opinions. The Rand Corporation Rep. RM-5888-PR, 87 pp. [Available online at http://www.rand.org/content/dam/rand/pubs/research_memoranda/2005/RM5888.pdf.]

Dance, S., E. Ebert, and D. Scurrah, 2010: Thunderstorm strike probability nowcasting. *J. Atmos. Oceanic Technol.,* **27,** 79–93.

Delbecq, A. L., A. H. Van de Ven, and D. H. Gusafson, 1975: *Group Techniques for Program Planning.* Scott Forester, 174 pp.

Dowdeswell, J. A., and Coauthors, 2008: Autonomous underwater vehicles (AUVs) and investigations of the ice-ocean interface: Deploying the Autosub AUV in Antarctic and Arctic waters. *J. Glaciol.,* **54** (187), 661–672.

Feather, M., and S. Cornford, 2003: Quantitative risk-based requirements reasoning. *Requir. Eng.,* **8,** 248–265.

Ferguson, J., A. Pope, B. Butler, and R. I. Verrall, 1999: Theseus AUV—Two record breaking missions. *Sea Technol.,* **40,** 65–70.

Griffiths, G., and A. Trembanis, 2007: Towards a risk management process for Autonomous Underwater Vehicles. *Masterclass on AUV Technology for Polar Science,* G. Griffiths and K. Collins, Eds., Society for Underwater Technology, 103–118.

——, N. W. Millard, S. D. McPhail, P. Stevenson, and P. G. Challenor, 2003: On the reliability of the Autosub autonomous underwater vehicle. *Underwater Technol.,* **25,** 175–184.

——, M. Brito, I. Robbins, and M. Moline, 2009: Reliability of two REMUS-100 AUVs based on fault log analysis and elicited expert judgment. *Proc. Int. Symp. on Unmanned Untethered Submersible Technology,* Durham, NH, AUSI. [Available online at http://eprints.soton.ac.uk/69184/1/UUST09_Griffiths_et_al_Reliability_of_REMUS%5B2%5D.pdf.]

Jenkins, A., P. D. Stanley, S. Jacobs, S. D. McPhail, J. R. Perrett, A. T. Webb, and D. White, 2010: Observations beneath Pine Island Glacier in West Antarctica and implications for its retreat. *Nat. Geosci.,* **3,** 468–472.

Kaplan, E. L., and P. Meier, 1958: Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.,* **53,** 457–481.

Kaplan, S., 1990: 'Expert information' versus 'expert opinions': Another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Reliab. Eng. Syst. Saf.,* **35,** 61–72.

——, and B. J. Garrick, 1981: On the quantitative definition of risk. *Risk Anal.,* **1,** 11–27.

Keeney, R. L., and D. von Winterfeldt, 1989: On the uses of expert judgment on complex technical problems. *IEEE Trans. Eng. Manage.,* **36,** 83–86.

Kleiner, A., J. Cheramie, J. Dean, and R. Raye, 2011: Ice class AUV development. *Proc. Arctic Technology Conf.,* Houston, TX, Offshore Technology Conference, 22121. [Available online at http://www.cctechnol.com/uploads/IceClassAUVDevelopment.pdf.]

Kynn, M., 2008: The 'heuristics and biases' bias in expert elicitation. *J. Roy. Stat. Soc.,* **171A,** 239–264.

McPhail, S., 2009: Autosub6000: A deep diving long range AUV. *J. Bionics Eng.,* **6** (1), 55–62.

Mosleh, A., V. Bier, and G. Apostolakis, 1987: A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliab. Eng. Syst. Saf.,* **20,** 63–85.

Nicholls, K. W., and Coauthors, 2006: Measurements beneath an Antarctic ice shelf using an autonomous underwater vehicle. *Geophys. Res. Lett.,* **33,** L08612, doi:10.1029/2006GL025998.

O'Hagan, A., 1998: Eliciting expert beliefs in substantial practical applications. *Statistician,* **47,** 21–35.

——, C. E. Buck, A. Daneshkhah, J. E. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, 2006: *Uncertain Judgments: Eliciting Expert Probabilities.* Wiley, 328 pp.

Phillips, L. D., 1999: Group elicitation of probability distributions: Are many heads better than one? *Decision Science and Technology: Reflections on the Contributions of Ward Edwards,* J. Shanteau, B. Mellors, and D. Schum, Eds., Kluwer Academic, 313–330.

——, and S. J. Wisbey, 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 Rep. NSS/R282, 158 pp.

Prentice, R. L., and J. D. Kalbfleisch, 2002: *The Statistical Analysis of Failure Time Data.* Wiley, 462 pp.

Rothrock, D. A., and M. Wensnahan, 2007: The accuracy of sea ice drafts measured from U.S. navy submarines. *J. Atmos. Oceanic Technol.,* **24,** 1936–1949.

Spetzler, C. S., and C.-A. Sael von Holstein, 1975: Probability encoding in decision analysis. *Manage. Sci.,* **22,** 340–358.

Tversky, A., and D. Kahneman, 1974: Judgment under uncertainty: Heuristics and biases. *Science,* **185,** 1124–1131.

United Nations Oceans and Law of the Sea, 2011: Part VI: Continental shelf. Laws of the Sea Conventional Agreements. [Available online at http://www.un.org/depts/los/convention_agreements/texts/unclos/part6.htm.]

Winkler, R. L., 1967: The quantification of judgment: Some methodological suggestions. *J. Amer. Stat. Assoc.,* **62,** 1105–1120.