To the Graduate Council:

I am submitting herewith a dissertation written by Joshua Adam Thomas entitled "Innovations in Dam Safety Risk Analysis: Exploring Motivational Bias and Opinion Pooling in Subjective Expert Risk Estimation." I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial Engineering.

Mingzhou Jin, Major Professor

We have read this dissertation
and recommend its acceptance:

James Ostrowski

Andrew J. Yu

Shuai Li

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

To the Graduate Council:

I am submitting herewith a dissertation written by Joshua Adam Thomas entitled "Innovations in Dam Safety Risk Analysis: Exploring Motivational Bias and Opinion Pooling in Subjective Expert Risk Estimation." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial Engineering.

Mingzhou Jin, Major Professor

We have read this dissertation
and recommend its acceptance:

James Ostrowski

Andrew J. Yu

Shuai Li

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Innovations in Dam Safety Risk Analysis: Exploring Motivational Bias and Opinion Pooling in Subjective Expert Risk Estimation

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Joshua Adam Thomas

December 2024

I dedicate this work to my children, James and Riley. It is my sincere hope that this is only the beginning of a legacy of curiosity and exploration.

Acknowledgements

I would first like to acknowledge my advisor, Dr. Mingzhou Jin for shepherding me through this process. Research is more than an extension of coursework. It is a new way to consider the world and he gave me the insight required to do just that. I would also like to thank the members of my committee, Dr. Shuai Li, Dr. Jim Ostrowski, and Dr. Andrew Yu, for their guidance and support.

I must also acknowledge Dr. Scott Turnbow and Patrick Kiser and the ratepayers of the Tennessee Valley Authority. Without their encouragement and support, this would have likely remained a goal beyond my reach. I also thank Jennifer Dodd, Chris Saucier, Jeff Munsey for their assistance and guidance.

Further, I also thank my wife, Ginny, and my children, James, and Riley, from whom I have sacrificed much of my time and surely all of their patience. Without them, I would be lost.

Finally, I would like to thank Jim Davis, the finest natural-born engineer I've ever known, for my earliest engineering education, Jeanette Davis for her insistence that I become a little bit better everyday, Karen Davis and Christel Tech for fighting for my early educational opportunities, Dr. Jeffrey Norden and Dr. Daniel Badoe of Tennessee Technological University for refusing to allow me to remove my hands from the plow, David Thomas for his constant encouragement through this process, and Kathy Busch for being an arrow that has always pointed fearlessly north.

Abstract

The topic of motivational bias in expert risk estimation for dams in the United States under the control of several federal agencies is addressed in the guidance documents of those agencies. However, the description of motivational bias in those guidance documents is limited to those situations wherein an expert risk estimator has a vested interest in the outcome of the risk assessment, such as some financial incentive to perform remedial work at the facility being assessed. This description is consistent with the literature upon which the guidance documents are based. Subsequent literature provides alternate descriptions of motivational effects that may also influence the behavior of the expert estimators. To determine if this is the case, a survey was developed to answer research questions related to this subsequent work and was distributed to dam safety industry professionals. This work provides evidence that these effects are likely influencing the dam safety estimation process and suggests training and facilitation techniques to reduce or eliminate their impact.

No prescriptive process for aggregation of expert risk estimates is provided in the federal dam safety guidance documents. This work presents commonly used aggregation techniques and alternate techniques as described in the academic literature. Decision makers working in the dam safety industry were consulted to better understand the current state of practice and the decision makers' desires with regards to expert weighting, information loss in aggregation, and other practical considerations. Additionally, a measure for determining how well each expert is represented in the aggregated probability function using Kullback-Leibler divergence

is presented along with an optimization method to adjust expert weights to improve representativeness in the aggregated distribution. Finally, a method for determining the relative importance of outlying estimates based on the conservativeness of individual expert estimates versus the general tendency of each expert is provided.

Table of Contents

List of Tables

List of Figures

Nomenclature

α	Parameter for Hölder pooling function
$\bar{b}$	Average of the best case estimates from all experts
$\bar{m}$	Average of the most likely estimates from all experts
$\bar{w}$	Average of the worst case estimates from all experts
c	Normalizing constant for the Hölder Pooling Function
D_{KL}	Kullback-Leibler divergence
$f(x)$	A piecewise function defining the triangular probability distribution
n	Number of experts
P	Individual expert probability distribution
$P(x)$	The pooled estimate of the probability of event x
$p_i(x)$	The opinion of expert i on the probability of event x
Q	Pooled probability distribution
w_i	Weights assigned to each expert
X	Set of possible events
Z	Normalizing constant

Chapter 1

Introduction

When faced with analyzing a complex system subject to high-consequence but low-frequency failures, as is the case with large dams, traditional standards-based approaches may be insufficient to provide an appropriately nuanced understanding of the initiating conditions, failure progressions, and, ultimately, the probability of failure of the system. The relatively rare and case-specific conditions that can lead to such a failure may be better quantified using a risk-informed approach that can leverage the expertise and experience of expert risk estimators to quantify the likelihood of a failure and the risk to the public in terms of both life safety and economic impact. This is the position of the United States federal government and the agencies that are charged with the regulation of federally owned dams and is the foundation of the current state of practice for dam safety risk assessment.

The risk informed decision making process (RIDM) for federally regulated dams has been developed cooperatively by several U.S. federal agencies and is described in both Chapters 2 and 3. At its center is the process of estimating the risk of a dam failure. This is achieved by developing a series of event trees, each of which contains an initiating event and a series of following events that culminate in the uncontrolled loss of the impounded reservoir or some other undesirable condition. Each event, represented as a node in the event tree, is assigned some likelihood distribution by a team of expert estimators. By means of a Monte Carlo simulation, the likelihood distributions are then used to calculate an annual probability of failure for the facility. It is this risk estimation process on which this work will focus.

Introducing the human element to what, in most other industries, would be a purely technical exercise introduces complexity and several new problems. In the case of a dam, or any other sufficiently complex system about which the technical team has high levels of uncertainty, this is unavoidable. Experts bring a variety of experience and knowledge, but this prior knowledge can introduce biases of all sorts. Bringing multiple experts together can create conflict that can exacerbate this bias. Even considering case studies for comparison of a facility to some other that has experienced similar conditions in the past can create an emotional response that can, in turn, be used as cognitive information further biasing opinions. While cognitive bias is covered extensively in the existing federal guidance documents, the subtler motivational effects stemming from conflict and emotional responses are not well-covered. The federal guidance documents do define motivational bias in limited terms, but that definition was based on very little prior academic work. The concept was defined largely by the work of Steven Vick in his 2002 book, Degrees of Belief. Vick held to a definition of motivational bias as "when a probability assessor has some stake, something to gain or lose, in the outcome of the assessment". (Vick, 2002) The Federal Energy Regulatory Commission, did not formally defined motivational bias and only spoke of reducing bias in general terms. (FERC, 2016) The U.S. Bureau of Reclamation and the U.S. Army Corps of Engineers mention motivational bias but defer to Vick's definition in their jointly prepared document "Best Practices in Dam and Levee Safety Risk Analysis". (USBR, 2019) Other work, not specific to dam safety, defines motivational bias in similar terms. (Frangopol and Tsompanakis, 2014; Otway and von Winterfeldt, 1992; Armor, 1999) Even work by experts writing for the dam safety industry failed to specifically define motivational bias and only spoke in general terms. (Bowles et al., 1998)

Upon reviewing the literature more broadly, it became evident that Vick's definition, while accurate, was limited and incomplete. This lead to exploring literature in a variety of domains to expand the definition substantially to include other motivational factors beyond the "vested interest" definition and even beyond the category of cognitive bias into that of affect influenced bias. Much of this literature was either contemporary or subsequent to Vick's book and it appeared that the basis of the federal guidelines should be updated to include this subsequent work. Validating that the expanded definitions of motivational bias were appropriate for dam safety risk assessments and that the effects of such bias were likely present and unaddressed in the process was the goal of Chapter 2 and should provide meaningful insight related to these issues to industry.

The problem of aggregating the opinions of multiple experts is also introduced when using RIDM. No longer can a decision be made purely on the basis of a standards-based calculation. A decision maker or their designee must select a method of both behavioral and mathematical aggregation to pool expert opinions. This is not at all a straightforward problem and is only made worse when the possibility of unequally weighting expert opinions is introduced. There is no single standard for opinion pooling in the federal guidance documents and the methods vary in practice. The work of Chapter 3 identified a significant gap with respect to opinion pooling as currently practiced in industry. This gap is not indicative of a lack of understanding of any particular pooling method, as those are quite well understood. It is, instead, indicative of a procedural gap and a lack of clarity, instruction, or prescription by the agencies responsible for regulation of federally operated dams on this topic. Specific software packages are mentioned in the guidance documents as well as required qualifications for software operators and other participants who will perform or influence the supporting calculations for the probability distributions, Monte Carlo simulations, etc., but there is no instruction for a specific method for the aggregation of subjective expert risk estimates. As this data forms much of the basis for the risk assessment output, it provides an opportunity for, at best, inconsistency of practice, and, at worst, error that negatively influences the output. After discussing the state of practice with industry experts (see Appendix B.1), it was learned that there may be great variability in pooling methods depending on the software package used and the preference of the practitioner.

One of the primary goals of Chapter 3 was to address this gap by investigating the opinion pooling methods known to be used in current practice along with other well understood methods, apply them to a set of sample data, compare the results of those methods, and comment on their behavior and differences. An improved understanding of these methods could inform and standardize future practice. The literature for arithmetic and geometric averaging (linear and logarithmic pooling, respectively) was quite well understood and seemed to be obvious first choices. After some investigation, it was found in the literature that the Hölder opinion pooling function provided a single equation that represented both the linear and logarithmic pooling functions as well as a range of methods in between by changing a single variable in the range between 0 and 1, so this approach was chosen.

A second gap discovered was a lack of guidance and consistency in the practice of weighting of expert opinions during the risk estimation process. Neither the desirability of, nor the process for, assigning different weights to some or all experts participating in the risk estimation process is covered in the federal guidelines. In conversations with industry professionals, it was relayed that unequal weighting is sometimes practiced, but that the weights tend to be subjectively assigned and this subjectivity may impact the dynamic of the expert team due to a perceived lack of "fairness" in the weights, with the presumption being that equal weighting provided equal impact in the output. While this may seem reasonable at first glance, it is not obvious that this is the case. It was suspected that each expert's input would likely not equally influence the output.

After investigating the literature, it was found that expert weighting in situations where there existed large sets of data and in which expert estimates could be calibrated against some known values was well understood and quite mature in practice. However, this is not the case for dam safety risk estimation. In this instance, the exact value of the quantities elicited are unknowable. There is a dearth of objective data because dam failures are so rare and the material conditions that exist at a dam, particularly one constructed of non-engineered materials such as clay, are very difficult to know with a high level of certainty. As such, the second goal of the paper was to establish some process by which the relative impact of each expert could be measured. Once measured, a method by which "fairness" could be optimized could then be developed. There existed in the literature some work related to the relationship of individual expert estimates to the pool of all expert estimates by using the Kullback-Leibler divergence (KLD). This KLD approach seemed reasonable and was further investigated. The measurement of the KLD is straightforward and well understood, but the application of some method to maximize "fairness" could not be found in the literature. Some recent work by Abbas (2009) was the nearest to addressing this particular problem, but was more focused on optimally reducing the sum of the KLDs of the input distributions. A similar approach was taken here but the highest individual expert KLD was minimized. This served the purpose of reducing the KLD of the highest outlier and reduced the range of KLDs, yielding a more "fair" pooled function. The justification for this approach was that weighting the experts based on representativeness in the pooled function would serve to appropriately weight the expert opinions without causing undue negative impact on the team dynamic. This approach assumes an axiomatic desire by the decision maker to optimize for representativeness. This may not always be the case in practice, but exploration of this concept and presentation of this method to industry provides an option in this direction should decision makers choose to explore its implementation.

Finally, at the urging of industry professionals, a method was developed to provide some comparison of experts, both overall and at the node level, to better understand the relative importance of a dissenting node estimate. It is reasonable to suspect that if, after several rounds of behavioral aggregation, potentially followed by weight optimization, a single expert's estimate is still quite different to the pooled opinion, this dissenting opinion may contain very important information that should be considered more carefully by the decision maker. However, current practice has no method for comparing experts in this way. It is not clear if a decision maker could discern the importance of a dissenting estimate without understanding how that expert typically estimates compared to the rest of the group, and whether the outlying estimate aligns with or contradicts that expert's usual pattern. This method can provide decision makers with context that could allow them to shift resources to additional field investigations, further objective analyses, etc. to validate or repudiate a dissenting outlier estimate. This additional context could also provide improved justification in the narrative report to warrant action or inaction in response to the impact of the dissenting estimate on the final output. In short, there is a desire to understand when a dissenting opinion is the result of the natural tendency of some expert to provide a higher or lower likelihood estimate and when it is indicative of a passionately held belief that should be respected and explored.

Chapter 2

Motivational Effects in Expert Subjective Dam Safety Risk Estimation

2.1 Introduction

Following the failure of Teton Dam in Idaho in 1976, the United States government took a more active role in regulating the process by which dam safety was considered and monitored by federal agencies. In 1977, President Carter issued a memorandum to all federal agencies involved with the "planning, construction, operation, and ultimate disposal" of dams. The memorandum required, among other initiatives, that the agencies evaluate the degree to which a "probabilistic or risk-based analysis" could be incorporated into all aspects of dam siting, design, construction, and operation. To that end, a number of federal agencies began developing the framework for the current Risk-Informed Decision Making (RIDM hereafter) process as required by the U.S. Federal Emergency Management Agency. (FEMA, 2004)

This approach stood in stark contrast to the traditional standards-based approach (SBA hereafter), which assessed a dam's safety by comparing its design and condition, to well established rules and engineering guidelines. While the SBA was, and is still, a critical component of proper dam safety management, dams are notoriously difficult structures to design and evaluate. This is due to the often unknowable qualities inherent in the structure's materials, the frequency and magnitude of the loading to which they are subject, and the underlying geological conditions in their vicinity. Items such as human performance in operations and the reliability of mechanical components critical to proper operation in extreme events are also not reliably quantifiable. (FERC, 2016) There are also the issues of the constantly changing state of practice, evolving understandings of flood and seismic loadings not considered in the original design, and the inability to observe dam performance in potential extreme events. (Bowles et al., 1997) Reevaluation of an existing dam to a changing standard is potentially complex, expensive, possibly unnecessary, and, as geotechnical investigation methods can be quite invasive, not without risk of its own.

The risk-informed approach required by the government sought to employ methods to recognize that risk of undesirable performance existed, quantify that risk in some meaningful way, and compare it to some new standard by which acceptability of failure could be benchmarked. (USACE, 2014) The approach required by the RIDM framework was not entirely new. Engineers in the U.S., the United Kingdom, the Netherlands, and Hong Kong had been using the principles of risk-based decision making for societal and personal safety issues since the 1960s. (FERC, 2016) However, the agencies were faced with developing a method that could be scaled and readily deployed across an array of structures to reasonably access their condition, the likelihood that a structure would suffer some unfavorable event, and the likely consequences of that event. Bayesian decision analysis using subjective probabilities had been used in other industries since the 1950s. (Vick, 2002) This provided a method by which the uncertainty of dam safety parameters could be quantified. However, unlike other industries where this method can provide an expected value that can be compared to other measurable outcomes over time, the low incidence rate of dam failures and the potential for the extreme consequences that may accompany a failure dictate that the field of dam safety cannot similarly benefit from a large history of past failures or expected results based on precisely defined parameters. (Bowles et al., 1998) Thus, a high level of uncertainty is present in any dam safety risk analysis. This uncertainty is accepted as unavoidable. Each federal agency has adopted some variation of a common risk management framework, within which exist processes to account for uncertainty. (FEMA, 2015)

A typical risk management framework is arranged as a cyclical process as shown in Figure 2.1 wherein a dam is regularly reevaluated. A list of potential failure modes is compiled during a potential failure mode analysis (PFMA hereafter), an estimation of the dam's risk is performed, an evaluation of the acceptability of that risk level is performed, action is taken to reduce or monitor the risk level, then the process repeats. The risk estimation portion of the process attempts to quantify the likelihood of difficult-to-predict events that would lead to a failure. Because of the uncertainty surrounding many of the events considered in the risk estimation process, it often requires extensive use of subjective probability estimates to quantify the likelihood of failure.

2.2 Risk Estimation

Federal agencies have modeled their RIDM programs around the FERC Risk Informed Decision Making Guidelines. Risk estimation is performed by means of a formal process often called either a Quantitative Risk Assessment (QRA hereafter) or a Semi-Quantitative Risk Assessment (SQRA hereafter) depending on the desired complexity of the analysis. During these assessments, each failure mode identified in the PFMA process is analyzed in turn. The risk estimation team, lead by a facilitator, identifies all required consecutive events in order for a failure to occur. These events are then used to populate an event tree with each event represented by a node in the tree. Probabilities or probability distributions are then assigned to each node, either through assignment of values by means of objective engineering analysis or by

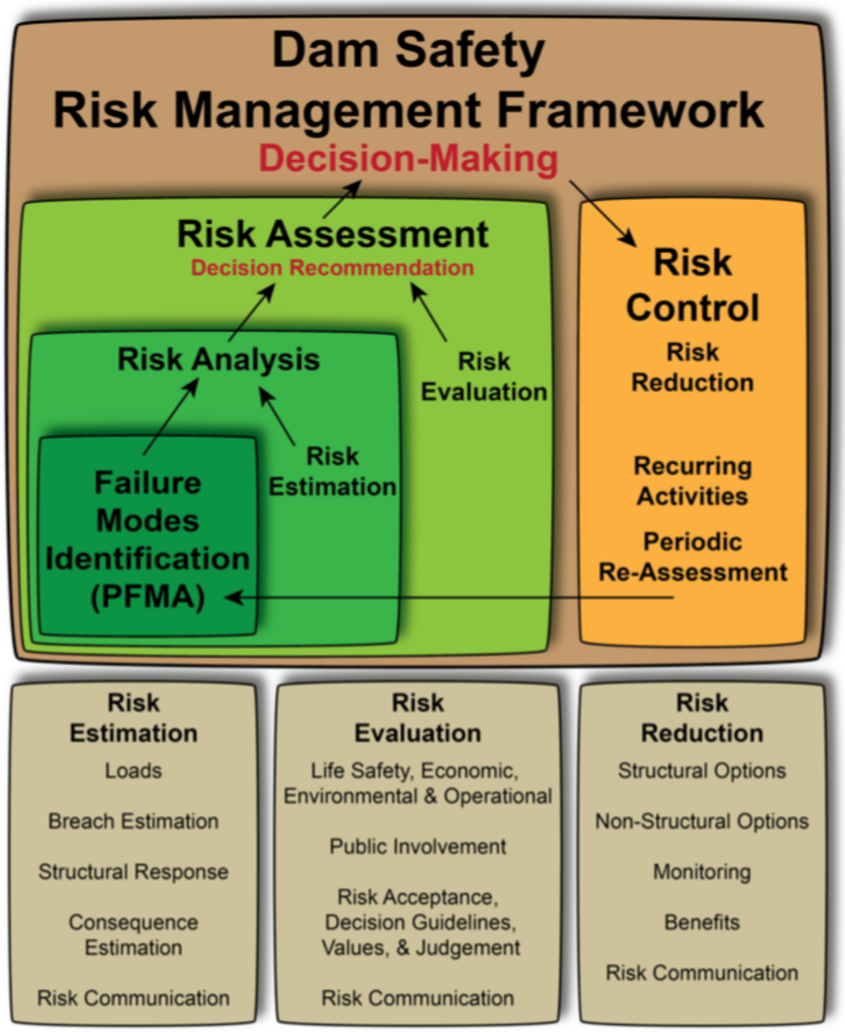

Figure 2.1: Typical Risk Management Framework From Federal Guidelines for Dam Safety Risk Management by FEMA in the public domain.

subjective expert estimation. Once all nodes have been considered, a Monte Carlo or similar analysis is performed to determine the annual likelihood of failure for each failure mode. These likelihoods are aggregated and an annual likelihood of failure for the dam is established. A similar process expands upon the potential failures to assess the consequences of failure. (USBR, 2019)

While the probability of an event at some event tree nodes can be reasonably supported by traditional engineering analysis, it is often the case that the likelihood that an event will occur is not reasonably predictable. This uncertainty can be broadly divided as aleatory or epistemic. (Frangopol and Tsompanakis, 2014) The former describes the unpredictability of the natural world, such as the frequency of seismic or extreme rain events. The latter deals with a lack of knowledge on the part of the assessors. Geologic conditions, soil properties in areas that cannot be reasonably sampled, groundwater flow regimes, and other physical characteristics of the dam that cannot be readily quantified fall into this category. (USACE, 2014)

2.3 Motivational Effects

2.3.1 Motivational Bias as Described in Current Federal Guidelines

Broadly defined, cognitive bias is considered to be any number of common faulty cognitive processes of well-meaning assessors. (Tversky and Kahneman, 1975) The potential for experts in the risk estimation process to make poor judgements due to cognitive bias is an obvious concern and is addressed in the guidelines. These cognitive effects are well understood and the guidelines consider them to be appropriately mitigated through a combination of estimator training and facilitator-led discussions. However, the subset of motivational bias has historically been less understood. Motivational bias in the dam safety risk management process is discussed only in general terms in the guidelines and policies of the implementing federal agencies.

In the USACE/USBR jointly-prepared ”Best Practices in Dam and Levee Safety Risk Analysis” (USBR, 2019), motivational bias is acknowledged and described as an estimator having a vested interest in the outcome of the risk estimation. The proposed resolution for this is for a trained facilitator to identify biased views and attempt to encourage opposing views in an effort to eliminate the effect of the bias. FERC (2016) lists a number of potential types of bias to avoid during expert elicitation, several of which describe behaviors related to motivational bias. However, the descriptions are brief with no specific mention of estimator motivation as a cause of error. Similar to the USBR recommendation, the facilitator is expected to recognize and minimize the negative effects.

Both USBR (2019) and FERC (2016) reference the seminal book, ”Degrees of Belief” by Vick (2002). Vick describes motivational bias in a bit more depth but acknowledges that cognitive research has little to offer on the subject. The work referenced by Vick similarly described motivational bias as the bias that occurs due to the assessor having some take in the outcome. (USBR, 2019; Vick, 2002; Otway and von Winterfeldt, 1992) While further discussion of the mechanism of motivational bias is limited, Vick does offer several stereotypical archetypes of expert estimators who may be compelled to influence risk estimation to serve their own self-interest. These archetypes are organized by their role in the project: designer, builder, investigator, etc. While all these archetypes, under more objective circumstances, would presumably fiercely defend their ethical judgement as professionals, Vick notes that the lack of an objective standard in this setting is an invitation for manipulation of the results to suit one's own needs. This paired with what (Layton Jr, 1986) described as the ”corporatization of engineering” may serve to decouple a professional's ethical obligation. Estimators may cease to view themselves as practitioners of the profession bound to ethical conservatism and take on the role of an organizational employee prone to providing a response driven by organizational loyalty or material self-interest. Likewise, the project sponsor can introduce error by way of their own motivational bias. As the project sponsor, one may be inclined to leverage the results of a risk estimation to advocate for some preferred outcome to the public, a regulator, or some other interested party. Vick warns of the potential for the failure of the RIDM process to protect the public interest due to the influence of unchecked motivational bias. A failure of this kind, he argues, would be the to the engineering profession what the Enron scandal was for the accounting profession.

While motivational bias as defined in the federal guidelines and Vick (2002) describe the phenomenon, they have been shown to be incomplete descriptions of all motivational effects in subjective assessments. Vick's work, and by extension the federal guidelines, frame motivational bias as an intentional behavior by an estimator to suppress information that could be damaging to the likelihood that their preferred outcome would be achieved. While such intentionality may exist, it is possible for motivation to play an outsized role in decision making without a willful choice to manipulate the process. A summary of some relevant definitions of motivational bias can be found in Table 2.1.

Kunda's foundational work on, what she called "motivated reasoning" provided an explanation by which "motivation could affect reasoning through reliance on a biased set of cognitive processes". (Kunda, 1990) This provides a middle ground between outright subversion of the risk estimation process by unethical actors and the purely cognitive biases for which the federal guidelines provide ample discussion and remedial strategies. This third path allows that it is possible that a well-meaning estimator could harbor unrecognized motivation to come to conclusions that are completely defensible in their own minds but that are biased heavily toward their favored outcome. This possibility is not directly addressed in the federal guidelines.

Table 2.1: Definitions of Motivational Bias in Current Regulations and Supporting Work

Source	Term Used	Definition	Proposed Remediation
Frangopol and Tsompanakis (2014)	Motivational Bias	"...when one or more of the members making estimates have a vested interest in the outcome"	"A facilitator can attempt to bring out opposing views when an opinion appears to be expressed from a motivational bias."
Vick (2002)	Motivational Bias	"when a probability assessor has some stake, something to gain or lose, in the outcome of the assessment"	"The assessor must act as a disinterested party to the outcome in providing probability judgements as fairly and honestly as possible."
Otway and von Winterfeldt (1992)	Motivational Bias	"the experts have a stake in the outcome of the analysis or the decisions that result from it"	"define clearly the nature of the judgments, to make explicit the assumptions and reasoning behind them, to expose their inherent uncertainties, and to counter possible biases"
USACE (2014)	N/A	Not defined.	States that the Dam Safety Program Management Tools (DSPMT) is to "provide unbiased data"
FERC (2016)	N/A	Not defined.	States that subjective probability should be "obtained by considering all available information honestly, fairly, and with a minimum of bias."
Armor (1999)	N/A	References (FERC, 2016)	See (FERC, 2016)
Bowles et al. (1998)	N/A	Not defined.	States that "improved procedures for eliciting professional judgments and minimizing biases which might exist in these judgments should be developed"

2.3.2 Developments in Motivational Bias Understanding

As research has progressed in this area since (Kunda, 1990), it has become evident that motivational bias and motivated reasoning are much more complex topics than the federal guidance documents might suggest. While motivational biases can manifest as deliberate efforts to further a personal agenda, it is also possible that people may be influenced unconsciously by external forces or by their own unrecognized desires. (Montibeller and von Winterfeldt, 2015) A number of salient motivational effects presented in recent literature are presented herein.

Affect Influenced Bias

The influence of affect on decision making can lead to unrecognized motivational bias. Affect is described by (Slovic et al., 2004) as a "faint whisper of emotion". That is to say, it is an automatic emotional response to a stimulus. People unwittingly use the feelings derived from this emotional response as additional information in the decision making process without regard to the rationality of such behavior. (Albarracín and Tarcan Kumkale, 2003) In fact, the impact of affect on decision making has been shown to provide results that are counter to the assumed behavior of rational individuals. Denes-Raj and Epstein (1994) Baybutt (2018) acknowledged an emotional component in their discussion of biases but provided no proposed remediation.

With regards to risk, it is reasonable to assume that risk and benefit of activities are independent of one another and exist across a range of values in both dimensions. Further assuming that activities with a high risk and low benefit would be eschewed or prohibited outright by society, it would follow that risk and benefit would correlate positively. However, risk and benefit have been shown to correlate negatively in the mind of research respondents. (Finucane et al., 2000; Fischhoff et al., 1978; Efendic et al., 2021) This suggests that the feeling of the respondents colors their feelings about an activity based on their opinion of its "goodness" or "badness" with a positive opinion of an activity resulting in a low risk-high reward ranking and a negative opinion resulting in a high risk-low reward ranking. This may be evidence of the "halo effect", which holds that people frequently judge all factors of a person, thing, or activity consistently with their overall perception of it. (Alhakami and Slovic, 1994) Furthermore, it has been shown that this relationship can be manipulated by providing respondents with information about an activity to be judged prior to their risk-benefit ranking. (Finucane et al., 2000) This may be particularly important for dam safety risk estimation due to the required discussion of potential failure modes, failure consequences, and case studies of other facilities. If those case studies described outcomes that were overwhelmingly negative they could shift the estimators' positions in an excessively conservative direction. Conversely, if case studies were presented solely for similar facilities that experienced failure mode initiating events but performed well or benefited from successful interventions, the balance could shift in an overly non-conservative direction.

In some cases, where objective data is readily available, experts are less prone to the influence of affect as a layperson might be. (Siegrist et al., 2007) While all of Vick's archetypes would, presumably, be knowledgeable and experienced enough to be considered an expert as required by the governing federal guidelines, they may not be fully immune to the influence of affect. In situations where the science is not fully understood, as would likely be the case when subjective expert estimation would be the only remaining option for risk quantification, the use of affect as information in lieu of reliable data is a possibility. (Sokolowska and Sleboda, 2015)

Myside Bias

Expert dam safety risk estimators may also be influenced by the entity that retained their services. While estimators are engaged by the federal agencies that own the dams (USBR, 2019), individual risk estimators may be engaged by functional subentities within those agencies or by other stakeholders. Those subentities and stakeholders are corollaries to Vick's archetypes and maintain the potential for motivational bias.

It is not unreasonable to suspect that estimators would be influenced by the desires of their functional employers and may be placed at odds with their peers, who may feel a conflicting allegiance. Furthermore, as the group formation process develops among the estimators, intragroup conflict will arise causing divisions among the estimators. (Tuckman, 1965) Once this conflict manifests itself in groups, people tend to tailor their evaluation of evidence to fit the position of the side with which they are aligned. (Stanovich et al., 2013) This cognitive failure is known as "myside bias".

It may be optimistically hoped that expert estimators, chosen for their intelligence, would not be as susceptible to myside bias. However, when tasked with developing counterarguments to challenge a prior belief, cognitive ability does not appear to be related to ability. (Perkins, 1985; Klaczynski and Gordon, 1996; Klaczynski, 1997; Klaczynski and Robinson, 2000) When this line of research was expanded to assess the quality of counterarguments rather than generation of such counterarguments, the results are similarly dissociated from intelligence. (Stanovich et al., 2013) There is, however, some evidence that resistance to myside bias may be a trainable skill. The quantity of one's post-secondary education, and the opportunities such an education provides to build experience that "brackets" one's beliefs, seem to reduce the effects of myside bias. (Toplak and Stanovich, 2003)

While years of post-secondary education may reduce the effect, highly trained forensic scientists have been shown to still be susceptible to myside bias. (Dror, 2020) Forensic scientists provide a suitable stand-in for dam safety expert risk estimators in that they have typically received years of specialized training and education and they perform highly technical but often quite subjective work on important cases that can have dramatic impacts on the lives of the public. (Pronin et al., 2002) Moreover, the stakeholders surrounding the forensic scientist are not unlike Vick's archetypes with social and professional pressures that are, presumably, quite similar. (Dror, 2017) Even forensic experts educated at the Ph.D., M.D., and Psy.D. level with significant experience in the field have been shown to highly favor the side of their presumed employers in an experimental setting when presented with identical data. (Murrie et al., 2013)

Adversarial Effects

The existence of myside bias within the estimator group provides the opportunity for compounding effects during estimate aggregation. As such, the preference of the federal agencies to encourage an iterative probability estimation process may pose more complex problems than are acknowledged in USBR (2019). The iterative process involves each estimator providing subjective estimates, sharing those estimates with the larger group, and the facilitator encouraging conversation or experts to justify extreme estimates. It is possible that this discussion may serve to create adversarial relationships, particularly if the divergence in estimate values falls along organizational lines. (Dror, 2020) While the current federal guidelines warn of the potential for adversarial effects in the form of bullying, this discussion is limited to the case where an estimator with a strong personality might sway the remaining estimators' opinions. This is certainly a possibility, but recent work suggests that once a group member senses competitive behavior from a counterpart, they often take an adversarial position that increasingly diverges from "conciliatory and cooperative courses of action". (Simon et al., 2020)

This response has been demonstrated to stem from the assumption of bias on the part of the outgroup member. People tend to recognize cognitive and motivational biases more readily in others than in themselves. (Pronin et al., 2002; Armor, 1999) This, in turn, results in an exaggerated adversarial response from that person beginning a cycle of escalation of conflict. (Kennedy and Pronin, 2008) This could explain why group interaction during consensus building tends to result in more extreme probability estimates. (Ouchi, 2004) This behavior is not as bounded as one might hope when in terms of rational response. For instance, parties in negotiations, even experts participating in high-stakes political negotiations, have been shown to act counter to their own self-interest once engaged in the cycle of conflict escalation. (Kelman, 2008)

The adversarial effects increasing group polarization could lead to more data gathering, as described in the federal guidelines. Such an effort may be undertaken to better quantify the event likelihood and reduce the large range of expert uncertainty. However, these efforts can be costly in terms of time and money and it is not always possible to more objectively determine the likelihood of an event. In such cases, it is unclear what solution may be available to the RIDM team to reconcile the broad spread in belief.

2.4 Aims of the Present Research

While the U.S. federal regulations and the research upon which they are based provides some limited discussion of motivational bias, research in the area has progressed substantially since their implementation. While individual practitioners facilitating risk estimation sessions may have incorporated a more sophisticated understanding of motivational bias in their instructions, training, and interventions into group discussions or conflicts, it is unclear if the industry has incorporated a more thorough understanding of the body of knowledge into regular practice.

Firstly, this research aims to determine if the "vested interest" definition of motivational bias is typically the sole definition provided to risk estimators during elicitations. An accurate description of the term, which includes the subtle emotional responses of the estimators and their effects, may serve to improve the quality of the subjective probabilities. As such, it is desirable to understand if this is commonly provided.

The second aim of this research is to understand the relative frequency of case studies that describe negative outcomes, such as loss of the water retaining barrier, loss of life, substantial property loss, etc., versus case studies that describe satisfactory performance of the dam, successful interventions, or recent best practices that may improve the confidence of the estimators. It is possible that facilitators may be unwittingly biasing estimation teams towards more conservative scoring by focusing on rare negative events.

Finally, this work aims to understand how often conflict or disagreement exists in the consensus building or aggregation stage of the estimation process and to determine if conflict escalation is self-recognized and if conflict escalation is suspected by estimators of their peers.

2.5 Research Methodology

In order to assess the current state of practice with regards to the research aims, a series of questions was developed to gauge the experience of working professionals who participate in the risk estimation process as expert estimators or facilitators. (see Appendix A.1) These questions took the form of a statement prompt paired with response choices on a five point Likert scale to measure the perceived frequency of certain events ranging from "Never" to "Always" with an additional selection of "Don't Know / N/A". A number of demographic questions were also included in the survey to determine if education level, years of experience in the field, or other factors affected the respondents perceptions of the state of practice. The demographic measures for age, years of experience in the field, and number of expert estimations were gathered as categorical responses. The questions were combined into an online survey and sent to industry professionals at several federal agencies responsible to the RIDM process at federally regulated dams and levees and practicing engineers and consultants who regularly take part in risk estimations. The survey recipients were also asked to distribute the survey within their organization to others who also participate in the RIDM process.

2.6 Results and Discussion

2.6.1 Participation and Validity of Responses

51 responses were provided to the online survey. Of those, four were incomplete and were excluded from the results. Three of the respondents stated that they had not participated in the risk estimation process as an expert risk estimator. While it is possible that they may have participated in the process in other capacities such as technical advisor, oversight, or in some other role, further information to verify was unavailable, so these responses were excluded as well. This left 42 valid complete survey responses for analysis.

Of the valid responses, 21 were from non-federal agency employees. Those respondents work as independent consultants or for engineering companies that provide technical services for dam safety projects. The remaining 21 responses were from employees of various federal agencies with responsibility for federally regulated dams and levees. Federal employee respondents included 12 employees from the USBR, five from the Tennessee Valley Authority [TVA], two from the USACE, one from FERC, and one with experience with both USACE and FERC.

Overall, the respondents were highly educated and highly experience in the field. All were college graduates with at least a Bachelors degree. 35 of the 42 held an advanced degree with ten holding a doctorate. All respondents had at least 6-10 years experience in the field while seven had 10-15 years of experience, five had 15-20 years, with the remaining 25 having 20 years or more. The level of experience of respondents in QRAs or SQRAs for risk estimation varied from only 1-5 assessments to 20+, however the non-federal employees tended to have much less experience with only 33% having been an expert estimator in 6 or more risk estimations while that minimum experience level was reported by more than 90% of federal employees. Detailed demographic data for respondents is presented in Tables 2.2 - 2.6

Table 2.2: Years of Experience

		How many years of experience do you have in the dam safety industry				Total
		6 – 10	10 – 15	15 – 20	20+	
In approximately how many quantitative or semi-quantitative risk assessments for have you participated as an expert estimator?	1 – 5	3	2	1	10	16
	6 – 10	1	1	0	3	5
	10 – 15	0	0	2	0	2
	15 – 20	1	2	1	4	8
	20+	0	2	1	8	11
Total		5	7	5	25	42

Table 2.3: Highest Level of Education

		What is the highest degree or level level of education you have completed in a dam safety related field?			Total
		Bachelor's degree (e.g. BA, BS)	Master's degree (e.g. MA, MS, MEd)	Doctorate or professional degree (e.g. MD, DDS, PhD)	
In approximately how many quantitative or semi-quantitative risk assessments for have you participated as an expert estimator?	1 – 5	1	10	5	16
	6 – 10	2	3	0	5
	10 – 15	0	2	0	2
	15 – 20	1	7	0	8
	20+	3	3	5	11
Total		7	25	10	42

Table 2.4: Respondent Age

		What is your age?				Total
		35-44	45-54	55-64	65-74	
In approximately how many quantitative or semi-quantitative risk assessments have you participated as an expert estimator?	$1 - 5$	1	5	5	5	16
	$6 - 10$	2	1	2	0	5
	$10 - 15$	2	0	0	0	2
	$15 - 20$	4	1	2	1	8
	20+	1	3	4	3	11
Total		10	10	13	9	42

Table 2.5: Respondent Gender

		What is your gender?			Total
		Male	Female	Other	
In approximately how many quantitative or semi-quantitative risk assessments have you participated as an expert estimator?	$1 - 5$	12	3	1	16
	$6 - 10$	5	0	0	5
	$10 - 15$	1	1	0	2
	$15 - 20$	8	0	0	8
	20+	9	2	0	11
Total		35	6	1	42

Table 2.6: Respondent Employer

		Employer		Total
		Federal	Non-Federal	
In approximately how many quantitative or semi-quantitative risk assessments have you participated as an expert estimator?	$1 - 5$	2	14	16
	$6 - 10$	3	2	5
	$10 - 15$	2	0	2
	$15 - 20$	6	2	8
	20+	8	3	11
Total		21	21	42

2.6.2 "Vested Interest" Definition of Motivational Bias

When asked how frequently the topic of motivational bias is formally discussed either as part of the risk estimation process or in associated bias training, 83% of respondents reported that at the topic was addressed at least sometimes. (See Figure 2.2) The number of risk estimates in which a respondent has participated and their response to the relative frequency of formal discussions of motivational bias were compared using Fisher's exact test and were found to be significantly related (p=0.002). This suggests that, while motivational bias is a common topic during risk estimations or associated bias training, its discussion is either not ubiquitous or is perhaps not sufficiently intensive for less-experienced estimators to recall its presentation. In these cases, estimators may be more susceptible to motivational bias since they may not have received a recent reminder of its impact on decision making or of potential debiasing techniques. This condition may lead to a lack of focus on the possibility of motivational bias and allow estimators to fall victim to the bias without a recent reminder to be thoughtful and intentional in implementing debiasing strategies in their own thinking as they formulate their subjective estimates.

More complex definitions of motivational bias beyond the common "vested interest" definition as described in the regulations and in Vick do not appear to be broadly well understood. 47% of respondents stated that either they did not know if more complex definitions had been provided or that they had never been provided. Of those who stated they had been presented with a more complex definition, 11 stated that they could not recall the specifics. Of the remaining 11, only four provided specifics that were not either largely similar to the "vested interest" description or were not descriptive of other cognitive biases covered elsewhere by the regulations. However, these four responses included discussion of unconscious cognitive bias, relationship to the project owner, a conscious desire to disrupt the process, and the risk of taking a side with a portion of the estimation team. These are all topics found in the more recent literature. These concepts have found their way into practice, but

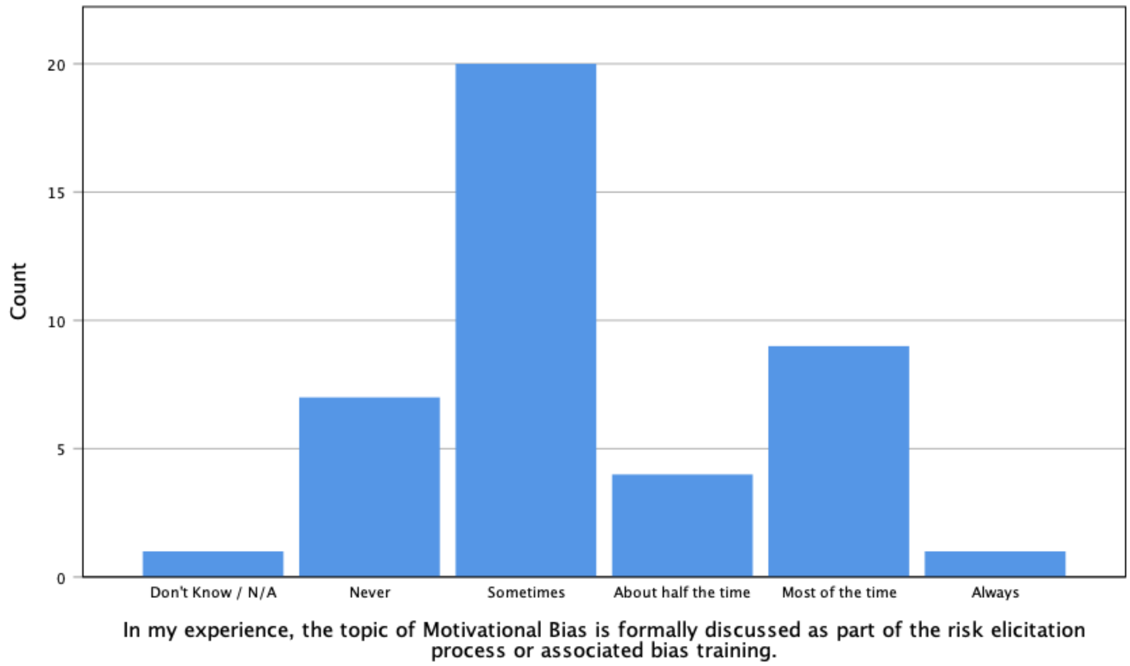

Figure 2.2: Motivational Bias Discussion Frequency

only in a limited way, and it is not clear that they have been discussed intentionally as part of the estimation process. It is also unclear if this exhibits a formal understanding of the topics or an intuition based on personal experience or knowledge from other domains. While the understanding of motivational bias as a subtler and more complex issue may be slowly finding its way into practice, it seems to have done so in a very limited manner, and without formal training to extend the reach of the definition, it is unclear that significant improvements in understanding can be assured in the short term and it is unlikely that a universal definition will gain consensus.

2.6.3 The Role of Affect

When asked about the case studies of similar facilities presented during risk estimations, 54% of respondents stated that the presented case outcomes following failure mode initiation were mostly or always negative. 38% stated that the case studies presented were evenly divided between positive and negative outcomes with less than 5% of respondents claiming mostly positive outcomes. (See Figure 2.3)

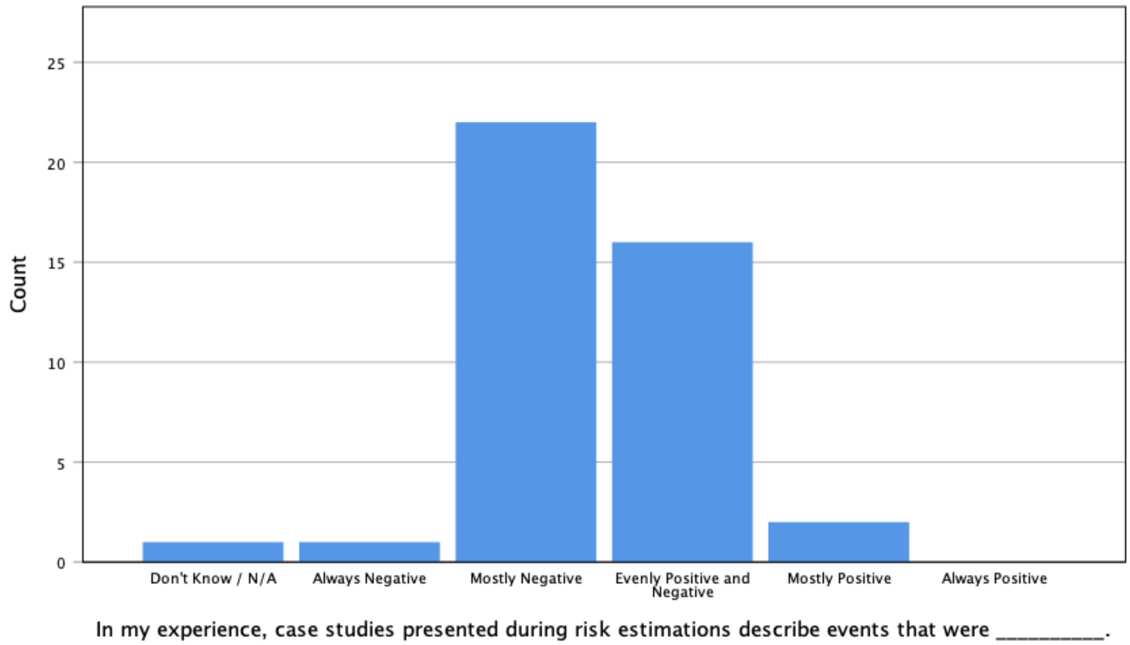

Figure 2.3: Case Study Types

Statements regarding relative frequency of case study types did not relate to years of experience, number of risk estimations, public vs. private sector employment, or specific agencies in a statistically significant way.

This largely negative presentation of similar cases could serve to impact the response of estimators. The case studies are commonly presented with video or still images of failed facilities and stark descriptions of the consequences of failure in human, financial, and reputational terms. These negative images and narratives may serve to induce an exaggerated risk perception in the minds of estimators and lead them toward unrealistic estimations. This is of particular concern in instances where the risk estimation process is used to compare two competing designs. A particularly negative related case study could not only increase the risk of an option in the minds of the estimators but reduce its perceived benefit as well.

Half of the respondents stated that there was never formal discussion, either in risk estimations or in related bias training, regarding the impact of emotion in decision making while 30% stated that it was formally discussed sometimes and just under

17% responding that is was discussed about half or most of the time. The responses to this question were related significantly to the number of risk estimations in which the respondent had participated (p=0.034) with more experienced estimators claiming a higher likelihood of a discussion of emotional effects. Additionally, 12 of the 21 respondents who replied "Never" had only participated in 1-5 risk estimates as expert estimators.

This, as with discussion of motivational bias as a whole, is likely related to inconsistent training or discussion by facilitators. Although the study of affective rationality is still relatively new and little is known with regards to minimizing or limiting affect influence in a risk estimation, it would likely be beneficial to facilitators and estimators to be aware of the possibility of affect influencing what is typically considered to be a rational cognitive process.

2.6.4 Myside Bias

Respondents were asked a series of questions regarding conflict within expert estimating teams. For the sake of consistency among the respondents, conflict was defined in the survey as interpersonal conflict among participants during the risk elicitation process. Escalation was defined as the hardening of one's own position when faced with interpersonal conflict during the consensus-building or aggregation phase of the risk estimation process.

The data suggests that conflict is common in risk estimation processes, with 88% of respondents reporting witnessing conflict at least sometimes, and nearly 24% experiencing it at least half the time. (See Figure 2.4) Furthermore, 59% of respondents stated that they never escalated conflict, while the remainder reported that they escalated it sometimes. In contrast, only 19% of respondents believed that their peers never escalated conflict, with the rest reporting they had witnessed their peers escalate at least sometimes. By this process, it was determined that nearly 43% of respondents claimed that the incidence of their peers escalating conflict was more

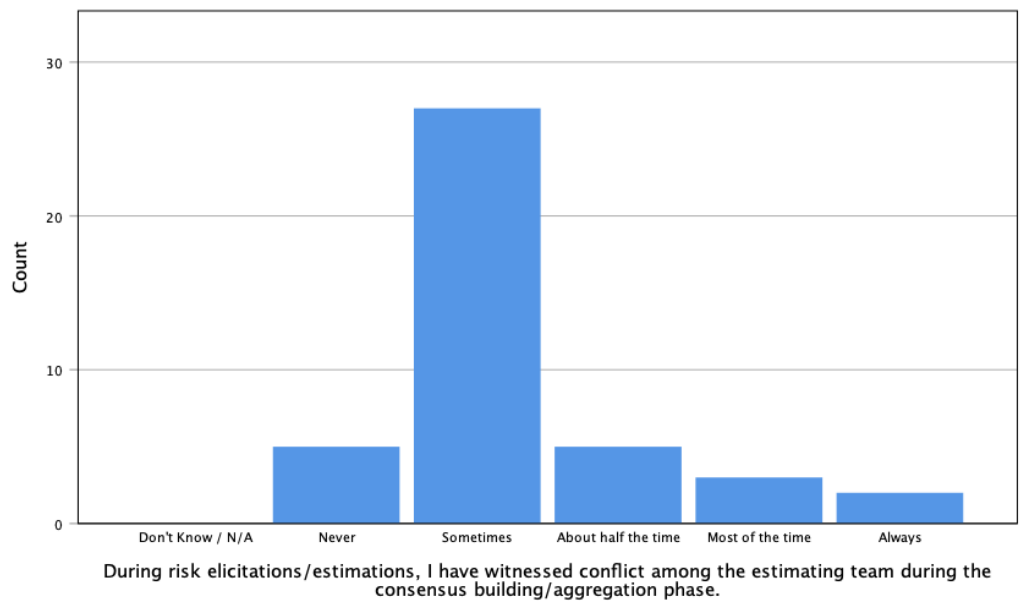

Figure 2.4: Witnessed Conflict Frequency

common than their own with all but one of the remaining respondents stating they believed their peers to be equally as likely as them to escalate conflict. (See Figure 2.5) The likelihood that a respondent felt their peers would be more likely, less likely, or the same as them with regards to conflict escalation did not correlate with age, years of experience, education level, employer, or number or risk estimations.

2.6.5 Adversarial Effects

Respondents were asked about both their own tendency and that of their peers to provide overly conservative and unconservative risk estimates to counter the positions of another estimator with an opposing position. Respondents stated that they never provided overly conservative or unconservative estimates at rates of 73% and 90% respectively with the remainder stating that they sometimes provide such estimates, save one who responded that they did not know. With regards to their peers, respondents were somewhat less willing to speculate, providing five responses of "I don't know". For those who provided responses for themselves and their peers, the same process was followed as was used for escalatory behavior and the relative

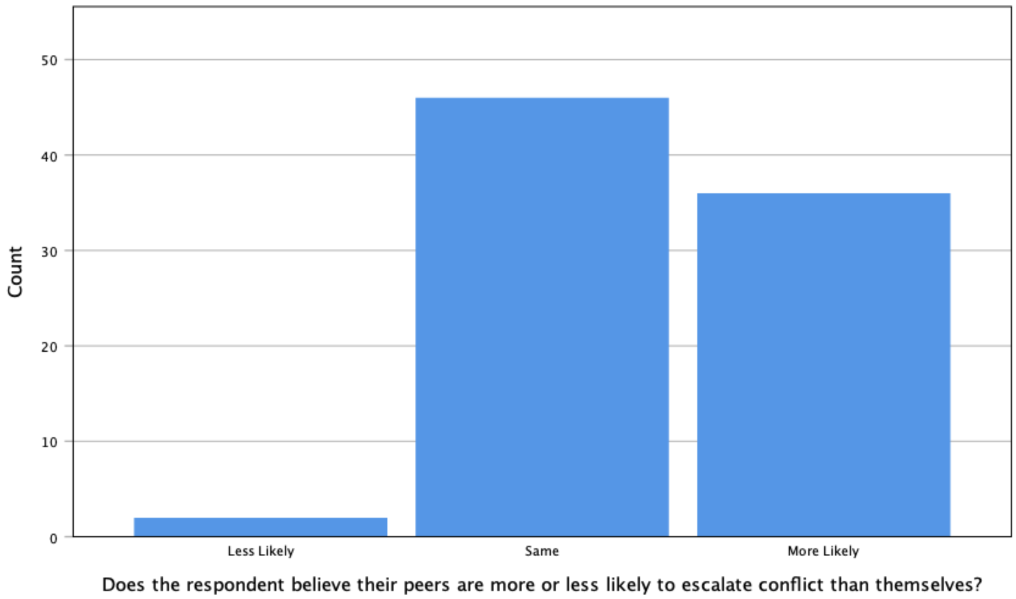

Figure 2.5: Perceived Willingness of Peers to Escalate Conflict vs. Self

likelihood of providing extreme estimates in response to a counter position was determined. From this data, 50% believed that their peers were more likely than themselves to provide overly conservative estimates to counter the position of another estimator. (See Figure 2.6) Similarly, nearly 53% believed that their peers would provide overly unconservative estimates. (See Figure 2.7) Years of experience in the dam safety industry is significantly related to the respondents' estimation of the relative likelihood that their peers will provide overly unconservative estimates than themselves ($p=0.047$). The same was not true for the overly conservative case.

2.6.6 Perceptions of Oneself and Others

The data suggests both that conflict is common in risk estimates and that there is either a lack of self-awareness among estimators, a distrust of their peers to behave rationally, or perhaps both. The effect of myside bias is likely influential in this behavior as it predicts both highly favorable opinions of one's own objectivity and suspicion of the objectivity of outgroup members. A limitation of this study is that the measure of unrealistic risk estimates was considered in terms of the members' risk

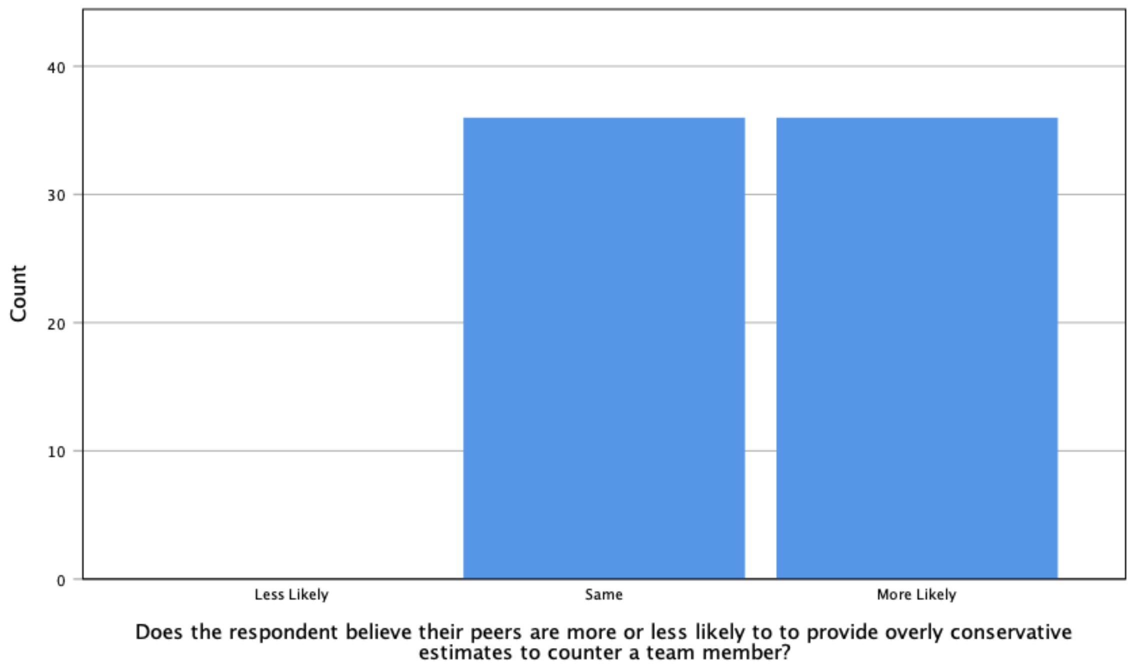

Figure 2.6: Perceived Willingness of Peers to Provide Overly Conservative Estimates vs. Self

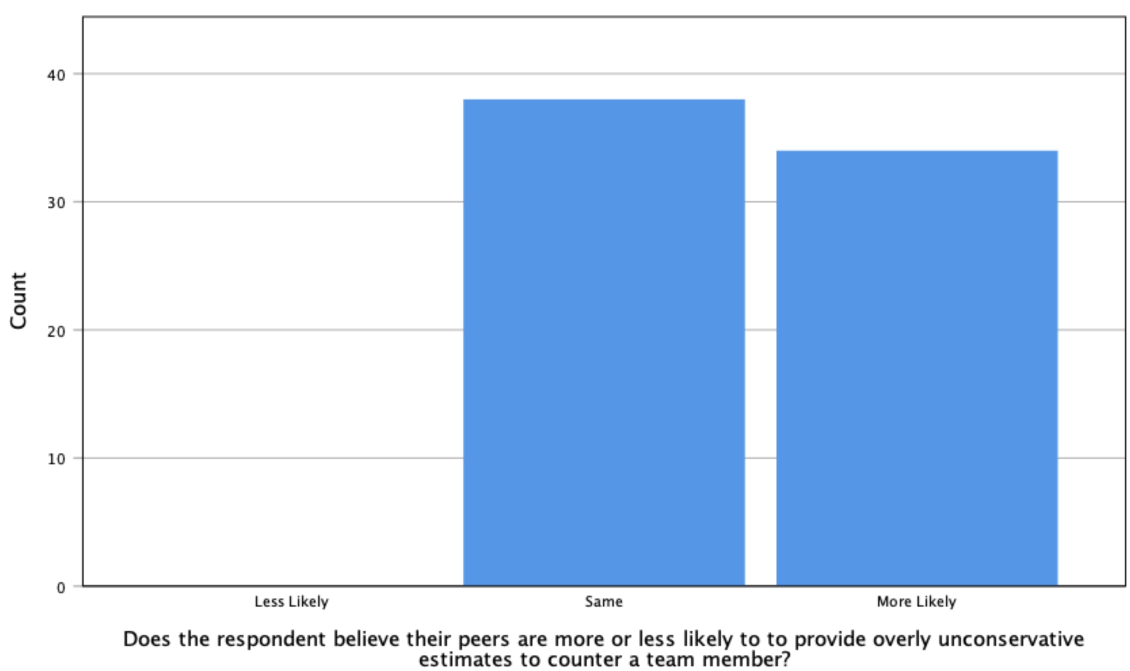

Figure 2.7: Perceived Willingness of Peers to Provide Overly Conservative Estimates vs. Self

position rather than by role or employer. A series of more detailed case studies of risk estimates would likely be required to collect data of sufficient granularity to gain further role-based insights.

2.7 Conclusion

The process of estimating the risk for dams and levees poses complex technical challenges due to the level of uncertainty present in their analysis but, perhaps more complex, are the subtleties of emotion and interaction among the estimators. Subjective risk estimation will remain an imperfect endeavor even for highly experienced and educated estimators but improvements in both process and understanding are possible and necessary. It is clear that the understanding of the motivational effects naturally present within estimators' thought processes are incompletely understood and incompletely remedied in current practice. Adequately redefining motivational bias among industry professionals from its present narrow description to a broader one will serve to improve awareness of effects currently unaddressed by regulation. Furthermore, shifting training and discussion away from purely cognitive biases to include the impacts of affective rationality and motivated reasoning may allow for improved and less biased estimates. It is also possible that identification of myside effects in group conflict and active management of conflict escalation by trained facilitators may serve to dampen extreme estimates, either conservative or unconservative, to ensure the most accurate risk estimate possible.

This work was broad in focus and likely has missed some measure of group, organizational, situational, or project-specific effects that could yield improved understanding. Future research in this area could leverage the findings of this work and design an observation process for determining the extent to which these effects impact aggregated risk estimates by observing group behavior, comparing individual node risk estimates before, during, and after aggregation, and monitoring the outcomes of presented case studies and their impact on individual node risk estimates.

Given the relatively low frequency of such risk estimates and the commitment of resources required, such work would surely prove difficult, but would likely yield significant process improvements that could be standardized by common practice or regulation.

Chapter 3

Subjective Expert Probability Aggregation for Dam Safety Risk Estimation

3.1 Introduction

In 1977, in the wake of the Teton Dam failure, U.S. President Jimmy Carter instructed all federal agencies with responsibility relating to the ”planning, construction, operation, and ultimate disposal” of dams to investigate using ”probabilistic or risk-based analysis” to evaluate the the safety of dams. This approach was counter to the standards based approach that was the state of practice at the time. In response, several federal agencies developed a framework for the Risk-Informed Decision Making (RIDM) process that now governs the risk quantification for federally regulated dams. (FEMA, 2004)

As part of the RIDM risk assessment process, the level of risk of failure of a dam must be quantified. However, by their nature, dams are complex, difficult to investigate, and objective comprehensive quantification of their condition is often impossible. As such, the risk of failure is largely quantified through subjective expert probability estimation. This subjective approach is particularly useful when analyzing systems where the rate of failures is very low, but the consequences of failure are very high, such as in the evaluation of dams and levees. (Bowles et al., 1998) The lack of objective data in such instances makes more robust evidence-based investigations infeasible and forces a reliance on subjective expert opinions to determine the probabilities of the events that could lead to a system or component failure. (Mosleh and Apostolakis, 1986; Dewispelare et al., 1995) This technique is not unique to the field of dam safety and is a critical component of the decision-making process in a wide variety of industries. (O'Hagan, 2006; Cooke and Goossens, 2008; Cooke, 1991)

3.1.1 The Dam Safety Risk Management Framework

For the specific case of quantifying risk probabilities for dams, the guidelines provided by a number of U.S. federal agencies with responsibility in the field provide an outline for the dam safety risk management framework. This is a circular process, wherein dams are periodically assessed under the direction of a decision maker. See Figure 3.1 The risks are first analyzed by a team of experts with experience with similar facilities, conditions, or construction techniques. The experts then perform a potential failure modes analysis (PFMA) where they posit a number of reasonable and appropriate failure modes for the facility being assessed. They then populate an event tree consisting of a series of discrete steps that must occur in order for the water retaining barrier to fail or some other undesirable behavior to occur for each potential failure mode. An example of an event tree as shown in DeNeale et al. (2019) is presented in Figure 3.2.

As part of the risk analysis process, risk estimation is performed to quantify the likelihood of failure for the dam. To begin, subjective probability distributions are provided by the experts for each node. The subjective probabilities typically take the form of most likely, best-case, and worst-case likelihoods for each node in the event

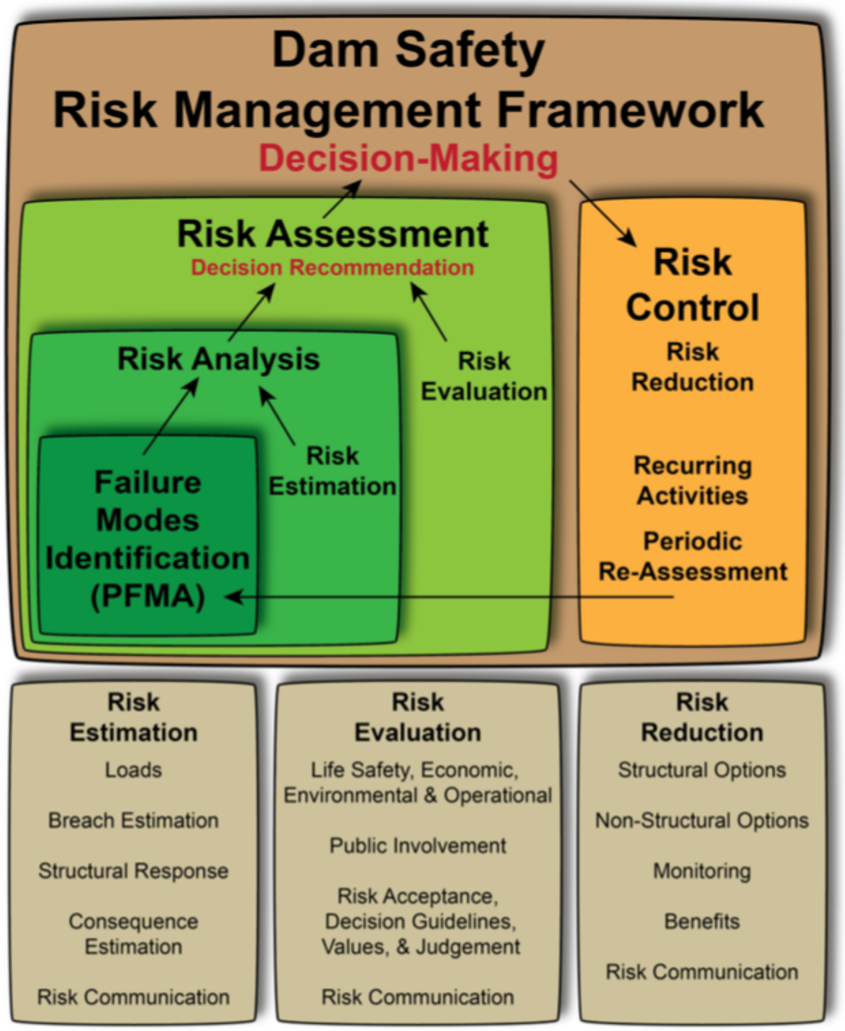

Figure 3.1: Typical Risk Management Framework From Federal Guidelines for Dam Safety Risk Management by FEMA in the public domain.

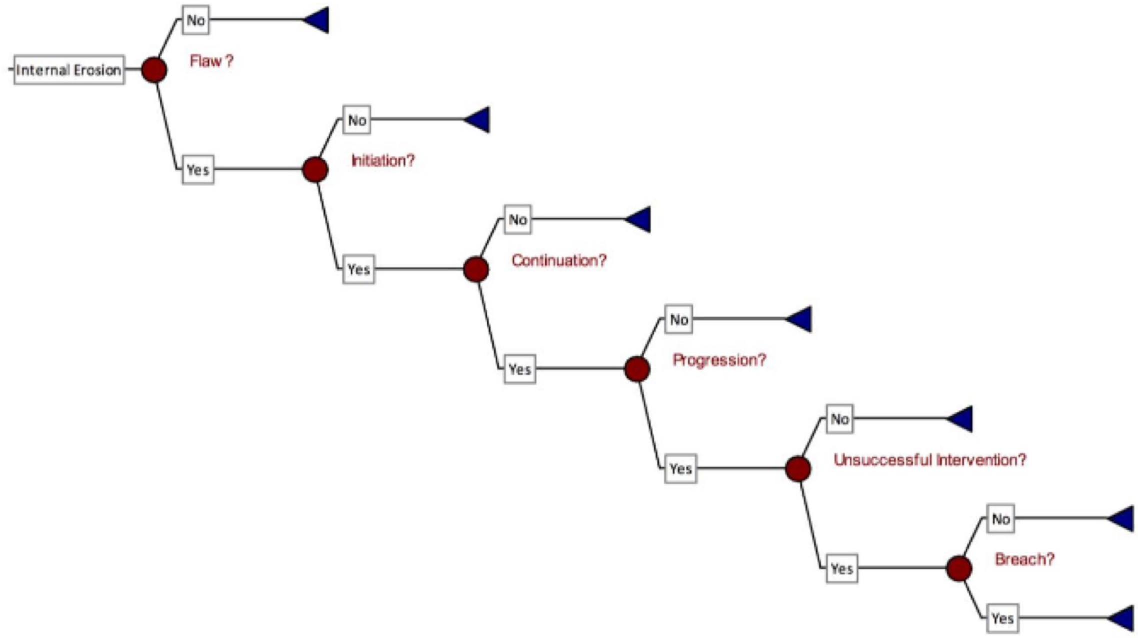

Figure 3.2: Typical Event Tree From Current State-of-Practice in Dam Safety Risk Assessment by DeNeale, et. al. in the public domain.

tree. These values are then used to generate probability density functions (PDFs) that describe each expert's estimated likelihood that an event in the event tree will occur. (USBR, 2019) These distributions are then aggregated and a system response probability curve, which is simply the aggregated cumulative distribution with the axes swapped, is generated for each node. A Monte Carlo simulation is then performed for all nodes of the event tree with values being sampled from the system response probability curves. The sampled values are then multiplied together for each potential failure mode and summed to calculate the annualized probability of failure (APF) for the facility. The Monte Carlo results are then plotted against average life loss or other negative societal impacts associated with a failure and compared to preestablished tolerable risk guidelines. An example of a f-N chart showing tolerable risk guidelines can be seen in Figure 3.3. This data is then combined into a comprehensive report including a narrative describing the risk analysis effort and recommendations for remedial action if required. This assessment is then used by the decision maker to

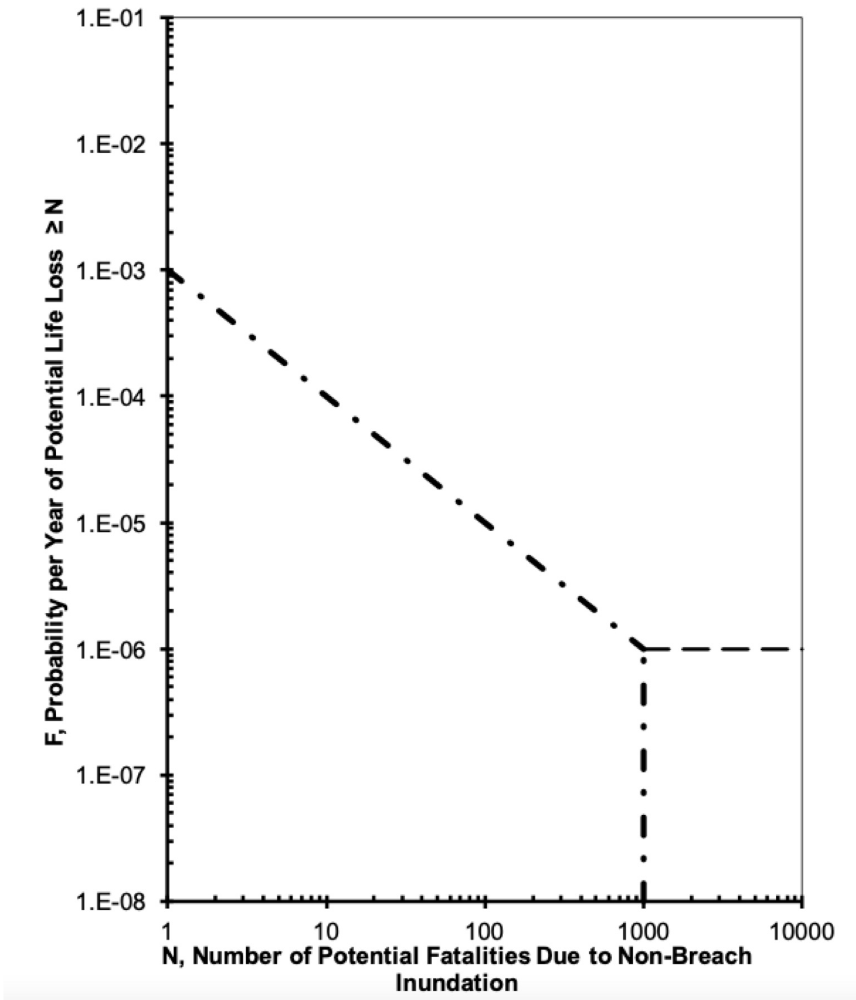

Figure 3.3: Typical Risk Management Framework From Safety of Dams - Policy and Procedures by USACE in the public domain.

inform risk control efforts and the process repeats at some regular cadence.(FEMA, 2004, 2015; FERC, 2016; USACE, 2014; USBR, 2019)

Similar processes are used to quantify marine accident risk (Nguyen, 2019), nuclear power accident risk (Sezen et al., 2019), the risk of military operations (Merrick, 2011), port security risks (Bakir, 2008), and risks in many other industries.

The risk estimation portion of the framework, particularly the aggregation of the experts' subjective probability distributions, will be the primary focus of this paper.

3.1.2 Behavioral Aggregation

As there are typically multiple experts providing subjective risk estimates, the question of how to aggregate the expert opinions must be answered. (Hammitt and Zhang, 2013) Two primary approaches exist for aggregation of expert opinions: behavioral and mathematical. Behavioral techniques encourage collaboration among the expert estimators to establish a group consensus opinion. (Clemen, 1989) A variety of methods are described in the literature. Among the earliest is the Delphi method, in which the experts anonymously rate the estimations of their peers then each expert has the opportunity to reassess their own estimate after considering the score given to them by their peers. This may iterate several times with the hope of reducing the range of estimates. (Ouchi, 2004) While Delphi was somewhat effective, similar techniques allowing direct interaction between estimators prior to re-estimation were shown to be superior. (Gustafson et al., 1973) This "Estimate-Talk-Estimate" process aligns with the iterative estimation process as described in USBR (2019). Allowing the interaction of experts within the process is not without risk of intimidation, manipulating, bluffing, or other effects of motivation or personality. (Genest and Zidek, 1986; Mosleh et al., 1988) The federal guidelines for RIDM require trained, experienced facilitators to monitor for such disruptive behaviors and control the process to minimize their effect.

3.1.3 Mathematical Aggregation

While behavioral aggregation is a useful technique for discussing, challenging, and updating experts' estimates, reaching a true group consensus by means of a single unified PDF for all nodes in an event tree using this method is extremely unlikely. Therefore, once the behavioral aggregation process has provided what benefit it can, the individual subjective estimates must still be combined into a single PDF by some mathematical means. Many methods for mathematical aggregation have been explored but, ultimately, they produce an aggregated point probability or PDF that combines all the expert opinions while allowing for the optional weighting of each expert's opinion based on some criteria. (O'Hagan, 2006)

According to an interview of industry professionals who act as decision makers for dam safety risk issues conducted in December of 2023 (See Appendix B.1), a particularly simple method is commonly used to aggregate risk likelihood estimates for each node. In this method, a facilitator collects the most likely, best case, and worst case likelihood estimates from each expert. Each of the three sets are then averaged and the resulting three values are then used to generate a single triangular distribution using the piecewise defined function:

$$f(x) = \begin{cases} 0 & \text{for } x < \bar{w} \\ \frac{2(x-\bar{w})}{(\bar{m}-\bar{w})(\bar{b}-\bar{w})} & \text{for } \bar{w} \leq x < \bar{m} \\ \frac{2(\bar{b}-x)}{(\bar{b}-\bar{w})(\bar{b}-\bar{m})} & \text{for } \bar{m} \leq x < \bar{b} \\ 0 & \text{for } x \geq \bar{b} \end{cases} \tag{3.1}$$

where $\bar{w}$ is the average of the worst case estimates from all experts, $\bar{m}$ is the average of the most likely estimates from all experts, and $\bar{b}$ is the average of the best case estimates from all experts.

This method will be referred to as unweighted initial averaging hereafter. This technique has the obvious benefit of simplicity. However, it can only provide a range of likelihood estimates equal to that of the experts if perfect consensus exists for the best and worst case likelihoods. In practice, the range of likelihoods is nearly always reduced, providing an artificially narrow range that causes the aggregated distribution to convey a higher level of certainty than actually exists within the group. While some decision makers interviewed understand the limitations of this method, it is regarded as acceptable for assessments of facilities where the issues being assessed are subjectively determined by the decision maker to be less complex or less serious in nature. As such, its ease of use and computational simplicity result in its continued common use.

The industry experts interviewed also stated that a more sophisticated linear pooling method is also commonly used. Linear opinion pooling is a popular method, dating back as far as Laplace, for aggregating subjective probability judgments from multiple individuals, experts, or sources into a single coherent probability distribution. (Bacharach, 1979) The method involves combining individual probability assessments by assigning weights to each judgment, and taking a weighted average of the probabilities. (Stone, 1961; Ranjan and Gneiting, 2010; O'Hagan, 2006; Genest, 1984a) It should be noted that expert weighting is not commonly practiced in dam safety risk assessments. In this case, the process is the same, but with equal weights applied to all experts.

Linear opinion pooling is described by the formula:

$$P(x) = \sum_{i=1}^{n} w_i p_i(x) \tag{3.2}$$

where $P(x)$ is the pooled estimate of the probability of event x, and $p_i(x)$ is the opinion of expert i on the probability of event x and w_i represents the weights assigned to each expert. The weights must satisfy:

$$\sum_{i=1}^{n} w_i = 1 \quad \text{and } w_i > 0 \text{ for all } i \tag{3.3}$$

While the calculation of the linearly pooled probability distribution is less intuitive and more complex than unweighted initial averaging and generation of a single unimodal distribution, it is easily achievable and computationally efficient. This method, as commonly practiced, is not without its downsides. While unweighted arithmetic averaging of PDFs does not result in a complete loss of information in some regions of the combined PDF as with unweighted initial averaging, the natural tendency of arithmetic averaging is to amplify those estimates nearer the mean and reduce the influence of those estimates that are relatively more or less conservative. This tendency may be desirable to some decision makers who are interested in a consensus decision. It may, however, serve to reduce the impact of some likelihood estimates that more closely match reality than the consensus but are only held by a minority of estimators. It is also the case that large numbers can dominate the arithmetic average, potentially providing dissenting overly-conservative expert opinions with an outsized influence. The resulting average at any point would still tend towards the mean, but more conservative estimates can tend to dominate less conservative estimates, particularly across multiple orders of magnitudes as is the case in dam safety risk estimates.

While not commonly used in dam safety risk assessments, Log-linear opinion pooling is another well-understood method for combining opinions from multiple sources. The method uses geometric averaging of the PDFs that represent the opinions of each expert, with the opportunity to apply weights to those experts' subjective probabilities. Log-linear pooling can ameliorate the dominating effect of large numbers as in arithmetic averaging. (Parkhurst, 1998) However, when the individual subjective probability density functions have values of zero or very nearly zero for large portions of the domain, these areas have a nullifying effect on their peer functions. Once multiplied through, these near-zero values produce an aggregated function with large peaks where the estimates overlap significantly with other regions being driven to zero or very nearly to zero when at least one estimate is at or near zero. (de Carvalho et al., 2020) It is not obvious that this is desirable. Given that the federal guidelines describe sample distributions that are triangular or rectangular in shape and industry practice is moving towards the use of PERT distributions, the most commonly used distributions result in substantial areas of zero, or near-zero, probability for each expert estimator. (USBR, 2019)

Log-linear opinion pooling is described by the formula:

$$P(x) = \frac{1}{Z} \prod_{i=1}^{n} p_i(x)^{w_i} \tag{3.4}$$

where $P(x)$ is the pooled estimate of the probability of event x, and $p_i(x)$ is the opinion of expert i on the probability of event x, Z is a normalizing constant, and w_i represents the weights assigned to each expert. (Genest, 1984b) The weights must satisfy:

$$\sum_{i=1}^{n} w_i = 1 \quad \text{and } w_i > 0 \text{ for all } i \tag{3.5}$$

These linear and log-linear approaches may be generalized in the form:

$$P(x) = c \left(\sum_{i=1}^{n} w_i (p_i(\theta))^{\alpha} \right)^{\frac{1}{\alpha}} \tag{3.6}$$

where:

$$c = \frac{1}{\int \left(\sum_{i=1}^{n} w_i (p_i(\theta))^{\alpha} \right)^{\frac{1}{\alpha}} d\theta} \tag{3.7}$$

This generalized form varies in output for different values of α with $\alpha = 1$ describing the linear opinion pooling function and $\alpha \to 0$ describing the log-linear pooling function. (Cooke, 1991) This generalized form is equivalent to the weighted Hölder mean and following the convention of Koliander et al. (2022), will be referred to as the Hölder pooling function. This form was selected for simplicity and flexibility. It allows for consistent calculation in a single computational tool for both linear and log-linear pooling by varying the value of α between values of 1 and $\to 0$ respectively. It can also provide otherwise unavailable intermediate aggregated PDFs, if desired, between both methods by varying the value of α between 0 and 1.

Other pooling methods may be used in lieu of linear or logarithmic pooling, such as a number of Bayesian techniques covered at length in Clemen and Winkler (1999). These methods rely on the decision maker to form their own distribution and update that distribution based on the input of the expert estimators. (O'Hagan, 2006) This process is not commonly used in dam safety RIDM and will not be explored here.

In addition to determining the appropriate aggregation technique, decision makers must consider the question of weighting the subjective opinions of the experts. This is not typically done in current practice, though the decision makers interviewed were open to the prospect of adjusting weights during pooling.

While many methods have been proposed for the weighting schemes, many well-understood techniques tend to focus on the weighting of each expert's subjective probabilities in a way that seeks to provide additional influence to the opinions of those experts who, by way of some calibration process, seem to provide the most accurate answers. (Clemen and Winkler, 1999; Genest and McConway, 1990) Methods such as the classical model Cooke (1991); Cooke and Goossens (2008) and other similar hierarchical techniques require either the elicitation of seed variables for expert ranking or a running history of expert forecast data to determine the relative quality of each expert's opinions. Such techniques are of little benefit for low rate, high consequence failures because the true likelihoods of events initiating or leading to a system failure are subject to high levels of aleatoric and epistemic uncertainty and rarely, if ever, provide a reference by which one could measure the quality of subjective expert opinions. (Vick, 1999) Additionally, a number of alternate weighting methods and more complex Bayesian approaches have been presented in the literature. However, as with the Bayesian pooling methods, these approaches are not commonly employed in RIDM for dam safety and will not be considered here. See de Carvalho et al. (2020) for detailed discussion of these techniques.

Perhaps the simplest weighting method is to equally weight all of the experts. This is often the method selected to ensure the perception of fairness. However, if it is the desire of the decision maker to equally consider the experience and expertise of each expert, any weighting approach that does not balance the relative impact of each expert on the aggregated distribution in some way, may provide results that fail to fairly represent the experience of each estimator. Any method that provides results counter to this desire is philosophically inconsistent with such a decision maker's approach, even if it is not well understood by the decision maker that the pooling process is providing such results. In this case, optimization techniques that minimize or maximize some objective function related to the difference between the individual expert's distributions and the aggregated distribution may be more appropriate than equalizing expert weighting prior to aggregation. Work in this area describes optimization methods to minimize the weighted average of various divergence measures including f-divergence and Kullback-Leibler (KL) divergence. (Koliander et al., 2022; Garg et al., 2004) Similarly, Abbas (2009) proposes the use of an optimization technique that maximizes a scoring function in the context of weighted average Kullback-Leibler divergence minimization. These methods seek to construct an aggregated probability function that reduces the total amount of disparity between each individual probability function and their aggregate. However, these approaches may not necessarily align with a decision maker's desire for "fairness" as it could be possible to construct a weighted aggregate that optimizes some KL-divergence metric but, in doing so, results in the dominance of a single or a small number of individual estimates. In such an instance, an alternate optimization approach that seeks to minimize the maximum individual KL divergence may be more appropriate.

3.2 Aims of the Present Research

This work will use a set of anonymized data from a recent dam safety risk assessment to compare the results of the unweighted initial averaging and unweighted linear pooling methods in common use in industry as well as unweighted log-linear pooling to determine if log-linear pooling should be considered for future dam safety risk assessments. Additionally, a pair of optimization techniques that will seek to minimize the maximum KL-divergence between the highest expert measure and the aggregated function for both the linear and log-linear pooling methods will also be explored.

The first goal of this work is to qualitatively understand the data loss and resulting distribution curve shape as well as the mean and range for the Monte Carlo simulation output for the entire event tree for each method explored. The second goal is to understand the relative performance of each method in terms of KL-divergence and the relative representativeness of each pooling method. Thirdly, the effects of the KL-divergence optimization techniques will be compared to the unweighted opinion pools and their suitability for consideration for future dam safety assessments will be explored. Finally, the decision makers interviewed expressed a desire for a tool to objectively analyze the experts' responses to determine if outlier opinions at the node level were indicative of a genuine extreme dissent or if these opinions were indicative of the experts' general tendencies. To satisfy that desire, the suitability of KL divergence analysis will be explored as a potential tool for understanding relative expert performance, determining critical nodes for further exploration, and crafting a more complete narrative to aid in the decision making process.

3.3 Methodology

The approach to comparing the various aggregation methods began with the generation of expert-specific probability distributions based on existing input data. These distributions were then aggregated using a combination of unweighted and optimized weighting techniques, with the goal of minimizing the divergence between individual expert estimates and the final pooled distribution. Various pooling methods were explored, and the KL divergences for each method were calculated to assess how representative the pooled PDFs were for the expert inputs. A modified optimization process also used to address cases where one expert's opinion dominates the aggregated result. The methodology concludes with a Monte Carlo simulation that samples the optimized pooled distributions to determine the expected ranges for the overall system's failure probability. Finally, a method for determining the importance of outliers opinions based on the general tendencies of the experts was developed and proposed. Each step in this process is described in detail in this section.

3.3.1 Distribution Generation

To develop and test the various methods, distributions typical of what experts would provide in practice were required. Distributions elicited from experts as part of a recent dam safety risk estimation were obtained from a federal dam owner. To provide anonymity for sensitive data, the owner, facility, failure initiating events, expert names, and other sensitive information was removed from the data. Only the expert event likelihood information and the event tree structure was retained. This provided a robust pool of recent estimations from which aggregated distributions could be generated. The input data took the form of three distinct values from each estimator at each node of the event tree: the worst case, best case, and most likely probability of the event occurring. It should be noted that the values used were the final values provided by the experts after multiple rounds of estimation with behavioral aggregation techniques applied by an experienced facilitator between rounds.

For the unweighted initial averaging method, the best, worst, and most likely values were used to generate expert-specific triangular distributions as described in Equation 3.1. For all other methods, those same values were used to define a normalized PERT distribution, as is becoming more common in current practice, rather than a simpler geometric distribution as described in USBR (2019) and as was used at the time of the risk estimation from which the data was collected. It should be noted that in some cases, experts provided the same value for all three of their data points. This was allowed during the risk assessment when the data was collected. Since data of this form will not produce a satisfactory triangular or PERT distribution, the best case and most likely estimates were reduced by 0.002 and 0.001 respectively to achieve a meaningful distribution. For instance, an estimate of (1, 1,

1) would become (0.998, 0.999, 1). This allows use of a technique preferred in the current state of practice without significantly altering the elicited values. Examples of individual expert triangular and PERT distributions are shown in Figure 3.4 and 3.5 respectively.

3.3.2 Equal Weighting, Pooling, and Divergence Calculation

For unweighted initial averaging, the output of the distribution generation process was the pooled PDF and no further aggregation was required. The KL-divergences were then calculated between each individual distribution and the aggregated distribution using Equation 3.8 and those results were recorded. It should be noted that, as the triangular distribution allows areas of the distribution to represent a likelihood of zero, a small Laplace smoothing constant of 1.0×10^{-9} was added to the aggregated distribution to prevent a division by zero error when calculating the KL-divergences.

For the unweighted linear and log-linear pooling methods, a vector was populated with equal weights as starting points for the pooling functions. The Hölder pooling function was then applied to the expert distributions and an initial aggregated distribution was obtained using both linear and log-linear pooling, employing α values of 1.0 and 0.01 respectively. The KL-divergences between each expert distribution and the aggregated distribution for each pooling method were calculated each using the equation:

$$D_{KL}(P_i \| Q) = \sum_{x \in \mathcal{X}} P_i(x) \log \frac{P_i(x)}{Q(x)} \tag{3.8}$$

where $Q(x)$ is the pooled probability density function.

3.3.3 Maximum Individual KL-Divergence Minimization

In order to more fairly represent the input of each expert in the aggregated probability density function, an optimization method was developed to minimize

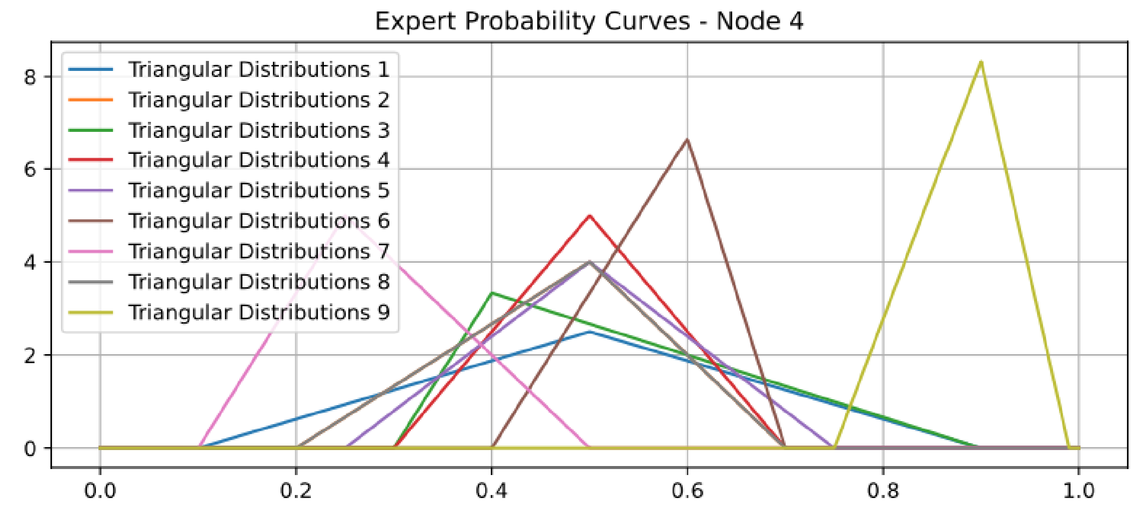

Figure 3.4: Typical Expert Triangular Probability Distributions

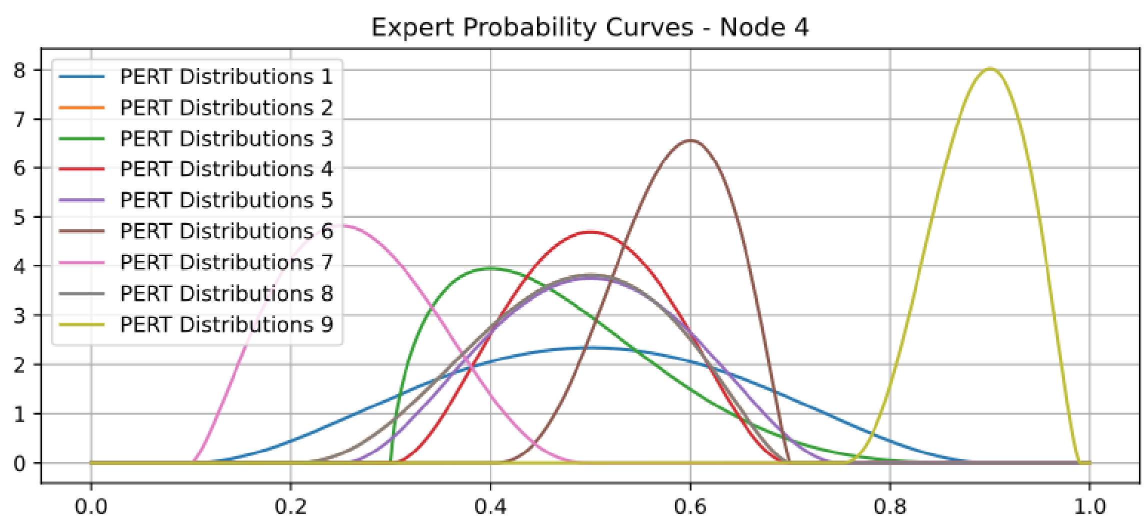

Figure 3.5: Typical Expert PERT Probability Distributions

the highest individual KL-divergence value when comparing each individual PDF to the weighted aggregated PDF. This was done by defining the difference between the maximum individual KL-divergence as the objective function and minimizing that function's value by varying the weights associated with each individual estimate in the generalized Hölder pooling function. The SLSQP (Sequential Least Squares Programming) method was used to perform the optimization. This approach was selected due to its inclusion in the commonly used SciPy library for the Python programming language and its ability to provide solutions for constrained decision variables with a highly non-linear objective function. The problem was defined as:

$$\begin{aligned} &\min_{w_1, w_2, \ldots, w_n} \quad \max_i \sum_{x \in \mathcal{X}} P_i(x) \log \frac{P_i(x)}{Q(x)} \\ &\text{subject to } 0 < w_i \leq 1 \quad \text{for all } i, \\ &\sum_{i=1}^{n} w_i = 1 \end{aligned} \tag{3.9}$$

where $Q(x)$ is the pooled probability density function.

Once the optimization algorithm found a solution, the optimized pooled PDFs and system response curves were plotted for each method. This technique often provided optimized solutions, wherein a single opinion dominated. To address this behavior, an arbitrary weight reduction limit of $0.5(w_i)$ was imposed as a lower bound to maintain a minimum level of influence for each opinion adn the optimization procedure was repeated. These revised weights were then used to generate the final aggregated PDF for the limited case and the system response probability curve for each node in the event tree. Examples of optimized linear pooling PDFs and system response curves, for both the limited and unlimited cases, are shown in Figure 3.6. A 1000-iteration Monte Carlo simulation then sampled event likelihoods at each node and the likelihood values were multiplied together to determine the expected ranges for the annualized probability of failure.

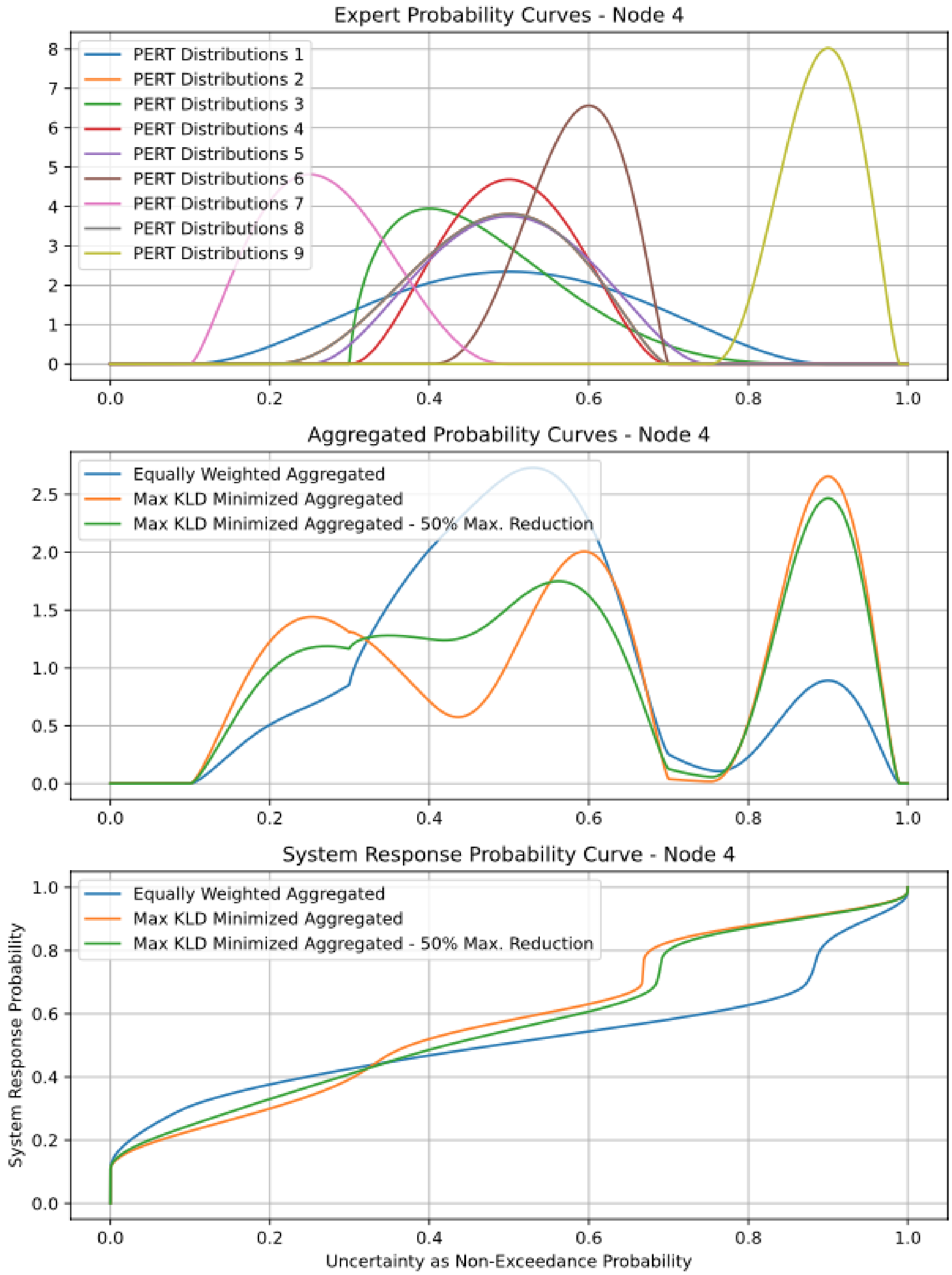

Figure 3.6: Typical Optimized Aggregated PDFs and System Response Curves

3.3.4 Expert Estimate Outlier Analysis

In an effort to satisfy the desire of industry experts to understand which outlier expert estimates merit further consideration, a two-part process was undertaken. First, the tendencies of each expert were compared to those of their peers in the broadest terms possible. To accomplish this, the mean value of the PERT distribution for each expert was calculated at each node. Those values were then used to determine the rank order of each expert at each node, with a lower rank order corresponding to a lower estimated likelihood of event occurrence. The node-specific rank orders for each expert were then averaged across all nodes to determine each expert's relative average rank order. From this set of of values, a corresponding set of Z-scores and percentiles based on the normal distribution were then calculated. (See Table 3.1.) This provided some indication of an expert's tendency to score some event more or less likely than their peers for the risk assessment being performed. Broader interpretation of such data across multiple risk assessments may be of some interest to decision makers who manage a fleet of facilities and undertake many risk assessments using a large set of experts. This may be possible, but the complexities of the interpersonal interactions during behavioral aggregation for each assessment may complicate the such an analysis and make a more universal understanding of expert tendencies unachievable. As this study does not have a broader set of estimates across multiple risk assessments, such an implementation was not considered here.

Once individual expert tendencies were determined generally, the node specific estimates were considered. For the purposes of this analysis, the KL-divergences between the individual node-specific PERT distribution estimates and the equally weighted linearly pooled distribution for that node were chosen as the measure of comparison. As with the PERT means, a set of Z-scores and percentiles corresponding to the KL-divergences were then calculated, with higher Z-scores and percentiles corresponding to more divergent estimates. Percentiles of 95% and higher were considered outlier and were flagged for consideration. Once an individual estimate

Table 3.1: Expert PERT Mean Rank Order and Percentiles

	Expert Average Rank Order	Z-Score	Percentile
Expert 3	3.41	-1.87	3.09
Expert 8	4.18	-0.96	16.83
Expert 5	4.64	-0.43	33.47
Expert 2	4.68	-0.37	35.43
Expert 7	5.27	0.32	62.56
Expert 4	5.41	0.48	68.45
Expert 6	5.45	0.53	70.32
Expert 1	5.91	1.07	85.71
Expert 9	6.05	1.23	89.02

outlier was identified, the mean of the PERT distribution for that estimate was compared to the group average of all PERT distribution averages for that node. If the value was lower than the group average, it was considered an unconservative outlier and, likewise, if the value was higher than the group average, it was considered a conservative outlier.

Once the outliers had been identified and their conservatism had been determined, that information was then compared to the individual experts' general tendencies. If an unconservative outlier corresponded to an expert whose general tendency percentile was above 50% for conservatism, the outlier was noted. This was also performed for the inverse case. These cases are worth investigation by the risk estimate facilitator and the decision makers as they represent a strong dissent from the group, even after multiple rounds of behavioral aggregation. They represent those cases where an expert not only dissents strongly from the group, but also provides an estimate in a direction counter to their own natural tendency. This may represent a difference in thinking, knowledge, or experience that may not have been sufficiently communicated or understood by the larger expert group. Facilitators and decision makers should investigate this issue and be certain that they understand the reason for the dissenting opinion. Such estimates may indicate the need for further discussion, engineering, or even field investigation. If no resolution can be found among the expert team, the risk report should include, in its narrative, a description of the condition and the outlying opinion. Such outliers may not be readily identifiable using data visualization.

3.4 Results and Discussion

Upon implementation of the various techniques described in Section 3.3 on a set of sample data, a number of tendencies were observed that distinguish the methods and the relative utility of their output.

3.4.1 Unweighted Initial Averaging

The unweighted initial averaging method, as expected, resulted in a rather significant loss of data. That is to say the aggregated distribution contained areas of zero probability where at least one of the estimators had assigned some positive value. The overall range of likelihood values was reduced, on average, more than 48% with range reductions above 60% in some cases. The reduction on the worst case end of the distribution averaged around 20%, resulting in a loss of conservatism. The increase on the best case end of the distribution averaged over 365% with multiple best case probabilities increasing by three orders of magnitude, thus increasing conservatism. (See Table 3.2) As the f-N plots typically span multiple orders of magnitude, this represents significant data loss. Once the Monte Carlo simulation was completed, the unweighted initial averaging method resulted in a mean value that was higher than the unweighted linear and log-linear pooling techniques. However, it is unclear if this data loss would reliably increase or decrease conservatism. The range of Monte Carlo results was much smaller than the unweighted linear results, but somewhat larger than the unweighted log-linear results. (See Figure 3.7)

Table 3.2: Node Specific Information Loss Using Unweighted Initial Averaging

Node	Expert - Worst Case Maximum	Expert - Best Case Minimum	Averaged Worst Case	Averaged Best Case	Expert Range	Averaged Range	Range Reduction	% Reduction - Worst Case	% Increase - Best Case
1	0.999	0.500	0.967	0.776	0.499	0.191	0.617	3.2%	55.3%
2	0.999	0.500	0.953	0.700	0.499	0.253	0.492	4.6%	39.9%
3	0.999	0.500	0.959	0.733	0.499	0.226	0.548	4.0%	46.6%
4	0.990	0.100	0.760	0.289	0.890	0.471	0.471	23.2%	188.9%
5	0.700	0.010	0.428	0.136	0.690	0.292	0.576	38.9%	1255.6%
6	0.900	0.100	0.628	0.233	0.800	0.394	0.507	30.2%	133.3%
7	0.900	0.100	0.611	0.239	0.800	0.372	0.535	32.1%	138.9%
8	0.800	0.100	0.650	0.211	0.700	0.439	0.373	18.8%	111.1%
9	0.500	0.010	0.361	0.107	0.490	0.254	0.481	27.8%	966.7%
10	0.500	0.010	0.400	0.111	0.490	0.289	0.410	20.0%	1011.1%
11	0.800	0.100	0.628	0.172	0.700	0.456	0.349	21.5%	72.2%
						Average:	0.487	20.4%	365.4%

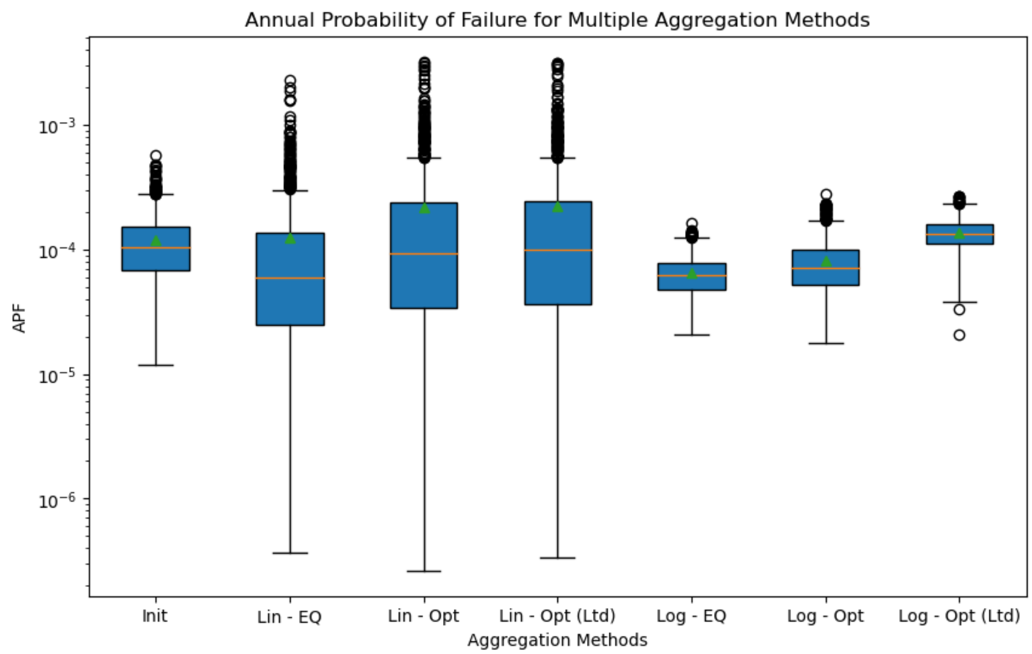

Figure 3.7: Annual Probability of Failure from Monte Carlo Simulation

These results suggest the unsuitability of this method for several reasons. Firstly, this narrowing of the output range could easily be incorrectly interpreted by a decision maker as a sign of a high level of certainty among the experts. This could also lead a decision maker to conclude that the likelihood of an event occurring as calculated by the Monte Carlo simulation never exceeds the tolerable risk guidelines as shown on the f-N plot based on the experts' likelihood estimations. Unweighted initial averaging also produced KL divergence values between the expert probability distributions and the pooled distribution that were quite high when compared to other methods considered, both in maximum value and also in their range. This shows that this method provides a pooled function that is neither particularly representative of any individual estimator nor the group as a whole.

3.4.2 Log-Linear Pooling

The log-linear pooling method produced aggregated distributions with substantial areas of near-zero likelihood. This was due to the fact that the individual PERT distributions contained regions of near-zero probability and, when aggregated using geometric averaging, the regions of very small probability dominated even the largest probability estimates in those regions, reducing them to a very small aggregated probability locally. For the PERT distribution or the geometric distributions described in the federal guidance documents, this method eliminates a significant amount of data that may represent real risk likelihood. This effect is shown in Figure 3.8.

This technique, when used with probability density functions with limited or no regions of zero or near-zero likelihood or with distributions with substantial overlap, may be quite appropriate. In such cases, the log-linear aggregation technique essentially serves to remove those regions of likelihood where the group does not have consensus agreement that failure likelihood exists. Activities or systems with similar ranges of expert uncertainty or where only the consensus view is of interest may benefit from such a technique. For this data set, in the context of dam safety risk estimation, the log-linear results naturally produced an aggregated PDF that resulted in large KLD values, both for the maximum values as well as the range of values. As with unweighted initial averaging, this meant that the aggregated PDF was not particularly representative of any single estimate nor did it produce a result that was particularly representative of the set of estimates.

The optimization techniques were able to produce aggregated PDFs that resulted in much lower maximum KLD values and KLD ranges, occasionally more than 80% reduction in both. However, the nature of the pooling method meant that even large reductions in KLD values did not produce changes that appeared meaningful, consistent, or particularly useful. Furthermore, even the optimized KLD values were still well more than an order of magnitude higher than those achieved with the linear pooling method. It should be noted that KL-divergences are not distance measurements, but significantly larger values suggest a less representative pooled function. When taking each node individually and comparing the rank of each estimate against the rank of each set of weights produced by the optimization

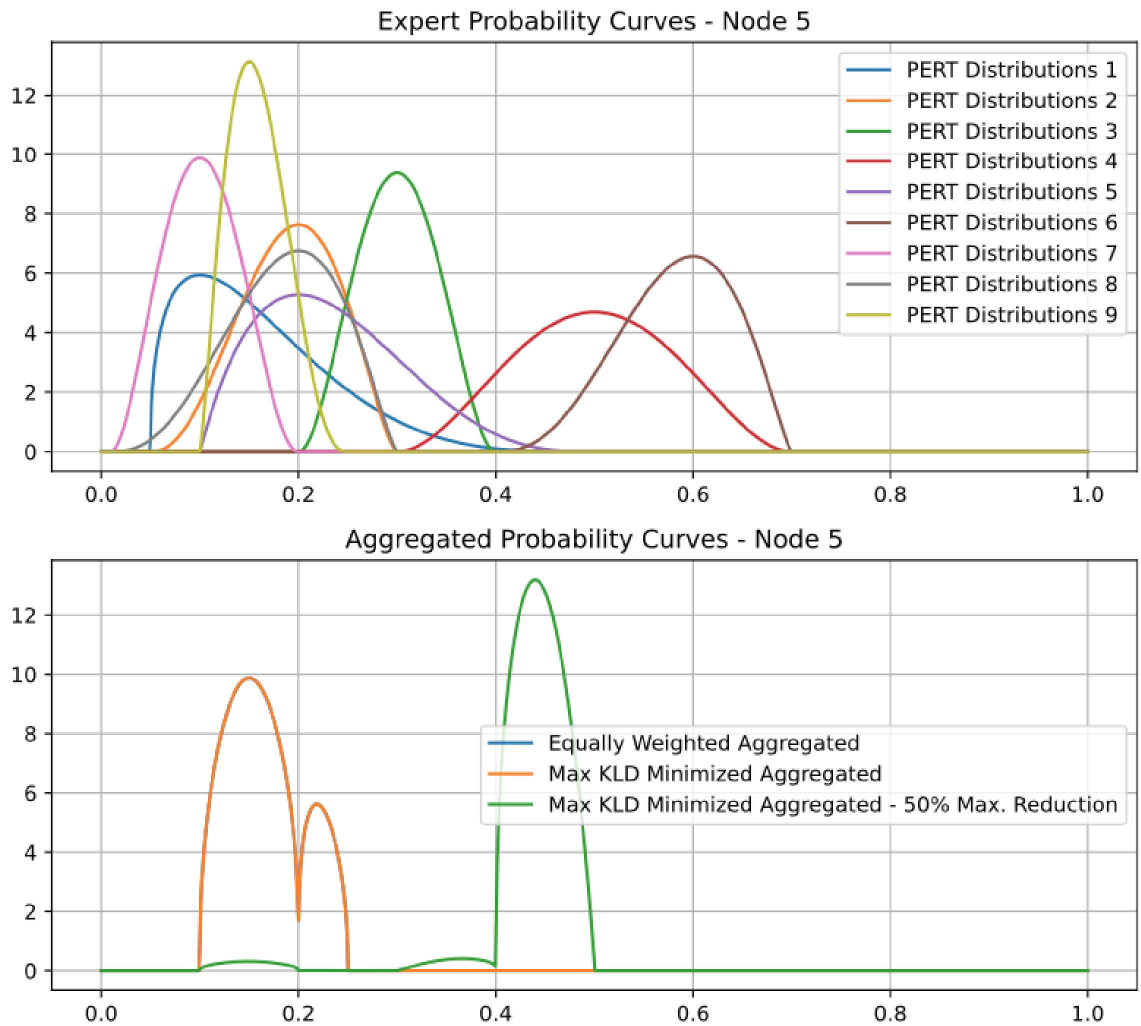

Figure 3.8: Typical Log-Linear Aggregated PDF

techniques, both limited and unlimited lower bounds, using Spearman's Rank Correlation Coefficient, there was only sporadic correlation and the results showed no discernible pattern in the redistribution of weights. See Table 3.3.

3.4.3 Linear Pooling

The linear pooling method produced an aggregated function that reflected the input of all individual estimates with likelihood values present in the aggregate at all locations present in the individual functions. However, the natural tendency of this method increased the maximum value of the aggregated function in those areas where there existed a high level of overlap between the individual estimates and served to reduce the maximum value of the aggregate in those areas where a single estimator provided a unique likelihood value. This tendency reduced the impact of relatively extreme estimates, even estimates with relatively high levels of certainty, in the pooled function. This effect is shown in the equally weighted aggregated probability curve in Figure 3.6. This may not be desirable for a decision maker as reducing the impact of a highly confident dissenting expert, unconvinced by multiple iterations of behavioral aggregation, relative to the consensus may eliminate valuable information. However, this effect does seem to have a limit. There existed in the data instances of extraordinarily high individual values as described in Section 3.3.1. In these instances, which essentially represented perfect certainty of the estimator, the linearly pooled function was highly influenced by this single estimate as is shown in Figure 3.9. However, these were the most extreme cases and are not indicative of the performance of this technique for more typical levels of expert certainty.

Table 3.3: Log-Linear Equally Weighted KLD Rank vs. Optimized Weight Rank Correlations

Case	Equally Weighted KLD Rank vs. Optimized Weight Rank		Equally Weighted KLD Rank vs. Optimized Limited Weight Rank	
	Spearman's Rho	Significance	Spearman's Rho	Significance
Node 01 - Log	0.627	0.071	-	-
Node 02 - Log	0.36	0.342	0.628	0.07
Node 03 - Log	-	-	0.142	0.715
Node 04 - Log	0.728	0.026	0.979	<0.001
Node 05 - Log	-	-	0.383	0.308
Node 06 - Log	-	-	-	-
Node 07 - Log	-	-	-	-
Node 08 - Log	0.642	0.062	0.343	0.366
Node 09 - Log	0.745	0.021	-	-
Node 10 - Log	0.477	0.194	0.695	0.038
Node 11 - Log	-	-	0.427	0.252

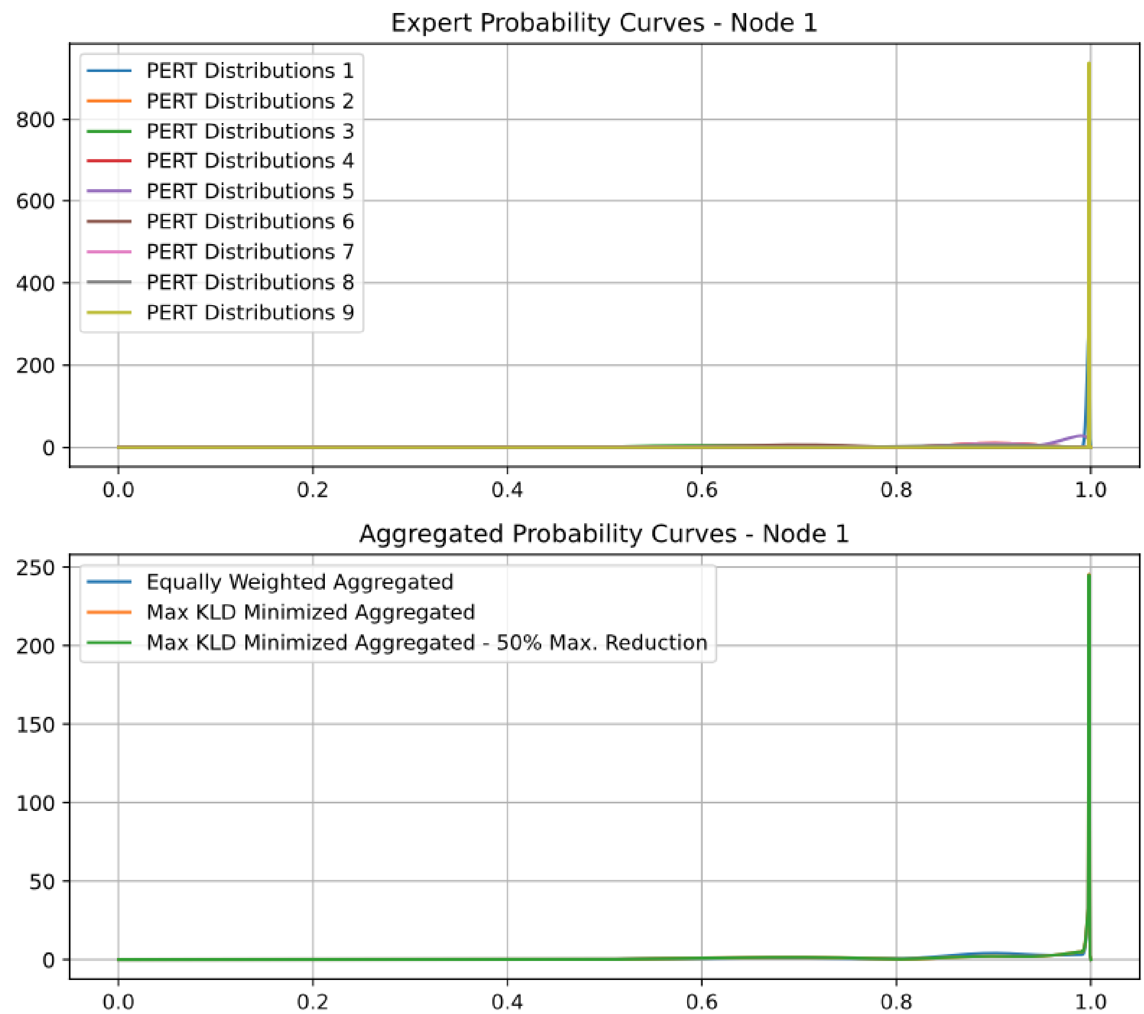

Figure 3.9: Expert and Aggregated Distributions with a Single Estimate of High Certainty

As expected, upon modifying the weights using the optimization procedure with no lower bound on weights, both the maximum KLD for each node and the range of KLDs across each node were reduced. This resulted in an aggregated PDF that more evenly represents the expert estimations that supports the presumed axiomatic desire for ”fairness” in the aggregated function. (See Figure 3.6) However, in all cases, at least one weight per node was reduced by at least 70%, which may be undesirable, as the process greatly reduces the input of at least one estimator. In some cases, all but three of the expert weights were reduced to a value approaching zero with nearly all of the remaining weight being divided among the other experts. For the optimization technique that was limited with an arbitrary individual expert lower weight bound of 50% of the equal weight, a noticeable reduction in both maximum KLD and KLD range were achieved. While not as pronounced as in the unlimited reduction method, maximum KLD reductions ranged between 80-99% as effective as their unbounded counterparts. Similar KLD range reductions were also achieved. The dominating behavior of a small subset of the expert weights was still present, but was limited by the additional optimization constraint. Further exploration of this behavior and a method for determining the best reduction limit may be warranted for a range of expert panel sizes in future work. The expert KLDs for the linearly pooled, equally weighted case can be seen in Figure 3.10 with the expert KLDs for the linearly pooled optimized methods, both unlimited and limited, may be see in Figures 3.11 and 3.12 respectively. Likewise, the weights for both linearly pooled optimized methods may be seen in Figures 3.13 and 3.14.

Those estimates that initially had a relatively high KLD value tended to have a higher weight after optimization. For the unlimited lower bound method, the rank of the initial KLD value positively correlated with the rank of the optimized weight for nine of eleven nodes ($p < 0.05$). The rate was slightly less for the unlimited reduction method with six of the eleven node positively correlating for the same ranks. See Table 3.4.

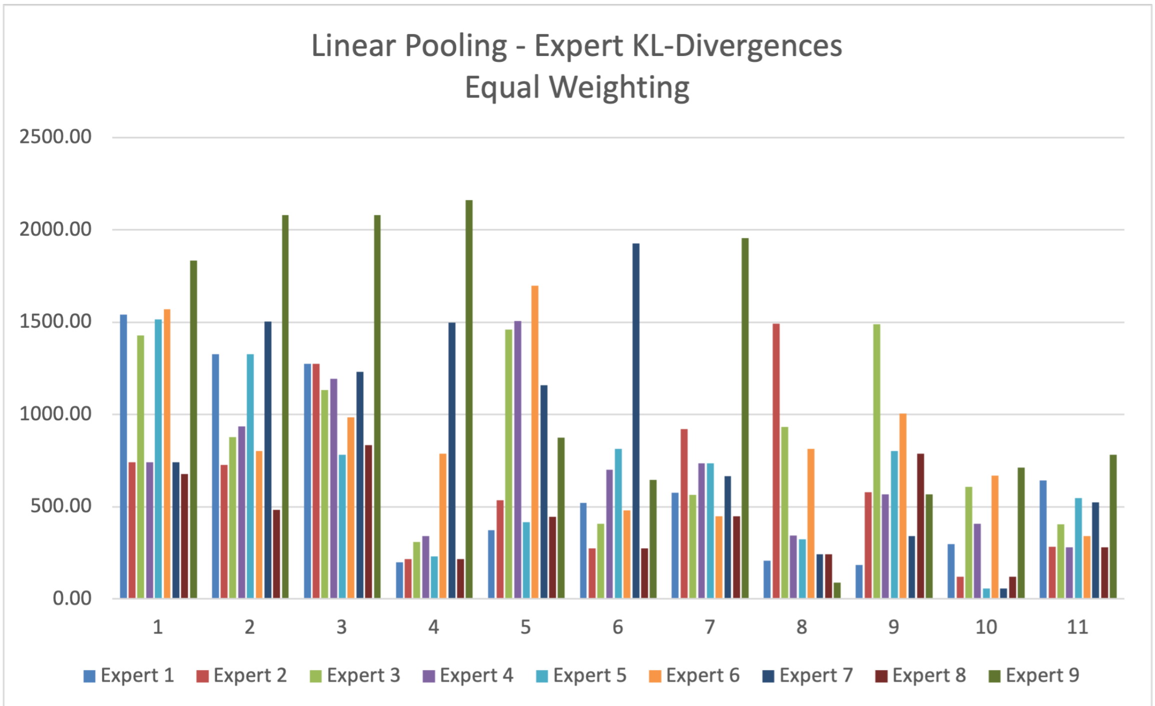

Figure 3.10: Expert KL-Divergences - Linear Pooling - Equal Weighting

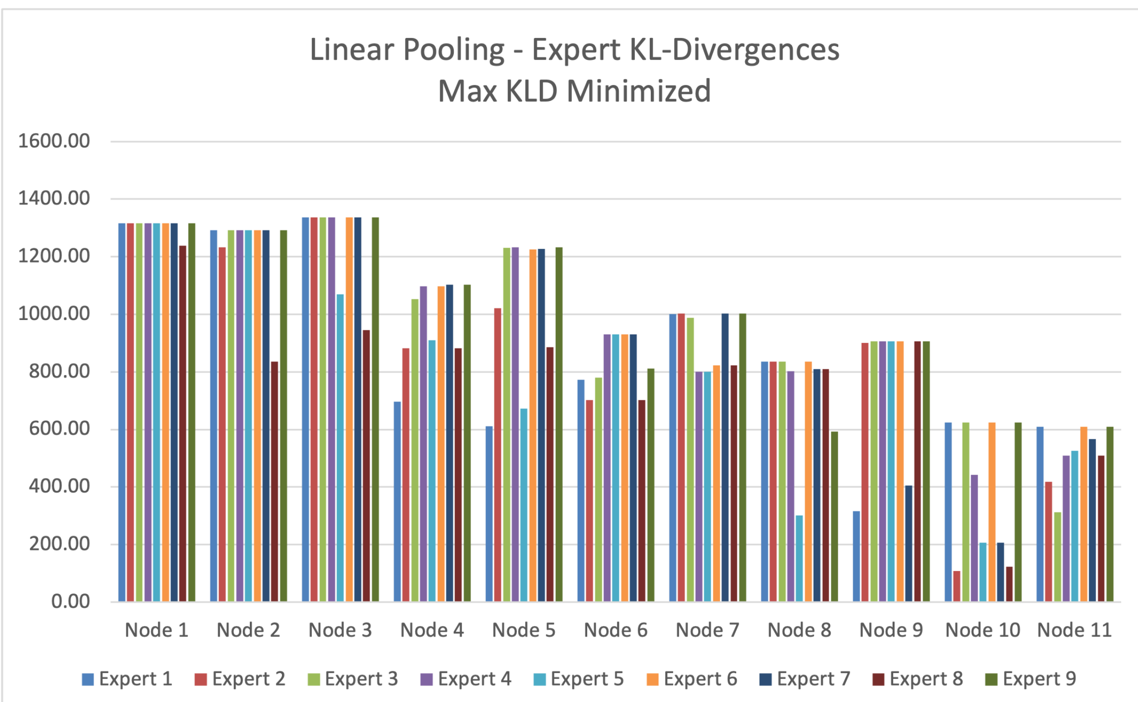

Figure 3.11: Expert KL-Divergences - Linear Pooling - Max KLD Minimization

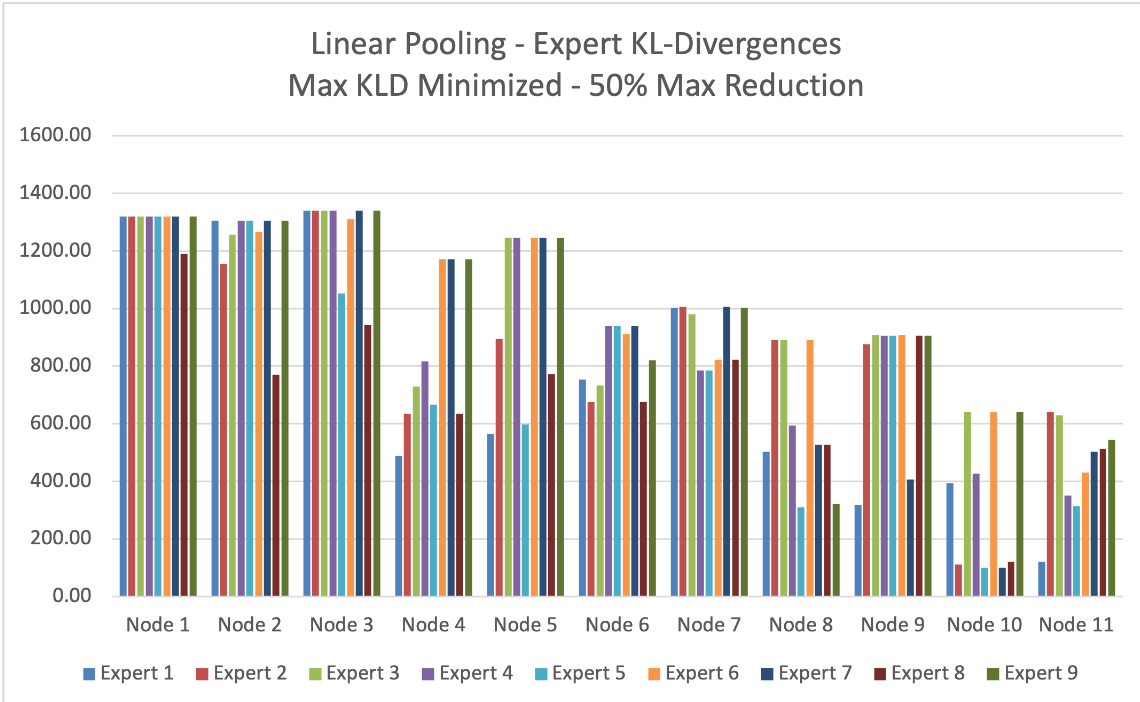

Figure 3.12: Expert KL-Divergences - Linear Pooling - Max KLD Minimization (50% Max Reduction)

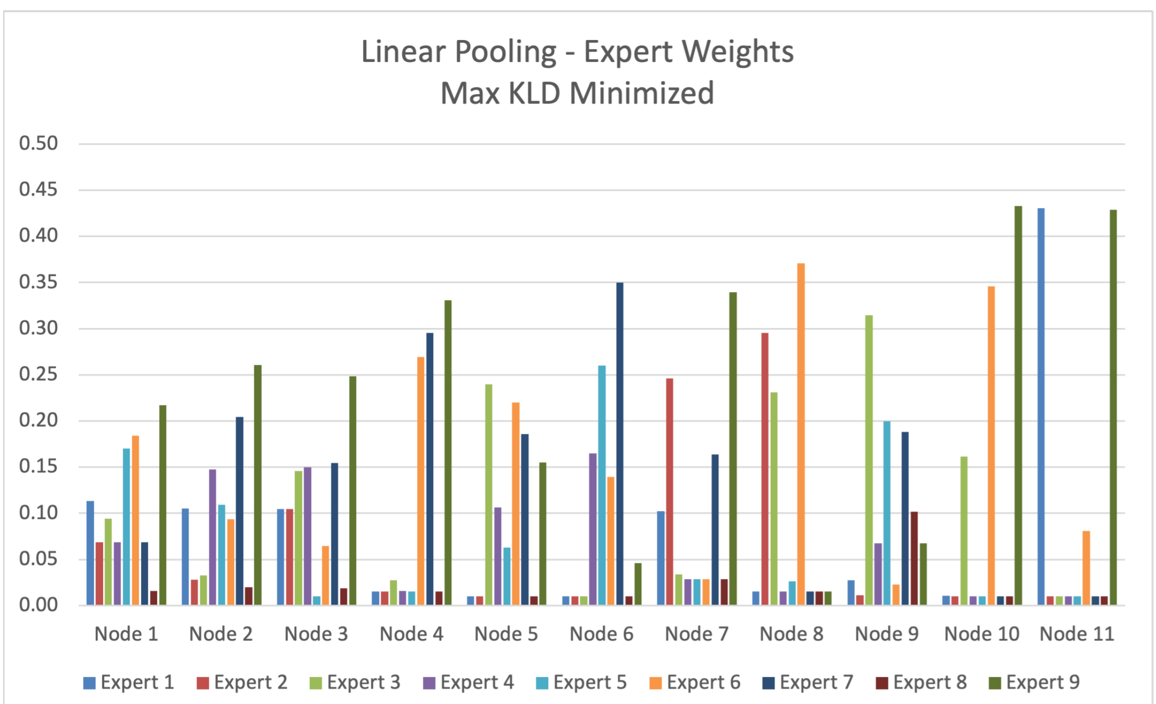

Figure 3.13: Expert Weights - Linear Pooling - Max KLD Minimization

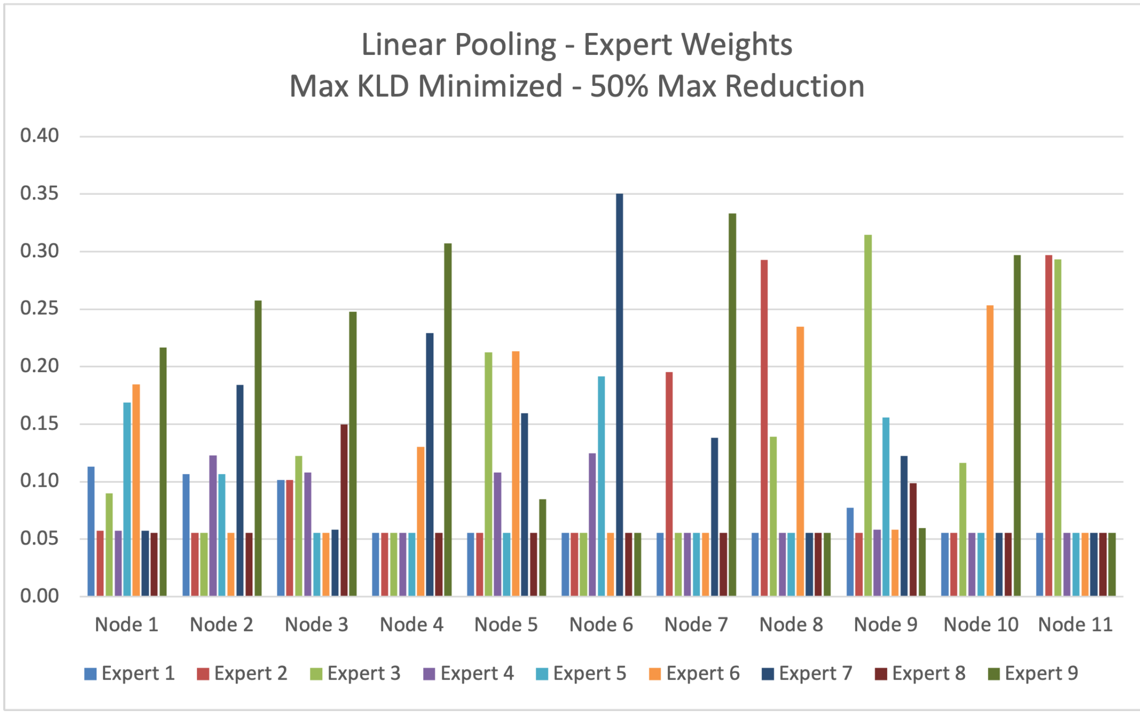

Figure 3.14: Expert Weights - Linear Pooling - Max KLD Minimization (50% Max Reduction)

Table 3.4: Linear Equally Weighted KLD Rank vs. Optimized Weight Rank Correlations

Case	Equally Weighted KLD Rank vs. Optimized Weight Rank		Equally Weighted KLD Rank vs. Optimized Limited Weight Rank	
	Spearman's Rho	Significance	Spearman's Rho	Significance
Node 01 - Linear	0.966	**<0.001**	0.966	**<0.001**
Node 02 - Linear	0.929	**<0.001**	0.812	**0.008**
Node 03 - Linear	0.745	**0.021**	0.377	0.318
Node 04 - Linear	0.929	**<0.001**	0.917	**<0.001**
Node 05 - Linear	0.800	**0.010**	0.850	**0.004**
Node 06 - Linear	0.921	**<0.001**	0.494	0.177
Node 07 - Linear	0.403	0.282	0.336	0.376
Node 08 - Linear	0.879	**0.002**	0.950	**<0.001**
Node 09 - Linear	0.293	0.444	0.285	0.458
Node 10 - Linear	0.895	**0.001**	0.979	**<0.001**
Node 11 - Linear	0.782	**0.013**	0.377	0.318

3.4.4 Comparison of Methods

When comparing the annual probability of failure (APF) results from the Monte Carlo simulation, log-linear pooling resulted in much lower APF values than any of the linear pooling results for all weighting schemes considered. (See Figure 3.7.) The APF ranges for all log-linear weighting methods were quite small and only varied within the same orders of magnitude as the equally weighted log-linear pooling results. This is due to the flat shape of the log-linear cumulative distribution function curve for each node. This flat shape provides little variability across a large number of Monte Carlo samples and the tendency of the pooling function for this data to result in near-zero values of likelihood kept the APF consistent across the sampling range. This confirms that the undesirable behavior observed at the node level also has undesirable effects on the APF of the entire tree and further suggests that log-linear pooling, no matter the weighting scheme, is likely unsuitable for input data with significant non-overlapping areas of estimated likelihood.

The Monte Carlo results for the linear pooling functions showed that both sets of the optimized results were nearly identical across the sampling range in spite of the fact that the weight reductions were arbitrarily limited to 50% of the equal weights. This result indicates that a decision maker can successfully balance the desire for equal weight and equal impact in the aggregated probability functions. In this instance, the results for both the optimized linearly pooled cases provided APF results that showed a higher maximum likelihood of failure than the equally weighted case. This was presumed to be due to the highest optimized unlimited weights being applied to relative outliers (i.e. those estimates that resulted in a high initial KLD) that represented a higher or more conservative estimated likelihood value; however, the variability in the data made confirming this presumption non-trivial. In order to confirm that this increased conservatism was due to the more conservative nature of outlier data, the entire calculation process for the linear pooling method was repeated using the inverse of the initial expert estimates. In this scenario, the expert estimates,

p, were transformed to $1 - p$ and the best and worst case values were swapped. This inversion of the estimates, as expected, resulted in APF values that were shifted downward indicating an increase in influence for less conservative outliers.

3.4.5 Expert Estimate Outlier Analysis

For the data set analyzed, there were two 95+ KL-divergence percentile outlier estimates than were counter to the expert's natural tendency. (See Table 3.5) When viewing the plots of the individual expert PDFs, it is not immediately evident that these estimates are particularly unusual. The Node 9 outlier distribution describes a curve with a high peak, indicating a relatively high level of certainty, but it fully exists within the group's range of likelihood values. The Node 11 outlier distribution also exists fully within the group's range of likelihood values, but with a peak only slightly larger then the other distributions. (See Figures 3.15 and 3.16 respectively) The subtleties of these visualizations may cause a facilitator or decision maker to miss the significance of such an estimate and fail to consider it as seriously they might if they were presented with the mathematical results.

This sort of analysis could be completed in real-time during the risk estimation process. The computations involved are not complex or computationally expensive and could easily be coded into the same tool that would calculate the KL-divergences and the adjusted expert weights.

Table 3.5: Individual Expert Atypical Outliers

Node	Expert	KLD Pct	Conservatism	Expert Conservatism Pct
Node 9	Expert 3	97.98	Conservative	3.09
Node 11	Expert 9	96.49	Unconservative	89.02

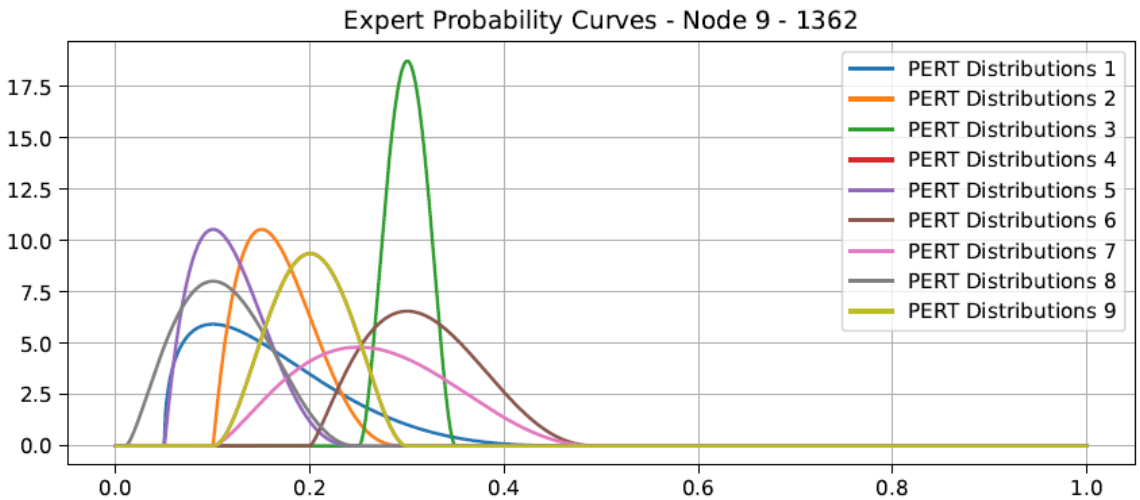

Figure 3.15: Node 9 Individual Expert PDFs

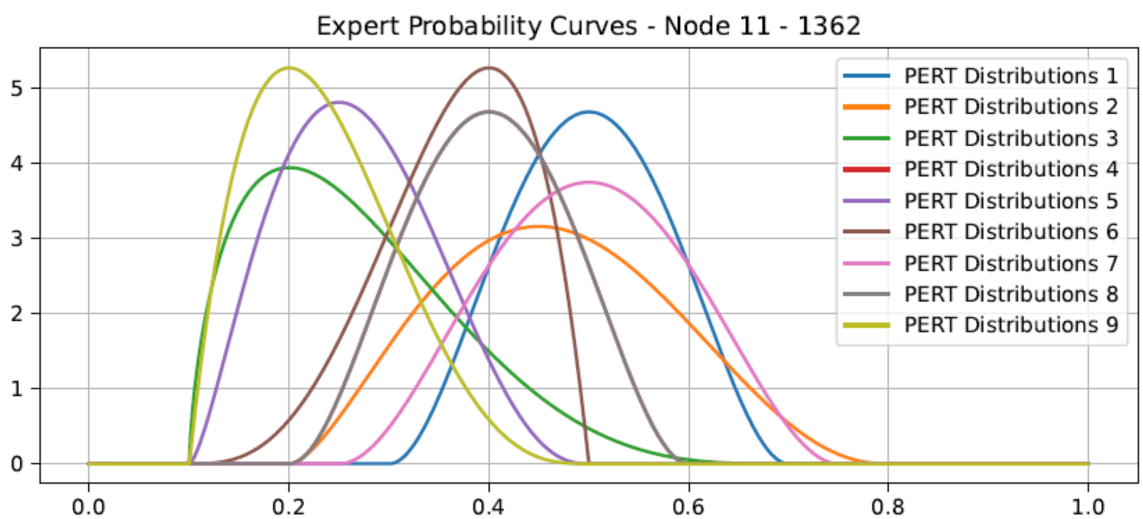

Figure 3.16: Node 11 Individual Expert PDFs

3.5 Conclusion and Recommendations for Future Work

Aggregation of subjective expert risk estimates using linear pooling with optimization of weights such that the maximum KL divergence is minimized, while limiting the reduction of weights to some subjective practical limit, is a valid method for those decision makers wishing to ensure that their aggregated risk estimates are as representative as reasonably possible with respect to each individual expert. Non-optimized linear pooling, as is currently common practice, is a reasonable method but can be improved using the optimization techniques. Unlimited optimization of weights provides improved divergence metrics, but the reduction of some experts' weights to zero or nearly zero is counter to the desire for representativeness, regardless of the KL-divergence improvements. Log-linear pooling is a poor fit for subjective estimate distributions with large areas of non-overlapping estimates. As this is likely a common occurrence in dam safety risk assessments, this method is undesirable, whether optimized or not. Unweighted initial averaging, while fast and simple, results in significant data loss and provides the appearance of an artificially low level of uncertainty among the experts.

Future work in this area should focus on the practical limit for the lower bounds for weights when using the optimization method. The maximum 50% reduction was arbitrarily chosen and it is possible that an alternate limit could yield satisfactory KLD maximum and range reductions while further limiting the weight reductions. Additionally, the α value in the Hölder pooling function could be varied to investigate hybrid aggregation methods that optimized for both α and weights making better use of the generalized form of the function.

Ultimately, risk informed decision makers must understand the complexities of opinion pooling. As there is no prescriptive method for aggregating expert opinion, decision makers may be relying on methods that are counter to their expectations for fairness and representativeness. These topics should be carefully considered and axioms for each industry, organization, or project should be determined prior to risk estimation and an aggregation method that suits the decision makers' needs should be applied to the data to ensure its proper use.

Chapter 4

A Discussion of Practical Implementation in Industry, Related Applications, and Recommendations for Future Work

4.1 A Discussion of Practical Implementation in Industry

Any change in an established industry, particularly a small one with a limited number of highly-qualified professionals, will prove challenging. Foremost among those challenges will likely be convincing the small pool of decision makers and facilitators that these changes are necessary, valuable, and worth the effort to implement. Making decisions regarding dam safety always means making decisions that affect the life safety of the public, financial impacts to the public, private businesses, government interests, and, perhaps most significantly to those intimately involved, their own reputations and the reputation of the implementing organization of which they are a part.

As such, it is expected that decision makers must be convinced of an increase in efficacy of the process as a result of the proposed changes. The results of Chapter 2 may not prove too challenging as they would only impact expert and facilitator training and would ultimately not impact the quantification of the results. However, for the pooling methods, implementation will likely prove difficult because the uncertainty levels in this sort of analysis are already so large and the optimization methods based on the KLD as proposed will likely always increase, rather than reduce, them. From a philosophical standpoint, this squares with the ability of the method to more "fairly" represent the set of diverse expert opinions, but also could lead to the perception of the decision makers that the experts are misaligned with regards to the pooled estimate. This increase in uncertainty in the aggregate could cause decision makers to believe this is a step in the wrong direction, even if they wish to have their expert team represented more uniformly in the output.

Additionally, the KLD is not a well-understood technique in industry. When discussing this work with industry professionals, they quickly understood the concept of the KLD when it was explained to them, but a strong working knowledge of the measure, its appropriate use, and the calculation of its value does not broadly exist. Some education, through publications, presentations, and professional conferences will be required to broaden the field's understanding.

The defensibility of the method in the event of an uncontrolled loss of the water retaining barrier or some other undesirable outcome at a facility will also likely be a concern of decision makers. Should some failure occur, any affected parties will surely demand detailed answers about how the safety of the dam was determined. The subjective nature of the estimation process would surely be attacked in such a scenario. It is unlikely that the public or politicians would understand the finer points of opinion pooling. Even the quite simple pooling methods that are currently employed may seem unnecessarily cryptic or complex to an aggrieved party that does not deal with mathematics of this sort regularly. The proposed methods have been discussed with industry professionals with significant levels of responsibility for the analysis of the safety of federally regulated dams and even they are not universally familiar with the specifics of the current methods. Opinion pooling seems to be a bit of a "black box" to them into which expert opinions are fed and out of which a f-N chart is generated. A decision maker may see adding another layer of rather complex mathematics in the form of optimization as further muddying these waters and making the defense of the process after some failure even more difficult.

This is not the case only for the public or only in a post-failure scenario. The decision makers for federally regulated dams are employed by the same agencies that own, operate, maintain, and upgrade these dams. As such, these quantitative risk assessments will be used to defend requests for funding and prioritization of projects and resources. The proper management of a fleet of dams, each with their own problems and risk profiles, requires the confidence of executive leadership to appropriately grant the funding necessary to address any dam safety issues identified as being of an unacceptably high risk during the risk quantification process. If the process is perceived as so opaque that it cannot be reasonably presented and defended, it is possible that the decision makers will lose the confidence of their executive leadership and the ability to secure funding for repairs or upgrades could be impacted. Worse yet, the complexity of the analysis could lead to a perception that the team responsible for the analysis is using unnecessarily complex tools to "game" the system to prioritize their facility.

While these potential problems are serious, they are not insurmountable. Any sufficiently advanced technical field is rife with complex issues and analyses that are not readily understood by the executive leadership of the practitioners or interested members of the public. Such is the case in the nuclear power industry. Likewise with the design of large bridges or commercial aircraft development and many other highly-technical industries upon which the public relies for its safety.

In interviews conducted for this work, professionals working in this space expressed the importance of narrative in the process. This is perhaps the best method for addressing all of these potential issues. Narrative is powerful, particularly in the absence of technical or mathematical experience. If a story can be crafted, even as far down as the event tree node-level, the mathematics of the process become far less important to those receiving the message. If the results can be accurately distilled in the risk assessment report in a way that captures the opinions of the experts and explains the reasons for their uncertainties and disagreements on the quantification of the risks, this may prove to be an even more powerful tool working in the service of the decision maker.

The development of the expert estimate outlier analysis as described in Section 3.4.5 to quantify the amount and direction of dissent for each expert, both in general and at the individual node level, can provide insight into those areas or events about which the experts largely agree and can highlight those areas or events where there exists more pronounced disunity. As currently practiced, this is all but impossible. Perhaps an experienced facilitator could make note of particularly contentious discussion between some experts during the behavioral aggregation process and could subjectively consider the expert likelihood distributions and provide some insight, but this still would not account for an individual expert's broader tendency to provide higher or lower estimates than the group. This means that the magnitude of this disagreement would be difficult to understand. The method described will be able to provide important context for relatively high levels of uncertainty in the pooled distributions and can point to particular nodes where a decision maker may choose to proceed with further investigation, allocate more resources, seek alternate opinions, data mine for case studies of similar situations, etc. to determine if dissent at some node is indicative of an important outlying opinion that should be given credence. It should be noted that such a tool could be employed either with or without the KLD-based optimization methods if those prove unpalatable. This approach will likely be more interesting to decision makers as it may not require changes to their current aggregation approach.

Of course, the above only describes the obstacles to implementation in human terms. In more practical terms, the most obvious hurdle is software. The software used for opinion pooling and Monte Carlo simulation ranges from rather simple Excel spreadsheets to bespoke packages specifically tailored to the dam safety industry. Each agency and consultant has their own preferences and standard processes. If implemented in practice, the methods presented here would require significant changes to the software currently in use. Excel would be a poor choice for either of the optimization methods. It is likely a poor choice for any more advanced techniques as linear, logarithmic, or Hölder pooling at anything like a reasonable resolution would be clumsy at best and poorly suited to the program. Any off-the-shelf packages almost certainly would not support the optimization methods without revision by the developer that would likely not be worth the investment for a niche industry. The bespoke packages could be revised to provide this functionality or standalone programs, similar to those built in Python for the purposes of this research could be developed. The methods are not computationally expensive for the relatively small datasets generated in these types of risk estimations.

It is likely that further vetting of these methods would be required prior to investing in full software implementation. A version of the Python code developed for this work could be easily adapted to "re-run" a larger set of historic risk estimates. This would allow decision makers to consider the outputs of real risk assessments with which they are already familiar and compare the results using their current approach and the proposed methods and determine if they wish to employ any or all of the new methods. This would also provide the opportunity to assess the viability of expert estimate outlier analysis to provide some context for dissent amongst the expert team. As they would presumably be aware of the people and personalities involved, this would give the decision makers, facilitators, and other involved staff the opportunity to understand how this information could have allowed them to better craft their narrative when compiling the risk assessment report. The agencies and consultants could then choose the portions of the work they would like to implement and modify or develop their own software accordingly.

The emergence of large language models may play an important role in future research in this area, particularly with regards to generating a narrative from the data. This is beyond the scope of this work at this time, but implementation of a large language model to craft a portion of the risk assessment report based on the data output by the optimization methods and the expert comparison tool could prove quite useful in distilling the complexities of the method in a way that provides value to the decision maker.

4.2 Applicability to Other Industries

The work presented, in Chapter 2 and Chapter 3, could be applied to any industry where subjective expert estimation is used. While the Chapter 2 specifically polled dam safety industry professionals to determine if conditions were present indicating the presence of motivational effects, it is reasonable to assume that these biases exist in other instances as well. As they are inherent to our own human nature, they are likely to present in most collaborative efforts, particularly those where high risk conditions or behaviors are discussed. The existing literature addresses these effects more broadly. This work is likely applicable to any decision making or estimating process that relies on subjective interpretation of data.

For instance, university admissions committees must choose between many similarly qualified candidates. High grades and standardized test scores are assumed to be a given at highly selective institutions with the differentiating factors being extracurricular activities, life experience, the impression made by an essay or letter provided with the application, or a recommendation from a teacher, professional, or alumnus. It is not clear how a committee should factor these intangibles into their decision making process. Subjective valuations must be made and reconciled among the committee members. There is no obvious opportunity for motivational bias to impact a committee members decision barring some personal connection to the applicant, but affect-influenced, myside, or other types of emotional biases could arise as a side effect of the group dynamic and unfairly impact their decisions.

In times of crisis or emergency, many of the same conditions are present as exist inherently in dams. There is a high degree of “unknowability” about the situation, the consequences of inaction or a poor decision may be quite high, and emotions are almost certain to impact the decision makers. It could be useful to have a facilitator trained in bias identification and control as are used in dam safety to quickly discuss biases with the emergency management team, if time allows, and help guide the conversations around critical decisions to reduce the use of emotional information as cognitive information. In this case, the negative emotion of being associated with a poor outcome would likely be difficult to overcome even with guidance, but having a neutral voice in the room could provide significant value.

The medical field regularly uses committees to review cases prior to significant interventions such as cancer treatments, complex surgeries, etc. This is another case where the uncertainty associated with the intervention and its efficacy lead to the use of subjective decision making. This is also an area with the potential for direct motivational bias. For instance, a surgeon on the committee may have a strong desire to perform an operation for the purposes of increasing their own knowledge or prestige while an oncologist on the committee may wish to pursue a non-surgical treatment for similar reasons. In these cases, a neutral facilitator could help guide the conversation productively and attempt to dampen the interpersonal conflicts that could lead to adversarial effects or myside bias as described in Chapter 2.

In terms of applicability to other industries, Chapter 3 applies less broadly. In most complex technical systems where a sophisticated process is used to determine likelihood of failure or some other negative outcome, there exists a large quantity of objective data from which to draw as input. Take, for instance, a nuclear power plant. Nuclear plants use a very similar event tree structure to that used in dam safety. These sophisticated models are populated with likelihood probabilities at the node level that are backed by a large dataset of subcomponent failure rates. One can imagine the volume of pumps, motors, valves, breakers, switches, sensors, etc. present in a nuclear plant. This is a much more complex system than a dam, but there is a high degree of certainty regarding the lifespan of each subcomponent. Not only are these subcomponents in wide use in the nuclear power industry, but in other industries as well. They are typically not a single point of failure, so failures can tolerated without overall adverse effects due to high levels of redundancy. This means that subcomponent failures can be studied and tracked over long periods of time and very reliable ranges of failure likelihoods can be used to populate the event trees. There is very little subjectivity required for such a model. As such, the models can be updated in real time to reflect the condition and availability of all critical components listed in the event tree and an overall likelihood of failure can be recalculated multiple times a day without any subjective influence. A similar high volume of objective data is also available when analyzing commercial aircraft, electrical power grids, chemical manufacturing plants, and many other systems that, if allowed to fail, would have a catastrophic effect on the public.

The ideal application for the techniques of Chapter 3 is a complex system about which there are high levels of epistemic and aleatory uncertainty and the consequences of its failure to the public, the economy, or the reputation of the owning or implementing organization are quite high. In this way, dam safety risk assessments are something of an extreme case. Dam failures are exceedingly rare and their condition is very difficult to assess. Even taking steps to investigate a dam's condition can cause hazards to its stability. For instance, invasive methods such as drilling or coring can create preferential pathways for erosion that can create a more serious problem than the drilling would have investigated in the first place.

Other applications for the subjective opinion pooling methods in Chapter 3 do exist, such as the security of infrastructure. The risk of attacks on infrastructure such as power transmission substations is a reasonably similar example. The consequences of failure are typically lower than a dam failure, but can be quite high in financial and reputational cost. The nature of such an attack is exceedingly difficult to predict without specific intelligence and therefore protecting against it is not a straightforward exercise. Small arms fire is different from an improvised explosive device, which is different still from a cyberattack. There are also many more such facilities than dams, so prioritization of upgrading or hardening of this kind of infrastructure is also possibly more complex than managing risk of failure in a fleet of dams. These factors lend themselves to subjective risk estimation. It is assumed that the methods for opinion pooling, optimized weighting, and expert estimate outlier analysis could be reasonably adapted to this process.

It is likely that the planning of highly sensitive military or related operations are similarly constrained by high levels of uncertainty, high consequences, and operation specific conditions that make past experience potentially unreliable. Expert judgement and decision making under uncertainty must play a role is such planning scenarios. As with the infrastructure example above, the methods of Chapter 3 could be adapted to any such group subjective risk estimation process.

4.3 Recommendations for Future Work and Conclusion

The work presented in Chapter 2 and Chapter 3 presents meaningful results, but the work of Chapter 3 was limited by relatively small sample size. If one of the federal agencies with responsibility for dam safety were to provide a much larger set of data for many facilities and many expert panels, the general tendencies of the methods could be more reliably demonstrated. There is some mathematical justification for the general behaviors of the techniques, but as this work was intended for practical implementation in industry, verification over the largest reasonable set of real-world data would be preferable.

Further, the ability of the Hölder pooling function to vary its nature between logarithmic and linear pooling, and even beyond for values of $\alpha < 0$ and $\alpha > 1$

respectively, offers opportunities for investigation of alternate values of α that may provide valuable pooling techniques, either equally weighted or weighted by the optimization techniques presented in Chapter 3.

The optimization methods could also benefit from a larger data set to hone the arbitrary constraint limitations placed on the solver to prevent excessively dominant solutions. For this work, an arbitrary constraint of a maximum weight reduction of 0.5 was selected and provided a reasonable middle ground between equal weights and a dominant weight. More work here to find a reduction value that consistently increased representativeness while still providing a palatable minimum weight value for decision makers should be pursued.

Finally, there is the possibility that some method may exist for reliably generating a probability distribution unrelated to the expert distributions except by way of the comparison of Kullback-Leibler divergence (KLD) values between the individual experts and the new distribution. Perhaps a neural network or some similar approach could develop a distribution to minimize some loss function related to the KLD that would not rely on the traditional pooling methods and would not require weighting. It is not obvious that this approach would produce non-trivial solutions, but it is an approach that may interest future researchers in this domain.

The techniques explored in this work offer both intriguing opportunities and notable challenges in improving expert opinion aggregation and decision making processes in uncertain environments. Despite these hurdles, there is considerable potential for these methods to enhance decision making in areas characterized by high uncertainty and limited objective data. Beyond their immediate application, the concepts presented may also hold relevance in broader contexts where subjective judgments play a central role, particularly when expert consensus is difficult to achieve. Combining technical rigor with techniques to improve unbiased thinking are key to realizing the benefits these approaches can offer.

Bibliography

Abbas, A. E. (2009). A kullback-leibler view of linear and log-linear pools. *Decision analysis*, 6(1):25–37. 5, 44

Albarracín, D. and Tarcan Kumkale, G. (2003). Affect as information in persuasion: A model of affect identification and discounting. *Journal of personality and social psychology*, 84(3):453–469. 15

Alhakami, A. S. and Slovic, P. (1994). A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk analysis*, 14(6):1085–1096. 16

Armor, D. A. (1999). *The illusion of objectivity: A bias in the perception of freedom from bias*. Thesis, University of California, Los Angeles. 2, 14, 18

Bacharach, M. (1979). Normal bayesian dialogues. *Journal of the American Statistical Association*, 74(368):837–846. 40

Bakir, N. O. (2008). A decision tree model for evaluating countermeasures to secure cargo at united states southwestern ports of entry. *Decision analysis*, 5(4):230–248. 37

Baybutt, P. (2018). The validity of engineering judgment and expert opinion in hazard and risk analysis: The influence of cognitive biases. *Process safety progress*, 37(2):205–210. 15

Bowles, D. S., Anderson, L. R., and Glover, T. F. (1997). A role for risk assessment in dam safety management. In *Proceedings of the 3rd International Conference HYDROPOWER*, volume 97. 8

Bowles, D. S., Anderson, L. R., and Glover, T. F. (1998). The practice of dam safety risk assessment and management: Its roots, its branches, and its fruit. In *its Roots, its Branches, and its Fruit. Proceedings of USCOLD 1998 Annual Lecture*. Citeseer. 2, 9, 14, 34

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583. 38

Clemen, R. T. and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk analysis*, 19(2):187–203. 42, 43

Cooke, R. (1991). *Experts in uncertainty opinion and subjective probability in science*. Environmental ethics and science policy. Oxford University Press, New York. Includes bibliographical references and index. 34, 42, 43

Cooke, R. M. and Goossens, L. L. H. J. (2008). Tu delft expert judgment data base. *Reliability Engineering & System Safety*, 93(5):657–674. 34, 43

de Carvalho, L. M., Villela, D. A. M., Coelho, F. C., and Leonardo Soares, B. (2020). Combining probability distributions: Extending the logarithmic pooling approach. *arXiv.org*. 41, 43

DeNeale, S., Baecher, G., Stewart, K., Smith, E., and Watson, D. (2019). Current state-of-practice in dam safety risk assessment. Technical report, Oak Ridge National Laboratory. 34

Denes-Raj, V. and Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of personality and social psychology*, 66(5):819–829. 15

Dewispelare, A. R., Herren, L. T., and Clemen, R. T. (1995). The use of probability elicitation in the high-level nuclear waste regulation program. *International journal of forecasting*, 11(1):5–24. 34

Dror, I. E. (2017). Human expert performance in forensic decision making: Seven different sources of bias. *Australian journal of forensic sciences*, 49(5):541–547. TAJFS.jpg. 17

Dror, I. E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical chemistry (Washington)*, 92(12):7998–8004. 17, 18

Efendic, E., Chanashekar, S. P., Lee, C. S., Yeung, L. Y., Kim, M. J., Lee, C. Y., and Feldman, G. (2021). Risky therefore not beneficial: Replication and extension of finucane et al.'s (2000) affect heuristic experiment. *Social psychological & personality science*, page 194855062110567. 15

FEMA (2004). Federal guidelines for dam safety (p-93). 2004368612 prepared by the Interagency Committee on Dam Safety. 28 cm. "FEMA 93, April 2004.". 7, 33, 37

FEMA (2015). Federal guidelines for dam safety risk management (p-1025). Catalog No. 14353-1. 9, 37

FERC (2016). Risk-informed decision making guidelines. 2, 8, 12, 14, 37

Finucane, M. L., Alhakami, A., Slovic, P., and Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of behavioral decision making*, 13(1):1–17. US National Science Foundation - No. SBR 9422754; No. SBR 9709307. 15, 16

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., and Combs, B. (1978). How safe is safe enough? a psychometric study of attitudes towards technological risks and benefits. *Policy sciences*, 9(2):127–152. 15

Frangopol, D. M. and Tsompanakis, Y. (2014). *Maintenance and safety of aging infrastructure*. Structures and infrastructures series, volume 10. CRC Press, Boca Raton ;. Includes bibliographical references and index. 2, 11, 14

Garg, A., Jayram, T. S., Vaithyanathan, S., and Zhu, H. (2004). Generalized opinion pooling. *Annals of Mathematics and Artificial Intelligence*. 44

Genest, C. (1984a). A characterization theorem for externally bayesian groups. *The Annals of statistics*, 12(3):1100–1105. 40

Genest, C. (1984b). Pooling operators with the marginalization property. *Canadian journal of statistics*, 12(2):153–163. ark:/67375/WNG-3C2H96C9-T. 42

Genest, C. and McConway, K. J. (1990). Allocating the weights in the linear opinion pool. *Journal of forecasting*, 9(1):53–73. ark:/67375/WNG-HRCWG9W4-F. 43

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical science*, 1(1):114–135. 38

Gustafson, D. H., Shukla, R. K., Delbecq, A., and Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, delphi groups, and nominal groups. *Organizational behavior and human performance*, 9(2):280–291. 38

Hammitt, J. K. and Zhang, Y. (2013). Combining experts' judgments: Comparison of algorithmic methods using synthetic data. *Risk analysis*, 33(1):109–120. istex:377E42DF76A513C2B5EE0B441E90833E5B385C8A. 38

Kelman, H. (2008). *A social-psychological approach to conflict analysis and resolution*, pages 170–183. Routledge [Taylor & Francis], London and New York. 19

Kennedy, K. A. and Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality & social psychology bulletin*, 34(6):833–848. 18

Klaczynski, P. A. (1997). Bias in adolescents' everyday reasoning and its relationship with intellectual ability, personal theories, and self-serving motivation. *Developmental psychology*, 33(2):273–283. 17

Klaczynski, P. A. and Gordon, D. H. (1996). Everyday statistical reasoning during adolescence and young adulthood: Motivational, general ability, and developmental influences. *Child development*, 67(6):2873–2891. 17

Klaczynski, P. A. and Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology and aging*, 15(3):400–416. 17

Koliander, G., El-Laham, Y., Djuric, P. M., and Hlawatsch, F. (2022). Fusion of probability density functions. *Proceedings of the IEEE*, 110(4):404–453. 42, 44

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3):480–498. 13, 15

Layton Jr, E. T. (1986). *The Revolt of the Engineers. Social Responsibility and the American Engineering Profession.* ERIC. 12

Montibeller, G. and von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Anal*, 35(7):1230–51. Montibeller, Gilberto von Winterfeldt, Detlof eng Risk Anal. 2015 Jul;35(7):1230-51. doi: 10.1111/risa.12360. Epub 2015 Apr 15. 15

Mosleh, A. and Apostolakis, G. (1986). The assessment of probability distributions from expert opinions with an application to seismic fragility curves. *Risk analysis*, 6(4):447–461. ark:/67375/WNG-VS9GRDJ7-J. 34

Mosleh, A., Bier, V. M., and Apostolakis, G. (1988). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety*, 20(1):63–85. 38

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., and Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological science*, 24(10):1889–1897. 18

Nguyen, H. (2019). Subjective risk estimation of the rare event. *Journal of KONES, internal combustion engines*, 26(1):103–110. 37

O'Hagan, A. (2006). *Uncertain judgements : eliciting experts' probabilities*. Statistics in practice. Wiley, London ;. Includes bibliographical references (p. 267-306) and indexes. 34, 39, 40, 43

Otway, H. and von Winterfeldt, D. (1992). Expert judgment in risk analysis and management: Process, context, and pitfalls. *Risk analysis*, 12(1):83–93. 2, 12, 14

Ouchi, F. (2004). Literature review on microfinance: Development research group. Policy Research Working Paper 3201, World Bank. 18, 38

Parkhurst, D. F. (1998). Arithmetic versus geometric means for environmental concentration data. *Environmental science & technology*, 32(3):92A–98A. 41

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of educational psychology*, 77(5):562–571. 17

Pronin, E., Lin, D. Y., and Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality & social psychology bulletin*, 28(3):369–381. 17, 18

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 72(1):71–91. http://dx.doi.org/10.1111/j.1467-9868.2009.00726.x. 40

Sezen, H., Hur, J., Smith, C., Aldemir, T., and Denning, R. (2019). A computational risk assessment approach to the integration of seismic and flooding hazards with internal hazards. *Nuclear engineering and design*, 355(2019):110341. USDOE. 37

Siegrist, M., Keller, C., Kastenholz, H., Frey, S., and Wiek, A. (2007). Laypeople's and experts' perception of nanotechnology hazards. *Risk analysis*, 27(1):59–69. 16

Simon, D., Ahn, M., Stenstrom, D. M., and Read, S. J. (2020). The adversarial mindset. *Psychology, public policy, and law*, 26(3):353–377. 18

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk analysis*, 24(2):311–322. 15

Sokolowska, J. and Sleboda, P. (2015). The inverse relation between risks and benefits: The role of affect and expertise. *Risk analysis*, 35(7):1252–1267. Polish National Center for Science - No. NN106363740. 16

Stanovich, K. E., West, R. F., and Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current directions in psychological science : a journal of the American Psychological Society*, 22(4):259–264. 17

Stone, M. (1961). The opinion pool. *The Annals of mathematical statistics*, 32(4):1339–1342. 40

Toplak, M. E. and Stanovich, K. E. (2003). Associations between myside bias on an informal reasoning task and amount of post-secondary education. *Applied cognitive psychology*, 17(7):851–860. 17

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological bulletin*, 63(6):384–399. 17

Tversky, A. and Kahneman, D. (1975). *Judgment under Uncertainty: Heuristics and Biases*, pages 141–162. Springer Netherlands, Dordrecht. 11

USACE (2014). Safety of dams - policy and procedures. 8, 11, 14, 37

USBR (2019). Best practices in dam and levee safety risk analysis. 2, 11, 12, 16, 18, 36, 37, 38, 41, 46

Vick, S. G. (1999). Considerations for estimating structural response probabilities in dam safety risk analysis. *US Bureau of Reclamation, Technical Service Center, Denver, CO.* 43

Vick, S. G. (2002). *Degrees of belief : subjective probability and engineering judgment.* ASCE Press, Reston, Va. Includes bibliographical references (p. 431-446) and index. 2, 8, 12, 13, 14

Appendices

Appendix A

Chapter 2 Supplemental Material

A.1 Survey Questions

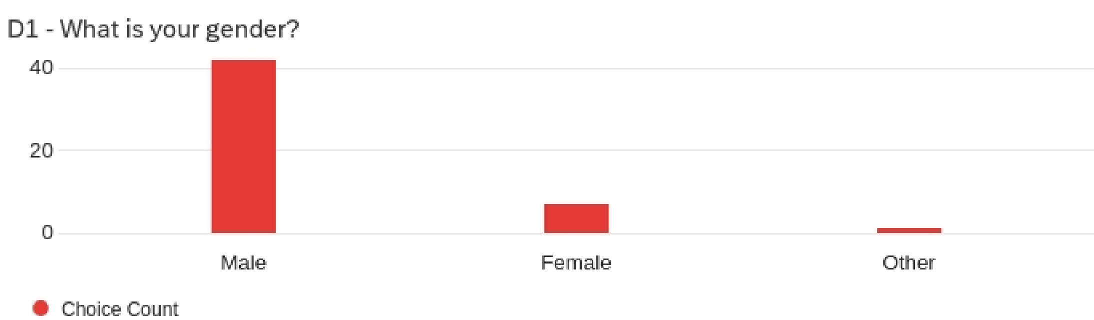

Figure A.1: Question D1

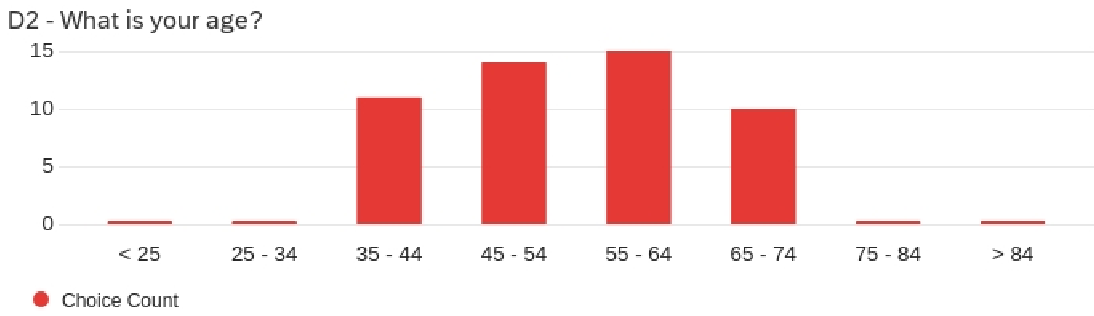

Figure A.2: Question D2

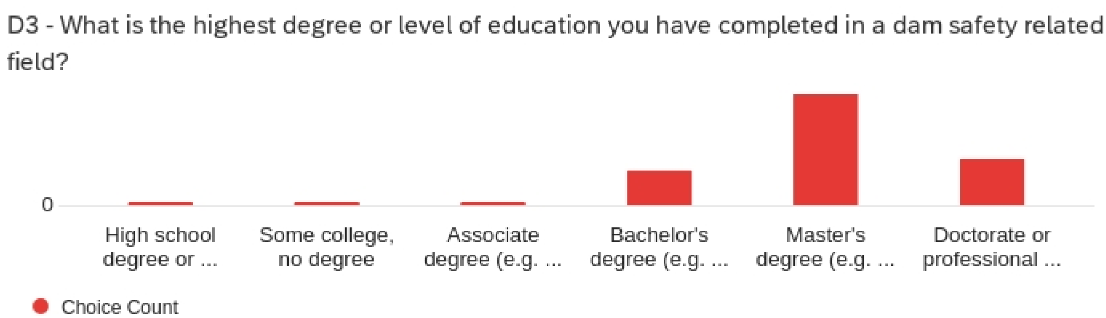

Figure A.3: Question D3

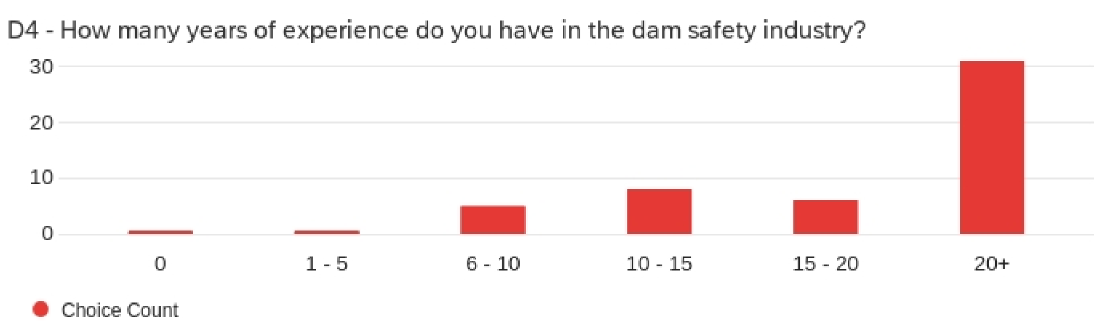

Figure A.4: Question D4

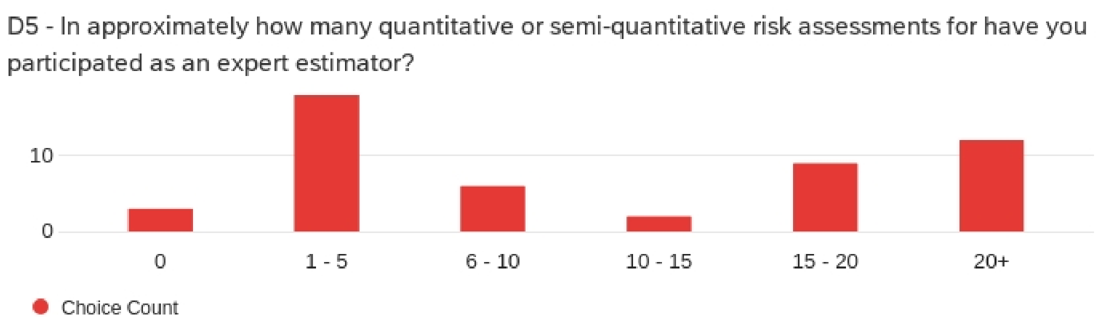

Figure A.5: Question D5

Figure A.6: Question D6

D6.1 - You responded Other/Multiple. Please briefly describe your most typical role in QRAs/SQRAs.

You responded Other/Multiple. Please briefly describe your most typical role in QRAs/SQRAs.

Program manager, operations, etc.

I have had a range of SME (earlier in my career), facilitator (mid-career) and now as a program manager for a region of dams. I no longer am a risk estimator- but that's hard to convey in the survey so far.

geotechnical SME for QRAs
both facilitator and SME for SQRAs

Client

Risk estimator, note-taker, report author, facilitator

Facilitator

I have been a facilitator, engineer of record and a SME

Risk analyst, system analyst, facilitator, risk quantification

Subject Matter Expert
Facilitator/Risk Analyst

SME most typical, have also been a facilitator on several occasions.

Regulator and oversight

As a risk cadre lead - facilitator, study technical and risk lead, HH lead, team lead, estimator.

As a Regulator, as an Owner, and as a consultant over the course of my career.

Figure A.7: Question D6.1

D7 - Who is your direct employer?

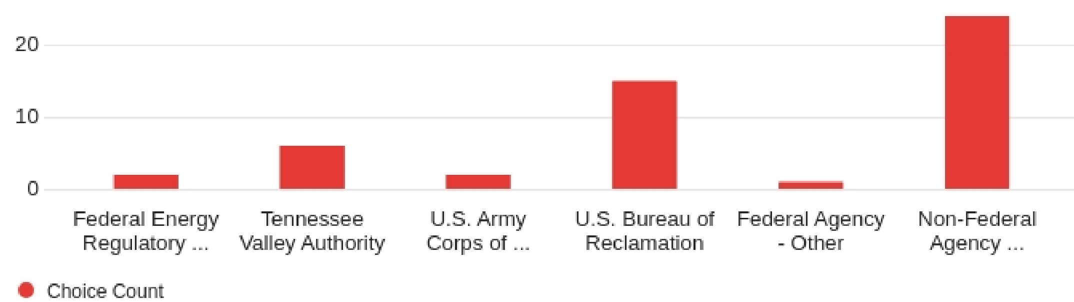

Figure A.8: Question D7

D7.1 - You selected Federal Agency - Other. Please list the agency name below.

You selected Federal Agency - Other. Please list the agency name below.

US Army Corps of Engineers for 21 years (12 as RMC risk cadre lead). Currently FERC RIDM branch senior civil engineer (May 2022).

Figure A.9: Question D7.1

Figure A.10: Question M1

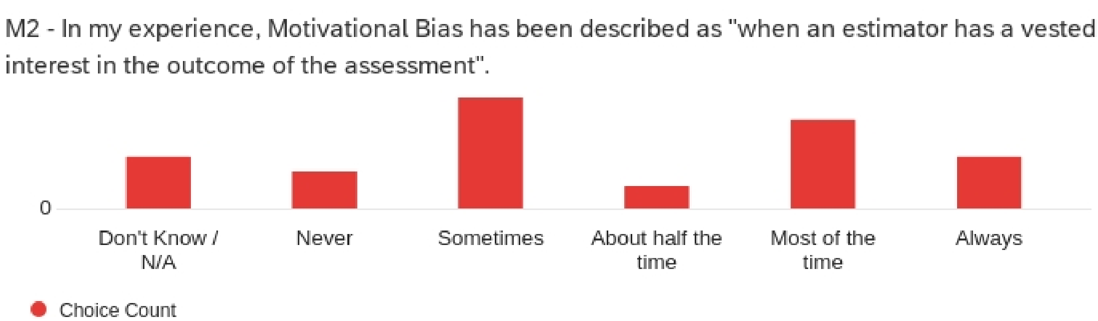

Figure A.11: Question M2

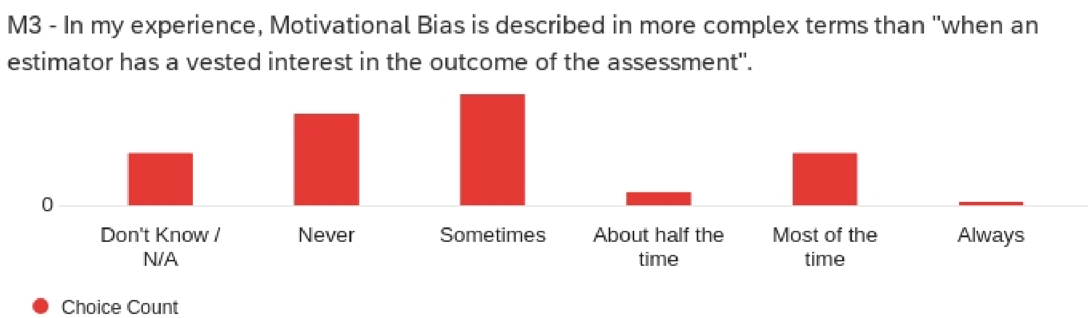

Figure A.12: Question M3

M3.1 - Please briefly describe the more complex definition(s) for Motivational Bias that you have seen presented during the risk elicitation process or associated bias training. If you do not recall them specifically, please respond "I do not recall" in the test box provided below.

Please briefly describe the more complex definition(s) for Motivational Bias that you have seen presented during the risk elicitation process or associated bias training. If you do not recall them specifically, please respond "I do not recall" in the test box provided below.

Do not recall

i do not recall

Explaining both conscious and unconscious elements of bias and explaining that a "vested interest" can be based on your preconceived idea of the outcome prior to the workshop. That is, you may be more likely to support an outcome that you have used/approved in the past, even when that may be a poor choice based on the standards of today.

There can be biases by the available data. In the case where there is an abundance of data I think there could be a bias on the part to reduce and simplify the data.

I do not recall

I do not recall

When a certain outcome is desired either knowingly or unknowingly and the results are skewed to match that outcome.

I do not recall

I do not recall

1) Picking the preferred design alternative, Say DMM versus buttress fill
2)Selecting SDF based on f/n charts, incremental risk and return period for SDF?

Typically due to historical areas of practice and personal experience.

Motivational bias takes many forms, including having a vested interest in the outcome. This bias exists in the utilization of data, models and methods preferred by the expert or the facilitator or the organization running the project or the owner.

I do not recall

Figure A.13: Question M3.1

Figure A.14: Question A1

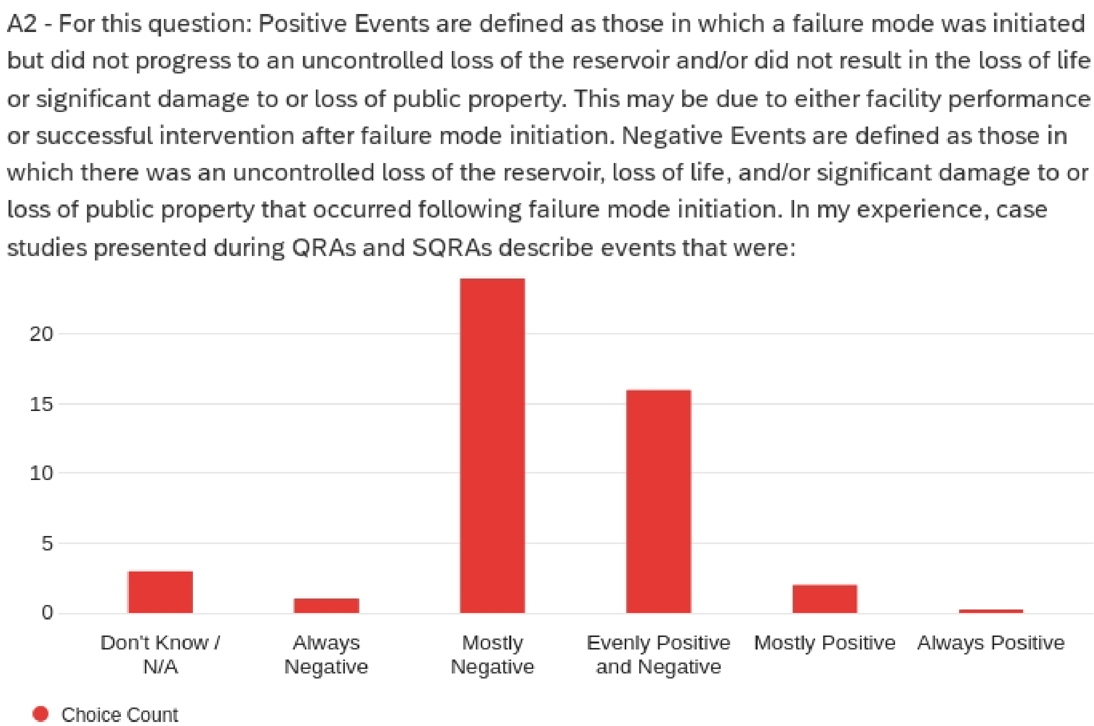

Figure A.15: Question A2

I do not recall the specific details, but many are related to having a vested interested in the risk estimation process in addition to the outcome.

I do not recall

I do not recall

I do not Recall

The "vested interest in an outcome" can be either positive (more work, more investment, more priority, more attention) or negative (the opposite of the above). It is important to recognize that motivational bias works in both directions and that governance needs to consider both types.

I do not recall except to say that the types of vested interests vary (Personal, professional, workload, etc.)

Motivation bias can occur at any point within an elicitation, regardless of the outcome, and can involve any number of cognitive biases the vested individual may use to argue his/her point. We are taught to recognize a broad number of biases within both our own arguments as well as those from others so that the most pertinent information is considered in any elicitation.

Figure A.16: Question M3.1 cont.

W1 - If you have any comments or observations regarding Motivational Bias or similar effects in the risk elicitation/estimation process, please provide them here.

If you have any comments or observations regarding Motivational Bias or similar effects in the risk elicitation/estimation process, please provide them here.

N/A

Thinking about your questions, one comment which comes to mind is the personalities of the facilitator and estimators. Changing one member of the group could result in better team cohesion/bonding where conservative/unconservative estimates are no longer thought to be needed.

From my experience, I have seen where one estimator's opinion can sway the other estimator's views (and risk estimates) which can impact the overall risk assessment outcome. This is more typical with stronger personalities vs. easy-going people that don't want to "rock the boat" or will agree just to keep the group moving. This may result in higher or lower risk estimates than the team would have developed in a different setting. Again in my view, one's personality and past experiences (ie. how they were brought up, etc) may lend to a "personality" bias that is rooted in whether the estimator is generally a conservative person or risk-taker. It's important that the team has a diverse group to avoid this from happening.

Would be good to add this influencing bias into the facilitator introductory presentation

Most engineers and scientist would reject that they bring motivational bias into their professional judgements. This makes it much harder to effectively resolve. We need to get away from how things have always been done and take a fresh look at old problems.

Motivation is biased to consensus and conservatism. Group dynamic is to avoid perception of professional incapacity. This creates environments where professional / experience based judgements are discouraged.

From what I have observed, designers in government agencies are good at objectively looking at a design (even if it is theirs) to critically evaluate it within a risk analysis.

Motivational bias is more common in joint risk analyses with other agencies or A/E firms; however, it is not often a driver in risk estimates and conflicts are most often difference of technical opinion and rarely escalate.

None

Figure A.17: Question W1

You should be aware that in Reclamation's expert elicitation model, the facilitator plays a strong and active role in the meeting. Our facilitators are not there just as referees, but are personally responsible for the quality of the risk estimates. While the facilitator could in theory have their own motivational bias, their job is to select the consensus range that makes the most sense in their experience, not the one that makes the most people happy. Also, estimators are expected to be able to explain their position rather than just throw out a number and dig in. From this perspective, it wouldn't really matter if somebody attempted to "counter" another person's estimate by going to the opposite extreme because the facilitator would likely intervene, and could decide to not use their estimate in the consensus range

I do think motivational bias exists and I've seen owners and regulators stake out opposite sides of an issue. However, my experience with QRA and SQRA is very limited. Most of my observations relate to PFMA workshops.

I have seen bias in terms of being overly conservative to not wanting to take on risk as they are either inexperienced, wanting there way, defensive, or have preconceived notion or objective.

I have not found these issues to have a major influence on risk assessment outcomes.

I have not found motivational bias to be a problem with the well qualified independent professionals I generally encounter in the risk assessments. I have seen occasional examples when owner or Engineer of Record representatives perform the role of elicitors. In many cases, this can be avoided by having owner or Engineer of Record representatives as full-time advisors to the team but not as elicitors.

I also often see anchoring bias impact an elicitation outcome, where people anchor to the first round of elicitation or first piece of key evidence presented. Another impact during an elicitation is sometimes the loudest voice in the room can dominate an outcome or someone in a position of authority over other SMEs can bias an elicitation outcome.

I believe the dam safety risk assessment community would be well advised to review the processes established as part of the nuclear industry's efforts to minimize bias during expert elicitations. See the Senior Seismic Hazard Analysis Committee (SSHAC) guidance documents (e.g., NUREG-2213, NUREG/CR-6372, and NUREG/CR-2117).

A major issue related to this is the inclusion of stakeholders to participate in risk estimation. They typically do not have the training to develope valid estimates and are almost always driven by motivational bias. Risk estimation teams should be as independent as possible and use stakeholders for providing information on the project rather than elicitation participants.

I have observed significant over sensitivity to failures, events leading to perceived near failure, and less recognition of the 10's of thousands of dams performing well for decades constructed under the state-of-practice at the time of construction. Over sensitivity to requiring filters in all instances vs factual phreatic data, material properties and regional/local dam performance. There are significant regional differences in dam construction, material and foundation conditions.

When motivational bias is connected to one's job performance (i.e. good performance = little risk present), it is almost impossible to recognize. This can happen at any level within an organization (senior leader to asset owner).

I haven't seen it often but sometimes motivational bias or just flat incompetency by one elicitator will trigger strong reactions from other team members. I do think these situations may cause some team members to harden their positions but I doubt it makes them be more or less conservative than they think is appropriate.

Figure A.18: Question W1 cont.

Appendix B

Chapter 3 Supplemental Material

B.1 Industry Experts Interview Notes

Discussion with Dam Safety Industry Experts on Expert Opinion Pooling - 19 December, 2023

The attendees at the interview were three dam safety experts with extensive experience in dam safety and risk-informed decision making. All three are employees of a U.S. federal government agency with responsibility for dam safety. Two of the employees are decision makers and one is responsible for the design of large projects related to safety and reliability of dams.

I introduced the topic of expert aggregation and gave a brief overview of my research interest related to the topic.

We discussed the typical makeup of an expert panel providing subjective risk estimates for a dam safety risk assessment. When asked if weighting experts by discipline or expertise might be desirable, the industry experts suggested that equalizing representation by discipline or role might be preferable to possibly eliminate dominance of a particular point of view that might be more prevalent in a single discipline. When asked about weighting the opinions of individual experts when calculating the aggregated probability distribution, the consensus was that equal weighting was standard practice. There was little interest in pursuing alternate weighting schemes.

During this discussion, I mentioned that linear pooling had been used on the project for which I had the data for my work, and it was mentioned that a simpler technique was often used. In this scenario, each expert would provide their best-case, worst-case, and most likely estimates, but rather than generating geometric or PERT distributions for each expert and performing linear pooling on the set of distributions, all the best-case point estimates were averaged, then the worst-case, and so on, and a single triangular distribution was generated. This was done due to a limitation of a commonly used software package for aggregation. When asked if multi-modal pooled distributions were undesirable, it was stated that they were not. The software limitation and the simplicity of the calculation was the driving force for the simpler method.

I was quite surprised to hear that such a method was used as it would eliminate a fair amount of data from the ends of the distribution. When I asked if the attendees were concerned about losing information, either with the simpler method or more sophisticated methods. The attendees did express a desire to somehow capture "outliers".

The focus then shifted to weighting individual experts as it relates to team dynamics and expert motivation. The group believed that applying disparate weights could cause experts to be negatively affected emotionally, thus affecting their participation.

When asked if the attendees believed that equal weighting produced a pooled distribution that is equally representative of the opinions of all the experts, the response was mixed. Two attendees thought that it should produce an equally representative distribution, while another did not believe that it would though none of the attendees were aware of a metric to measure representativeness.

When asked if such a metric would be useful, there was some discussion on its possible utility. One attendee questioned whether it could be used to spot key divergences of opinions during elicitation. Another suggested that it could be used to explain the influence of "outlier" opinions. Its use as a sensitivity tool to determine those nodes which may account for a large amount of variance in the final result was considered. There was some discussion of using it as a "QC" tool to determine if behavioral aggregation was sufficient to build consensus. One of the decision makers then explained that the narrative explanation of the numerical results was quite important in the process. It was stated that the narrative, rather than the numerical analysis, is the primary resource for decision makers.

It was then stated that this metric could be used to identify nodes where there may be a need for further engineering analysis or data gathering to improve the quality of the likelihood estimates. This would essentially serve as a way to prioritize those nodes relating to issues where, perhaps, more technical rigor could drive consensus by providing more objective data.

The attendees were comfortable using equal weighting despite the perception of some downfalls like information loss. The attendees were thanked for their input and agreed to provide written responses to follow-up questions.

B.2 Industry Experts Questions and Responses

- **Question 1:** With regards to taking the arithmetic average of the experts' best case, worst case, and most likely estimates then developing a single triangular or PERT distribution to describe the aggregated distribution for the group: Are you satisfied that this is an acceptable aggregation method?

  Response (Industry Expert 1): I would generally consider this an acceptable method. However, it does have the shortcoming of not fully reflecting the uncertainty of the experts. For decision-making regarding Dam Safety issues, I think it would usually be sufficient.

  Response (Industry Expert 2): While I find the current methodology of averaging risk "voting members'" adequate, it is far from giving an accurate profile of the potential risk weighted by the level of expertise. I would like understand better the differences in risk using more rigorous methods. It would be interesting to see the comparison to see if it would change our understanding of risk and ultimately our decision making.

  Author's Note: Industry Expert 2 provided a combined response to Questions 1, 2, and 3.

  Response (Industry Expert 3): I am satisfied that is an acceptable aggregation method, though I think there must also be some representation of uncertainty within the expert elicitation which accompanies the aggregation.

- **Question 2:** Is the natural tendency of this method to reduce the range of likelihoods a concern?

  Response (Industry Expert 1): Yes, but more from a purist's viewpoint than practical implications.

  Response (Industry Expert 3): Not necessarily, so long as a measure of uncertainty in the ranges is provided.

- **Question 3:** Would you like to see a comparison of this method to a more rigorous method of linear pooling of each expert's entire distributions rather than their values at three points?

  Response (Industry Expert 1): Perhaps. I think it depends on the complexity of the problem, the importance of the estimate, the number of experts providing the estimate, and how disparate their estimates are.

  Response (Industry Expert 3): I think that could be useful.

- **Question 4:** With regards to the more rigorous methods as described above: Is the resulting non-unimodal output distribution of this method a concern for you? Why or why not?

  Response (Industry Expert 1): A concern to the extent that it may overly complicate and lengthen the process, yes, but from a technical standpoint, no.

  Response (Industry Expert 2): I do understand the current methodology may be over simplified since lessor experts can sway the risk characterizations.

  Response (Industry Expert 3): I think the existence of non-unimodal results is an important thing to know, as it may represent a clear division of opinion into "schools of thought" around a particular idea. However, once the existence of non-unimodal result is known, there remains the problem of distilling those thoughts to a relatively succinct comparator for the issue in question. For this distillation, the numerical distribution (specifically, the distribution-specific measures of central tendency and distribution) may not be as important as knowing the existence of non-unimodality. That is, once non-unimodality is known to exist, one numerical representation of the result may be as good as any other, and neither may be as important as the narrative which captures the team's variability of thought on the issue.

- **Question 5:** Are you more comfortable with this method than the simpler method? Why or why not?

Response (Industry Expert 1): Purely from a technical standpoint, yes, I am more comfortable. However, I also recognize that explicitly capturing each expert's estimates may produce results that seem contradictory or confusing, and that has the potential to be an undesirable outcome.

Response (Industry Expert 2): I am open to piloting a real life elicitation by multiple methods to see the impacts on the resulting risk characterization. My biggest concern is sometimes more complicated can cause significantly more workload without significant changes in the risk characterization.

Response (Industry Expert 3): I think I would be equally comfortable with either method. However, I admit the "more rigorous methods" may make it easier to identify non-unimodality such that they may be useful in specific instances.

- **Question 6:** With regards to the "fairness" or "representativeness" of these or other opinion pooling methods: Do you feel that it is important for each expert's opinion to be equally (or as near as is reasonably achievable) represented in the pooled probability function? Why or why not?

  Response (Industry Expert 1): This is a difficult question to answer. In general, I would say "yes," but I know the implications of equal representation may be surprising. Also, trying to implement equal weighting assumes that all of the "experts" are essentially equally qualified, and I have seen firsthand more than once that this is clearly not always the case.

  Response (Industry Expert 2): Not fair since some expertise is not related to the failure mode but gets weighted equally.

  Author's Note: A portion of this response referencing a specific project was omitted for the sake of maintaining anonymity.

  Response (Industry Expert 3): I think it is important to represent each SME opinion and to weight them equally in numerical calculations. However, I think it is equally important that the narrative capture this influence in descriptive language, and the narrative may occasionally benefit from casting the results in a form which excludes or weights individual SME results. Exclusion or weighting are tools which should be used very sparingly, though. The bottom line is that there needs to be a mechanism by which a decision can be made, even in the face of non-consensus opinions from the SMEs, as this possibility will always exist. There is no need to force SME opinions into a distribution which may yield the appearance of greater consensus.

- **Question 7:** If provided with some measure of "fairness" or "representativeness" of each expert with relation to the pooled probability function, how would you use that information?

  Response (Industry Expert 1): I am not sure how this would be implemented. I don't think I understand the question well enough to answer.

  Response (Industry Expert 2): Easy - I would use it to make better informed risk reduction decisions.

  Response (Industry Expert 3): I am not sure how I would use this information. I think I may naturally develop a sense of confidence in the opinion(s) of one or more individual experts among a panel, and I may be interested in the opinions of those individuals in matters which are controversial. However, this sense of confidence in certain individuals might be influenced by a myriad of factors which I'm uncertain I could qualify in advance of an elicitation (i.e., I am not comfortable I could translate all factors which would give me confidence in an individual SME's opinion into a set of objective qualifications which must be met by each panel member to allow their participation, say).

- **Question 8:** If provided with this measure, would you be interested in adjusting the weights of each individual expert to try to construct a more representative pooled probability function?

  Response (Industry Expert 1): Again, without understanding how this would be done, it's difficult to answer. What I would be interested in is a tool that would help highlight individual expert bias, especially if it were a consistent bias and outside the expert consensus.

Response (Industry Expert 2): Yes.

Response (Industry Expert 3): In general, no. Though as mentioned above (see response to Item 3.a), on very rare occasions it may be interesting to quantify the team's judgments without the influence of a strong dissenting voice.

- **Question 9:** With regards to information loss: Are you concerned that either of the commonly used pooling methods as described above "lose" information when pooling expert opinions? (Information loss in this case being defined as generating a pooled probability function that contains values of zero in certain ranges wherein one or more experts provided non-zero likelihood estimates.)

  Response (Industry Expert 1): Yes, but my level of concern is proportional to the complexity and seriousness of what the estimate applies to.

  Response (Industry Expert 2): Not sure. I do think that those giving input and putting in zero values for those areas they are not experts in can alter specific failure mode risk characterizations.

  Response (Industry Expert 3): Again, this is difficult to gauge. I suspect this concern would be much greater when there are one or more likelihood estimates within the range of "frequent" hazards (nodal state estimates of 1:100 or higher or failure likelihoods more likely than 1:10,000). So, perhaps that is a concern over loss of information, though I must admit that I can't confidently say if the loss of information is always (or even often) significant.

- **Question 10:** If it was shown that either or both methods lost information, would you seek a method that eliminated or minimized information loss?

  Response (Industry Expert 1): I would not seek an estimate that eliminated the loss but would consider pursuing an estimate that minimized the loss of information.

Response (Industry Expert 2): Maybe. More accurate inherently better, but perfection is the enemy of good. I typically strive for the 80/20 rule (80% of the value with 20%of the effort). That last 20% is worth pursuing, but the benefits have diminishing returns. So ultimately, if the extra analysis, time to do that analysis, and the potential for different results all have to be factored into this consideration.

Response (Industry Expert 3): Not necessarily. I do think there is a pretty high level of importance in remaining aligned with some "industry standards" or "common peer practices" to minimize confusion between various agencies and ensure common quality levels between assessments. By necessity, these common practices often employ relatively simple mathematical distillations which do not always convey the richness of opinion. This is to say that the loss of information is insidious to some degree. A team needs to exert a reasonable effort to investigate variations and uncertainty and to capture that in the narrative supporting its decision. However, the tools used to make those investigations should probably not figure prominently in the final narrative of the decision.

- **Question 11:** How important is it to you to include the information from "outliers"? (Outliers in this context would not necessarily be defined in absolute statistical terms but as those experts whose estimates are to some degree more or less conservative than the group mean even after multiple rounds of behavioral aggregation, conversation, estimate revision, etc.)

  Response (Industry Expert 1): With all the caveats satisfied regarding trying to ensure outliers are legitimate outliers, I think it could be important to include outliers. But again, I would correlate the extent of my efforts to include outliers to the importance of the decision under consideration.

Response (Industry Expert 2): It is important to understand the outliers because it may lead to better understanding of the risk. What if we missed something and that outlier clues us into that?

Response (Industry Expert 3): I think it is important to represent every opinion. However, it is equally important to provide a mechanism for dissent (even strong dissent) to exist against a chosen path forward, so that decisions can be made based on the bulk of evidence.

- **Question 12:** How important are the objective results (f-N plot, Monte Carlo results) versus the narrative?

  Response (Industry Expert 1): This is an interesting question. Personally, I am such a numbers/data person that it is easy for me to focus on objective results. However, I do recognize that decisions are based on other factors than just risk, and context through a narrative is important to inform those decisions.

  Response (Industry Expert 2): These go hand in hand. The results have to have a basis informing the narrative which ultimately influences the decisions. At the end of the day, our goal is to reduce dam risk and you don't reduce risk until you take action! Both pieces contribute to the action and care needs to be taken since an uninformed narrative may drive inadequate action/inaction.

  Response (Industry Expert 3): I think representations of the central measure and distribution of the risk are required. In general, though, I place greater weight on the narrative which describes the holistic nature of a decision, regardless of the mathematical representations of risk. I think the value of the mathematics is largely in forcing the team to come up with a quantification. That is, forcing each team member to place a numerical value on their opinion is the thing which drives them to comprehensively consider all of the factors influencing a risk and the significance of each factor. This forced framework is the mechanism which ensures a well-considered characterization of risk will occur. But the numerical depiction of that characterization, as represented by the team's composite estimates of central measure and distribution of risk, are an overly-simplistic snapshot which must be given context through a narrative which defines the logic of the decision.

- **Question 13:** Would having more detailed information regarding information loss, the impact of individual estimators, etc., lead you to explore altering the aggregation process?

  Response (Industry Expert 1): Yes, especially if it were information that was available before the risk estimating process was underway.

  Response (Industry Expert 2): Possibly, dependent on how difficult it is to obtain better accuracy in risk characterization. Ultimately, as an owner, I want to make timely risk reductions while continuously making process improvements.

  Response (Industry Expert 3): I doubt I would alter the aggregation process used in "final metrics" of a study. However, it may provoke language related to exploring uncertainty by any of various means.

- **Question 14:** Would this information lead you to alter your description of the risk in the narrative with no changes to the aggregation process?

  Response (Industry Expert 1): Maybe, but by the time I get to writing a narrative, there have probably been so many estimates made on so many PFMs and nodes that I don't know how relevant it would be at that point or how I would incorporate that into a narrative. I think it would be important to note if one or two experts consistently estimated outside the consensus and what may be driving that.

  Response (Industry Expert 2): It could.

  Response (Industry Expert 3): Perhaps in certain circumstances, yes.

- **Question 15:** Would you be interested in retroactively comparing multiple aggregation methods using data from historic risk assessments and investigating possible process improvements?

Response (Industry Expert 1): This is an easy yes, assuming someone else will do it for me!

Response (Industry Expert 2): Yes. It would be an interesting pilot to understand the value of this improvement of accuracy informing dam risk.

Response (Industry Expert 3): I would definitely be interested in this to "learn for understanding", but not necessarily for altering the process. The process may need to retain some rigidity with respect to common practices shared by various agencies. While the exercise can lead to improved understanding, I am not certain it would lead to improved processes, as any improvement may need to be quite dramatic to motivate changes to a standard process being employed across multiple federal agencies.

Author's Note: A portion of this response referencing a specific agency was omitted for the sake of maintaining anonymity.

Author's Note: Minor typographical errors (e.g., capitalization, spelling) in the expert responses have been corrected for readability. The content and meaning of the responses have not been altered unless otherwise specified.

B.3 Opinion Pooling Results by Method and Node

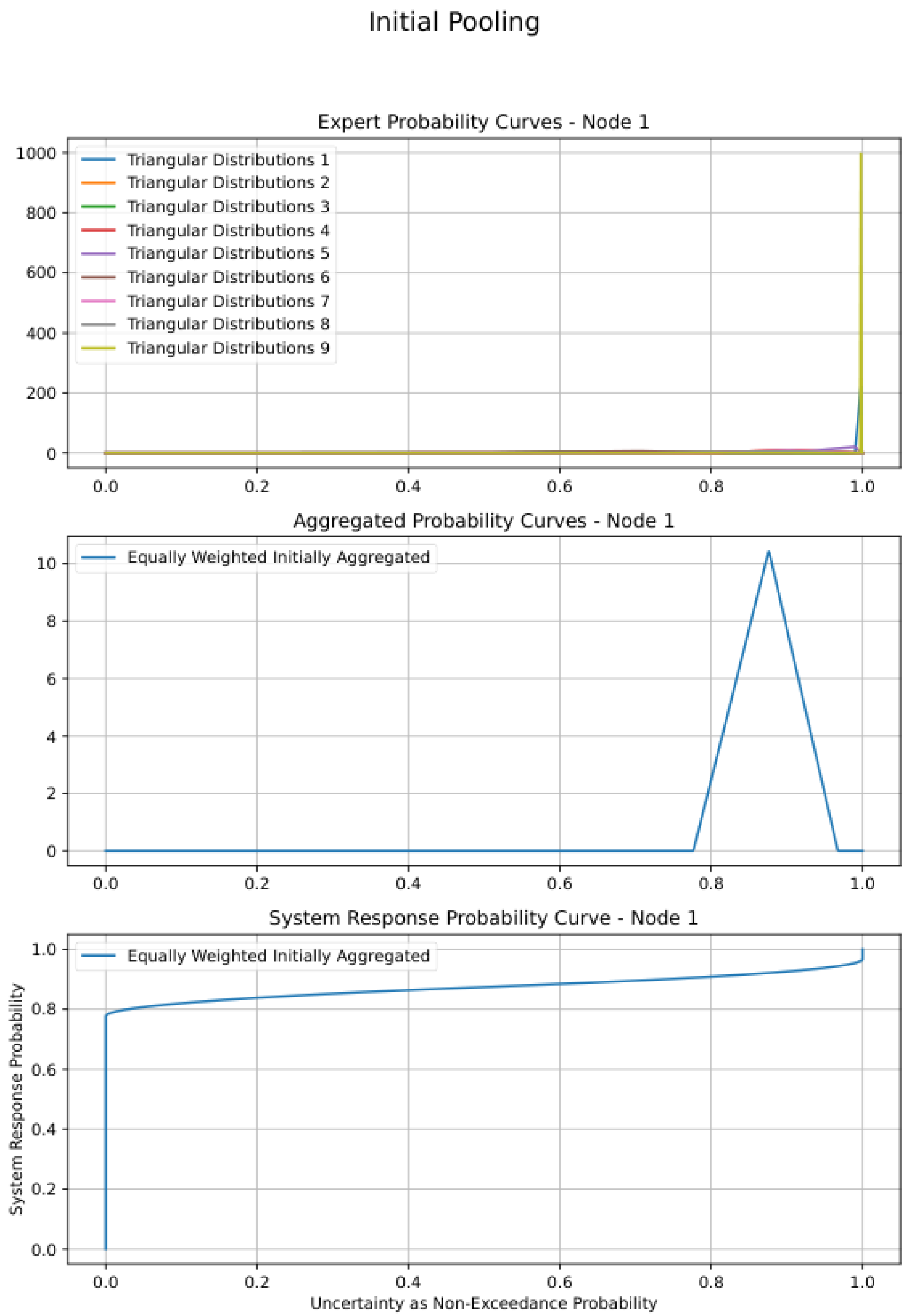

Figure B.1: Initial Averaging - Node 1

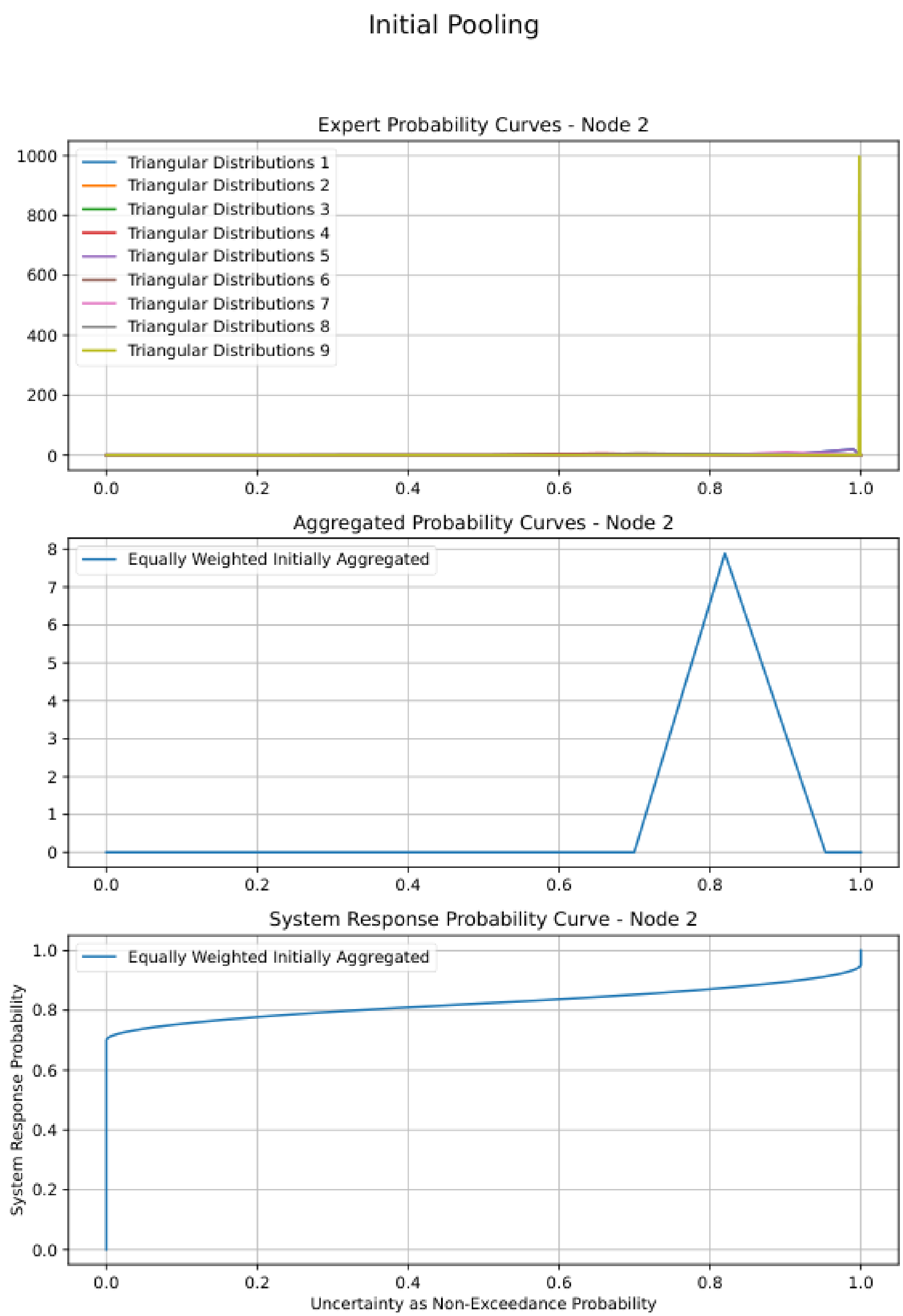

Figure B.2: Initial Averaging - Node 2

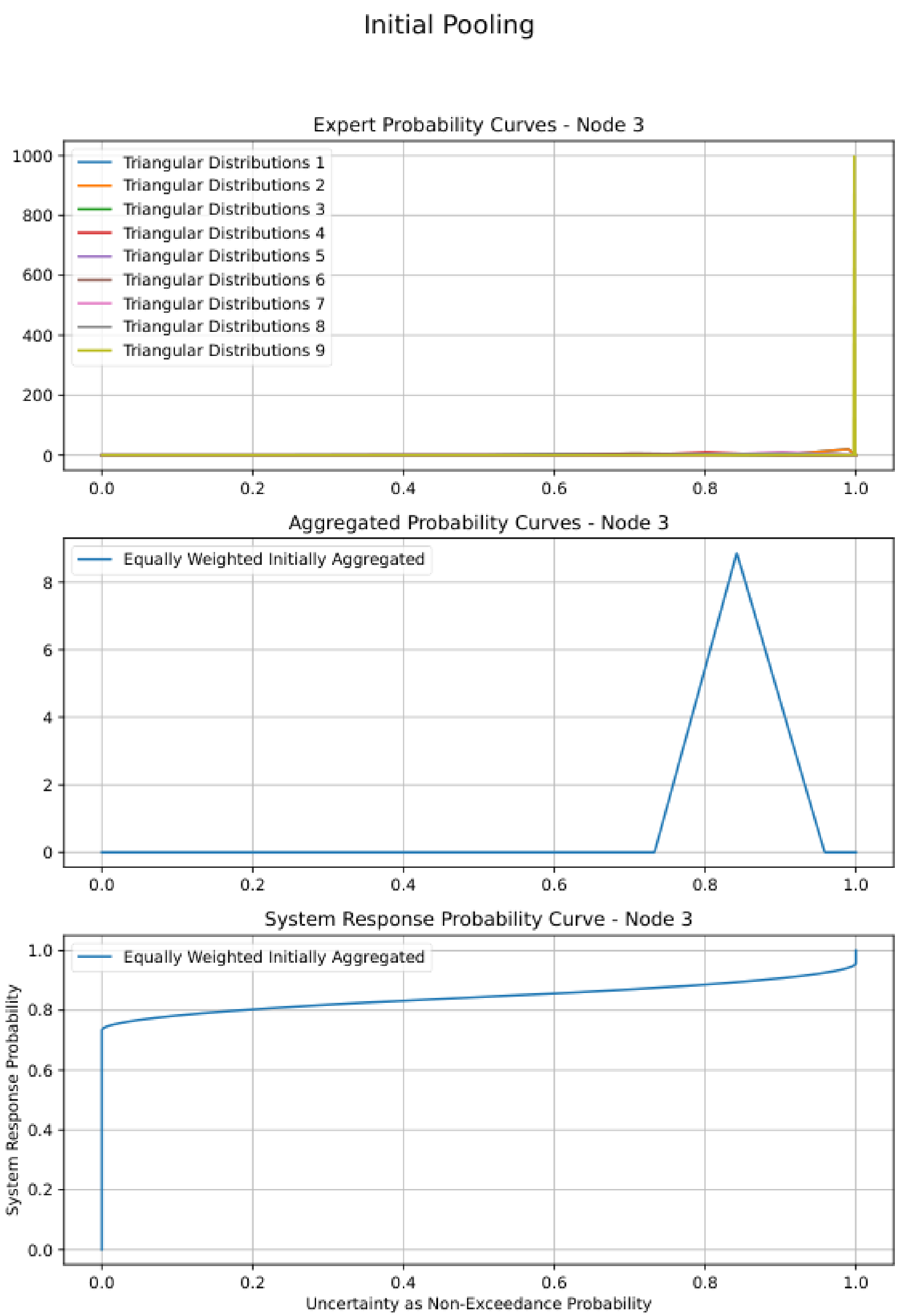

Figure B.3: Initial Averaging - Node 3

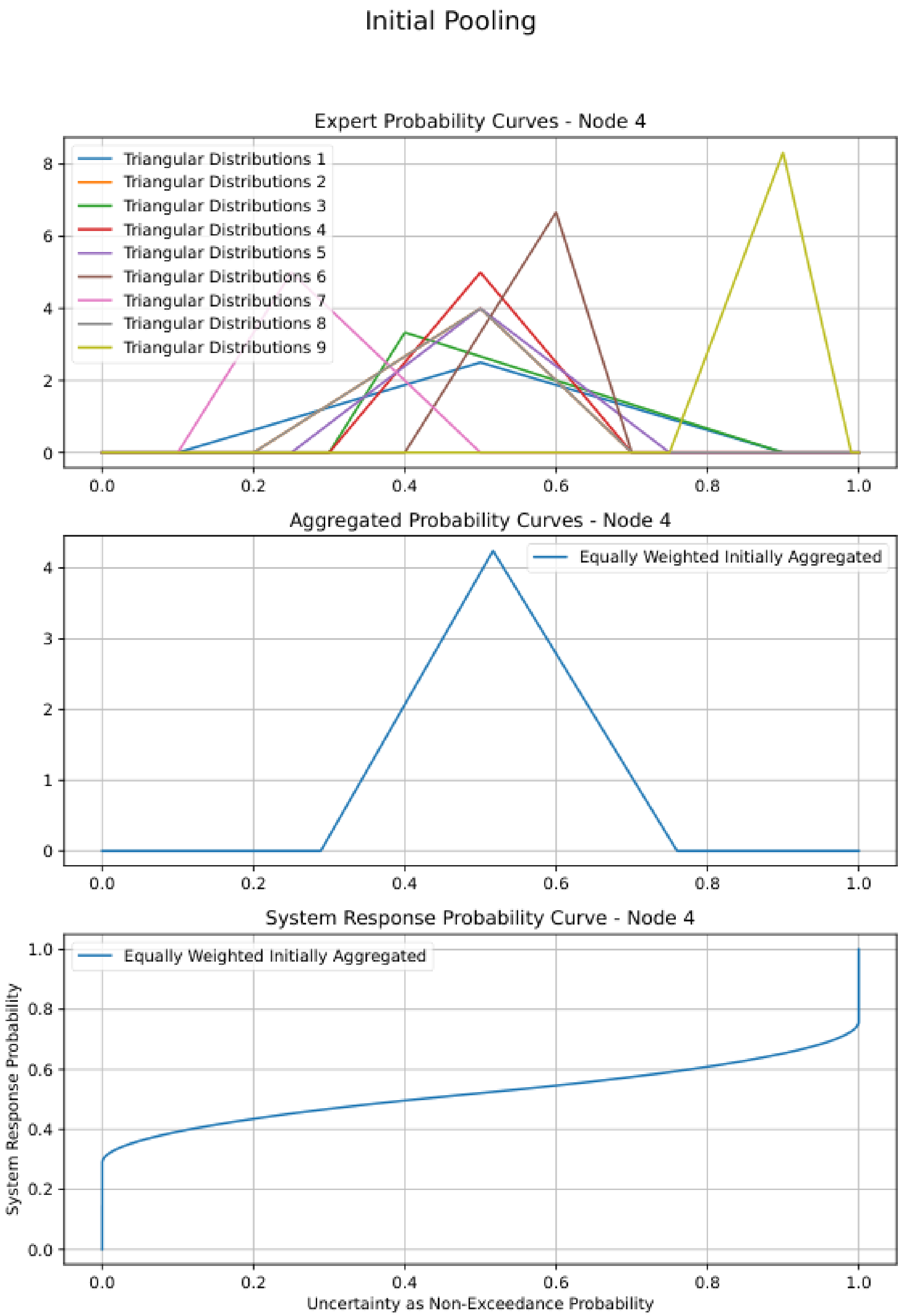

Figure B.4: Initial Averaging - Node 4

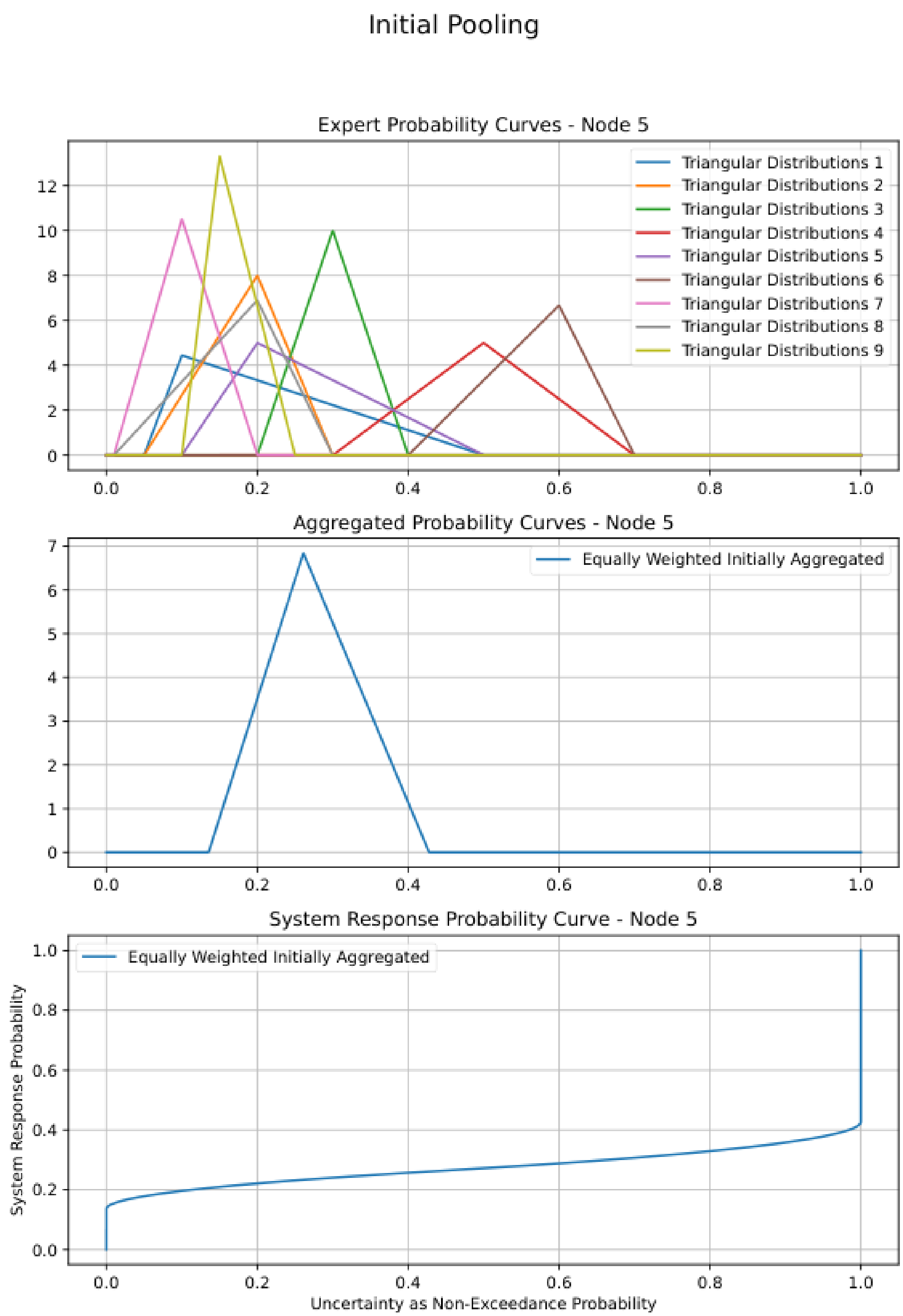

Figure B.5: Initial Averaging - Node 5

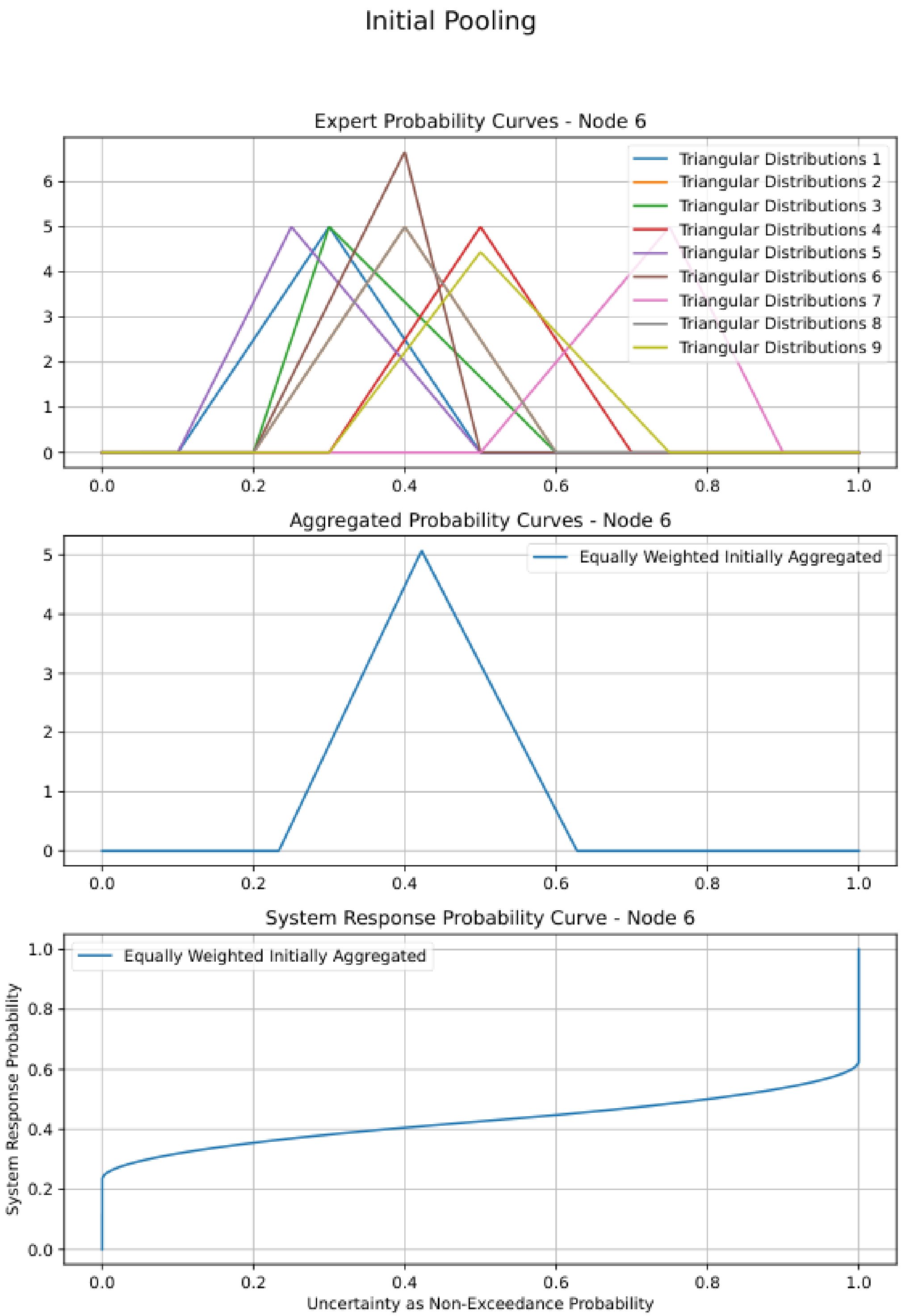

Figure B.6: Initial Averaging - Node 6

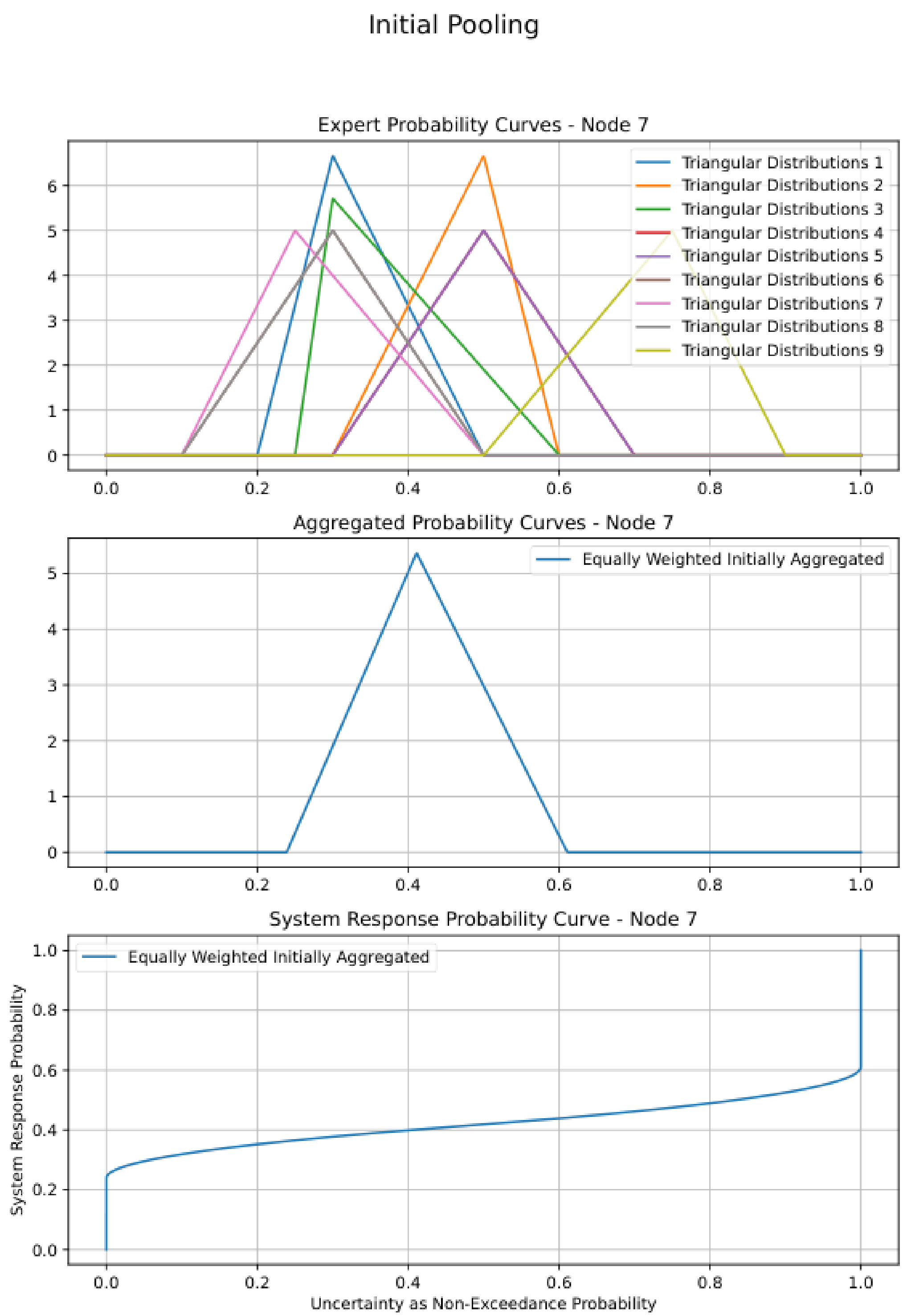

Figure B.7: Initial Averaging - Node 7

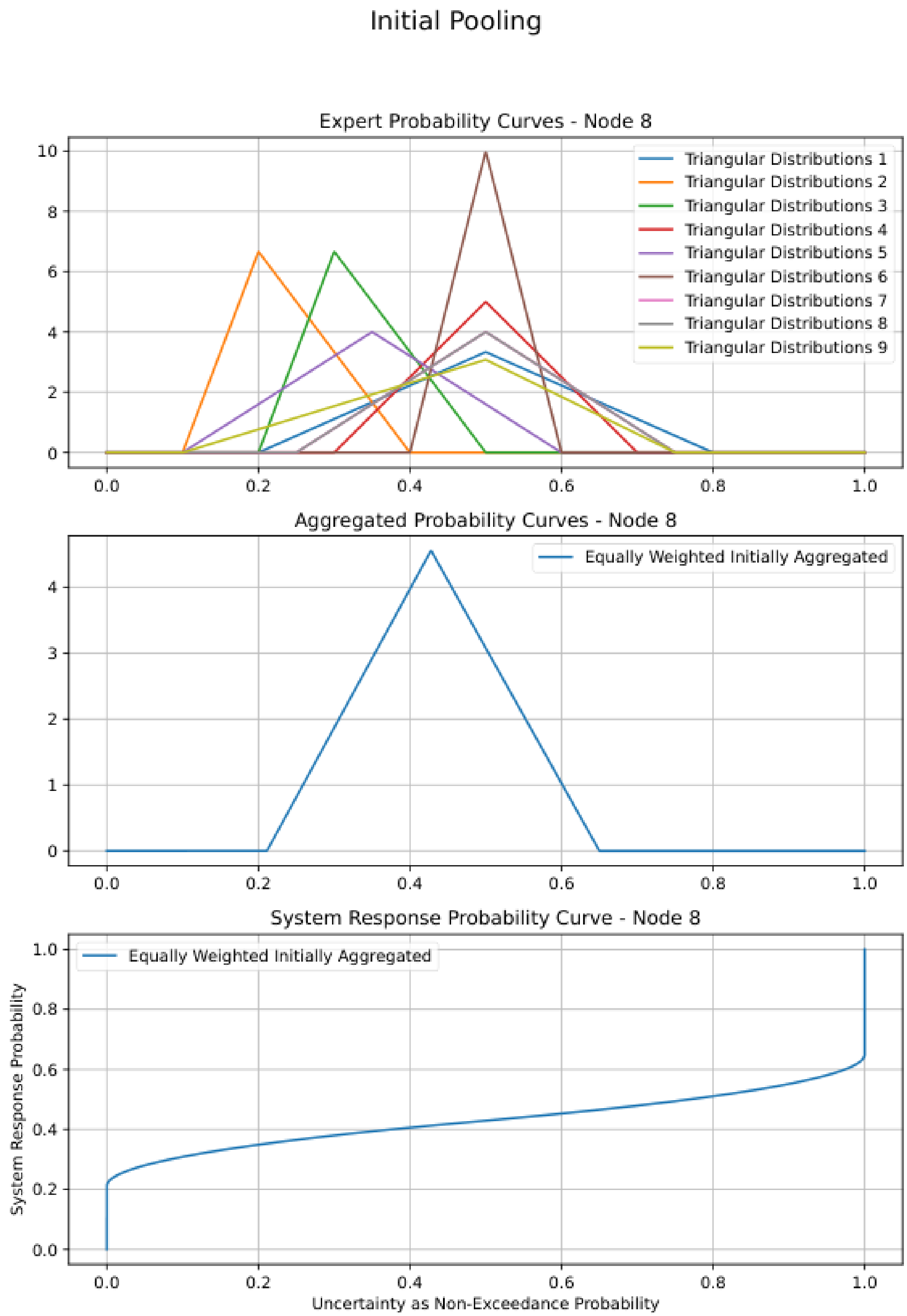

Figure B.8: Initial Averaging - Node 8

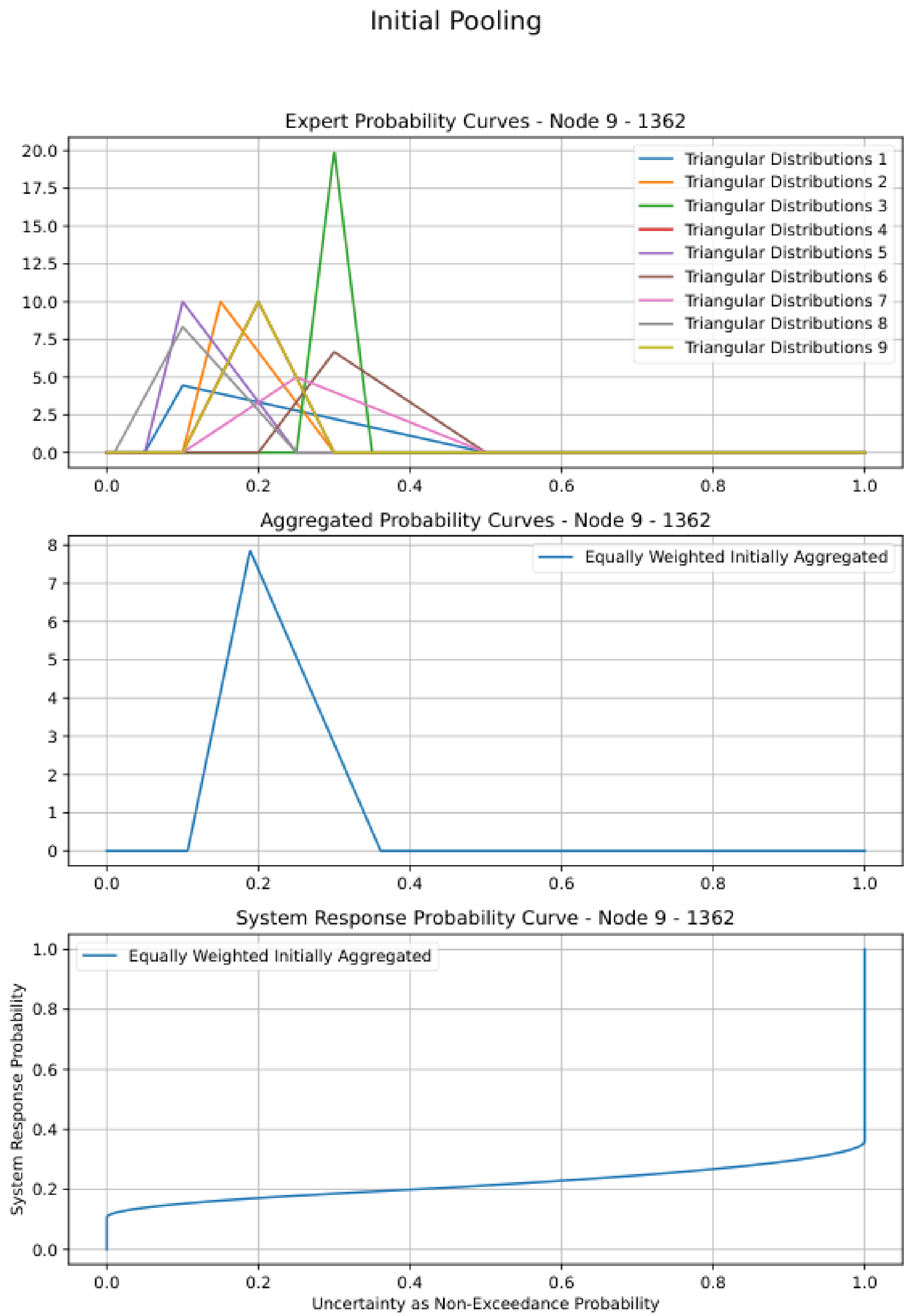

Figure B.9: Initial Averaging - Node 9

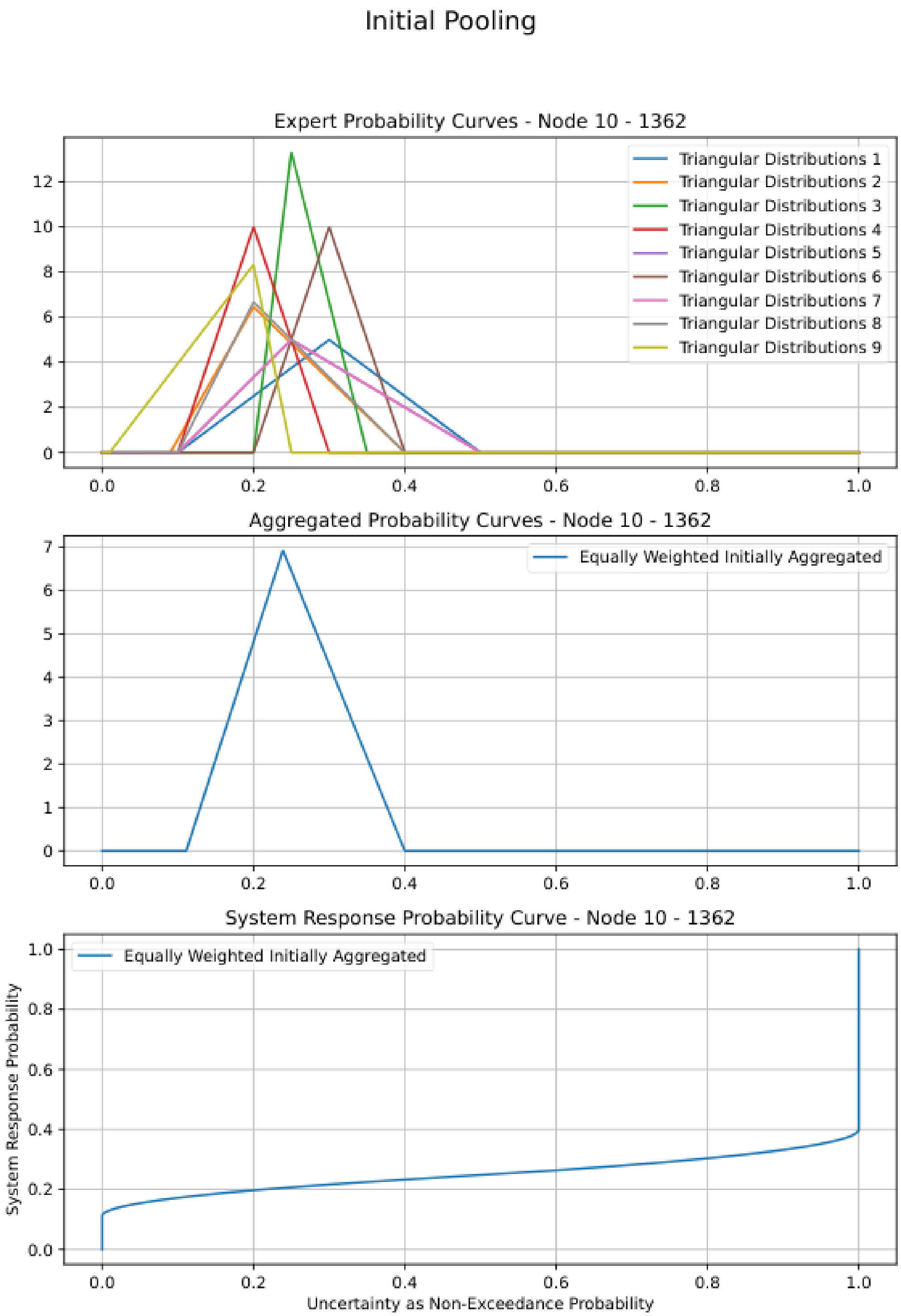

Figure B.10: Initial Averaging - Node 10

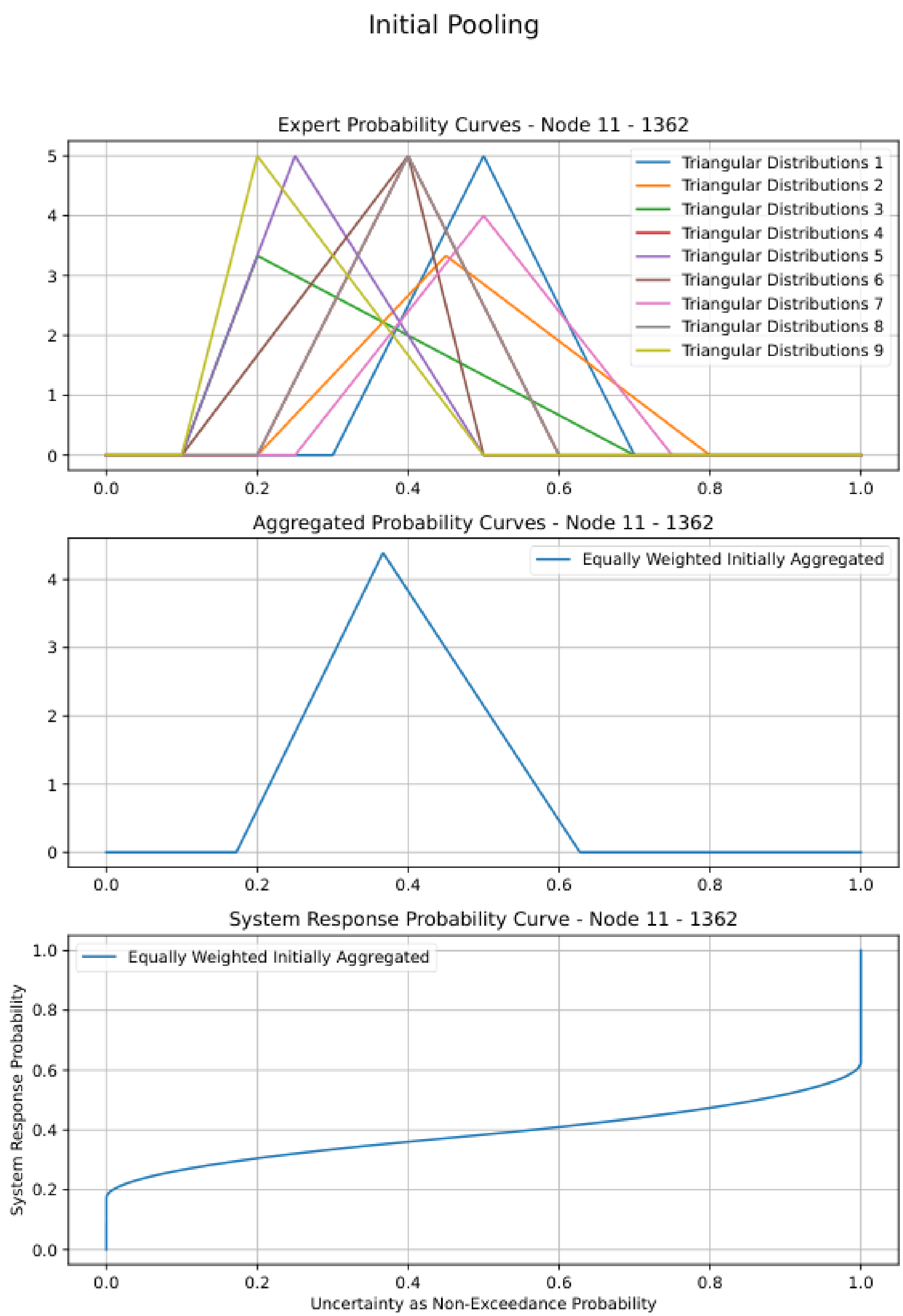

Figure B.11: Initial Averaging - Node 11

Initial Pooling

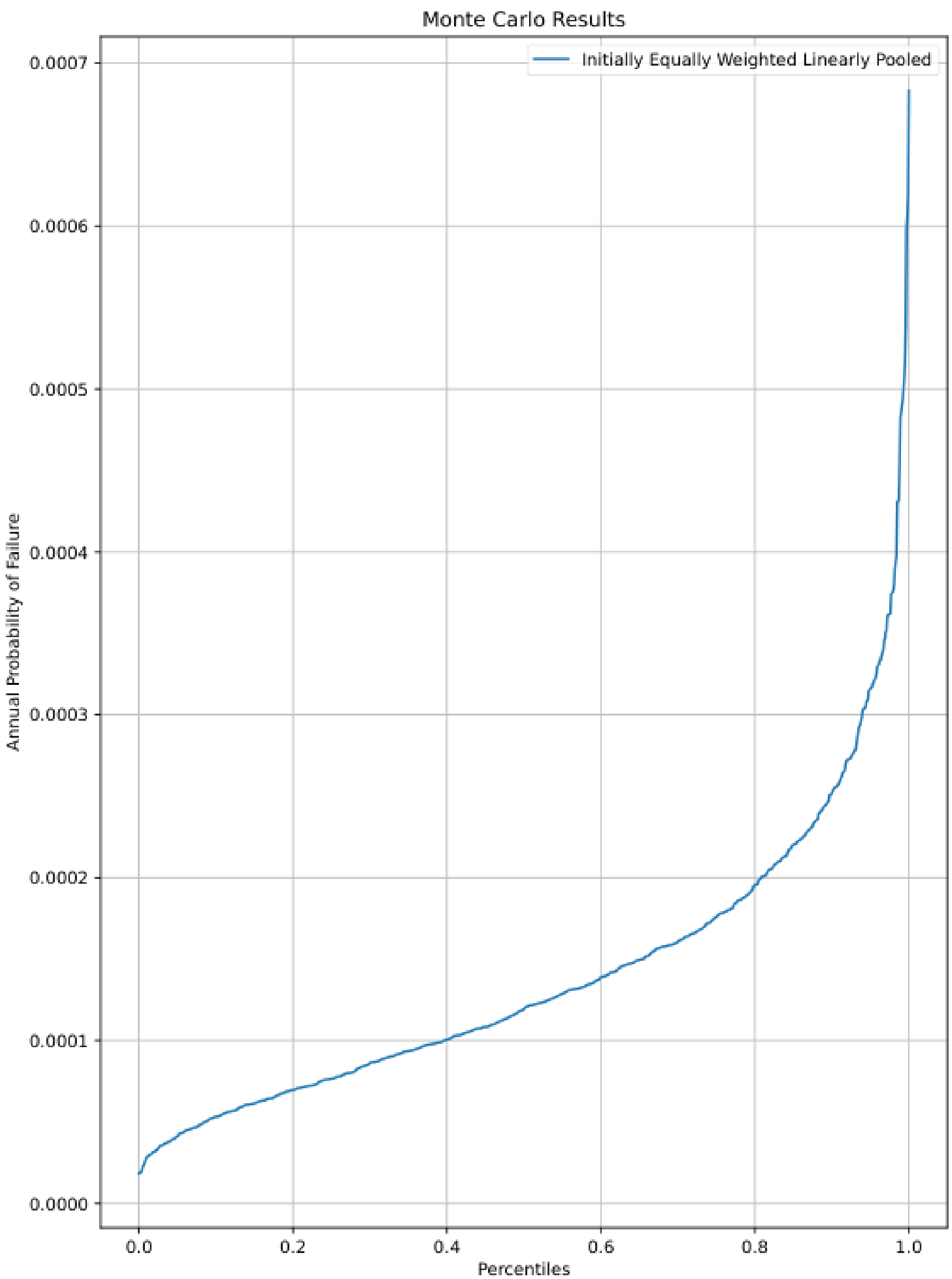

Figure B.12: Initial Averaging - Monte Carlo Results

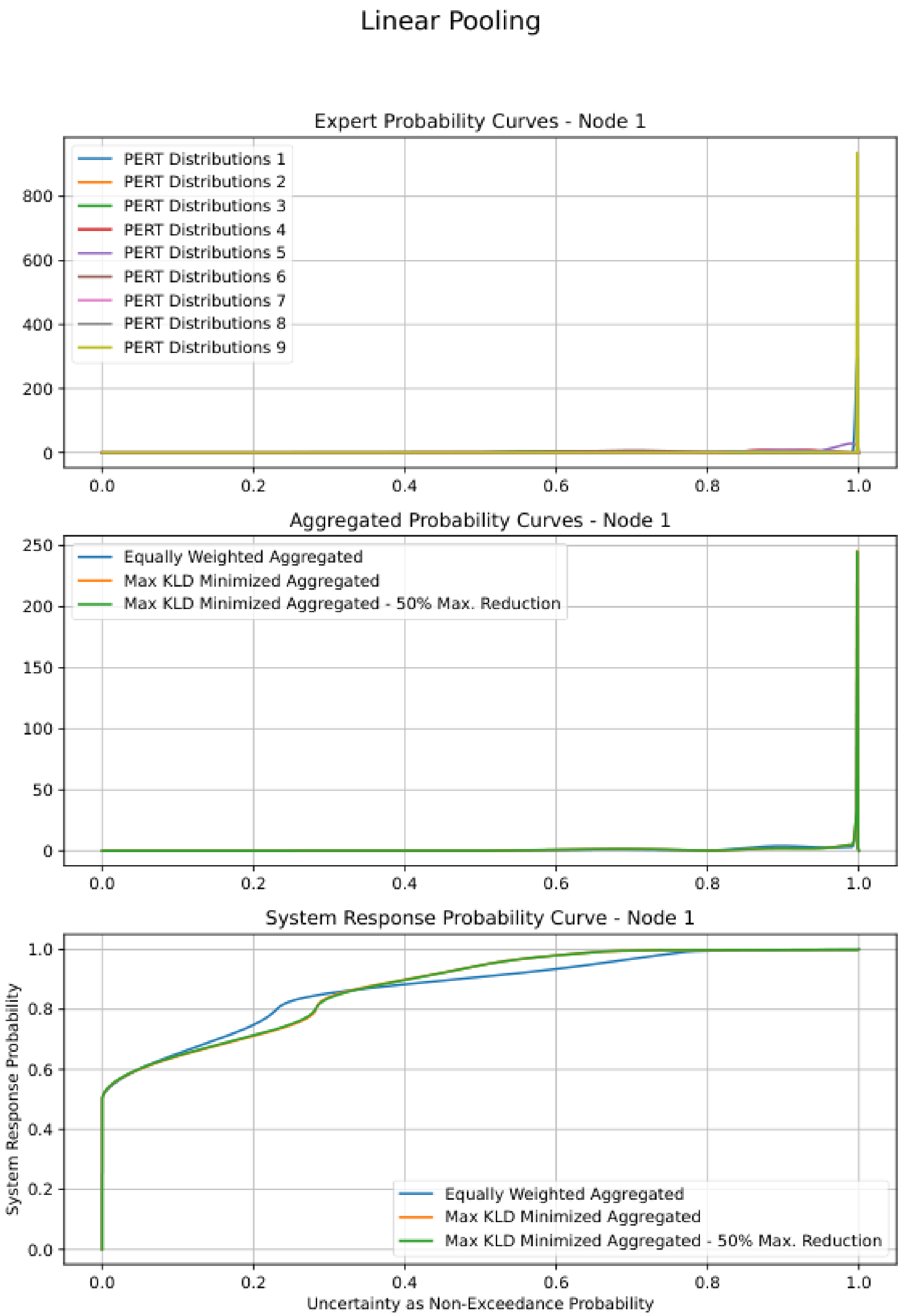

Figure B.13: Linear Pooling - Node 1

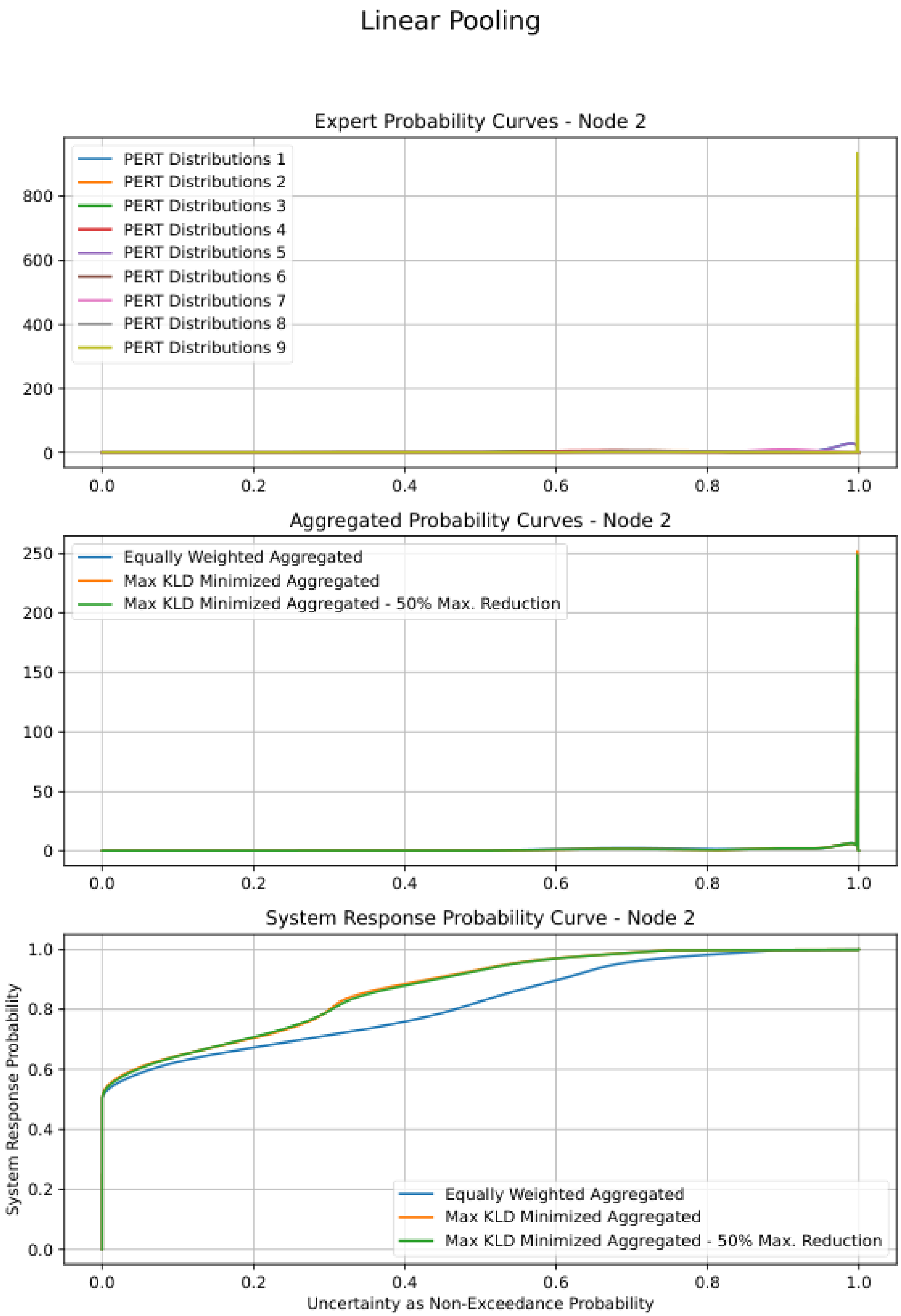

Figure B.14: Linear Pooling - Node 2

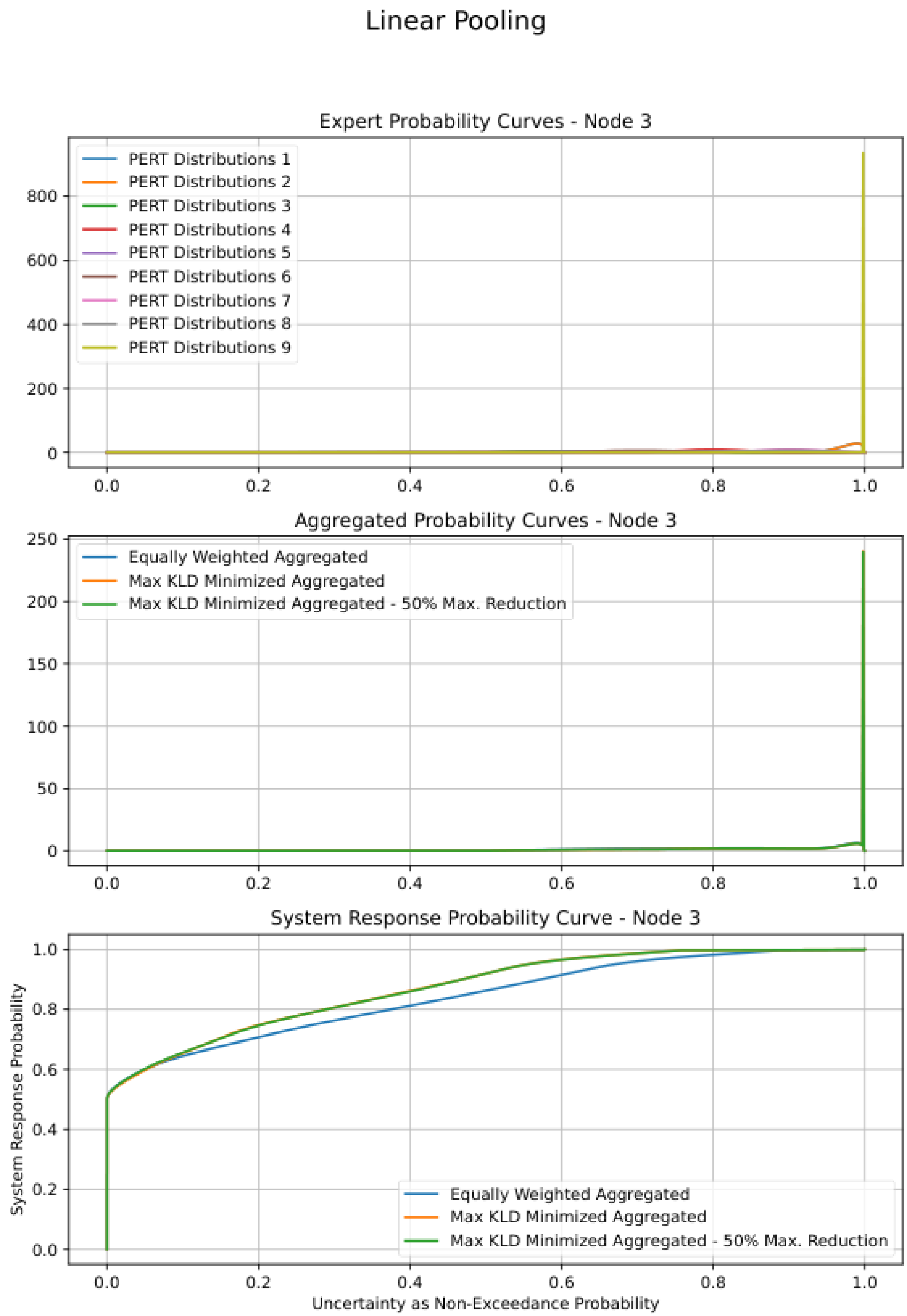

Figure B.15: Linear Pooling - Node 3

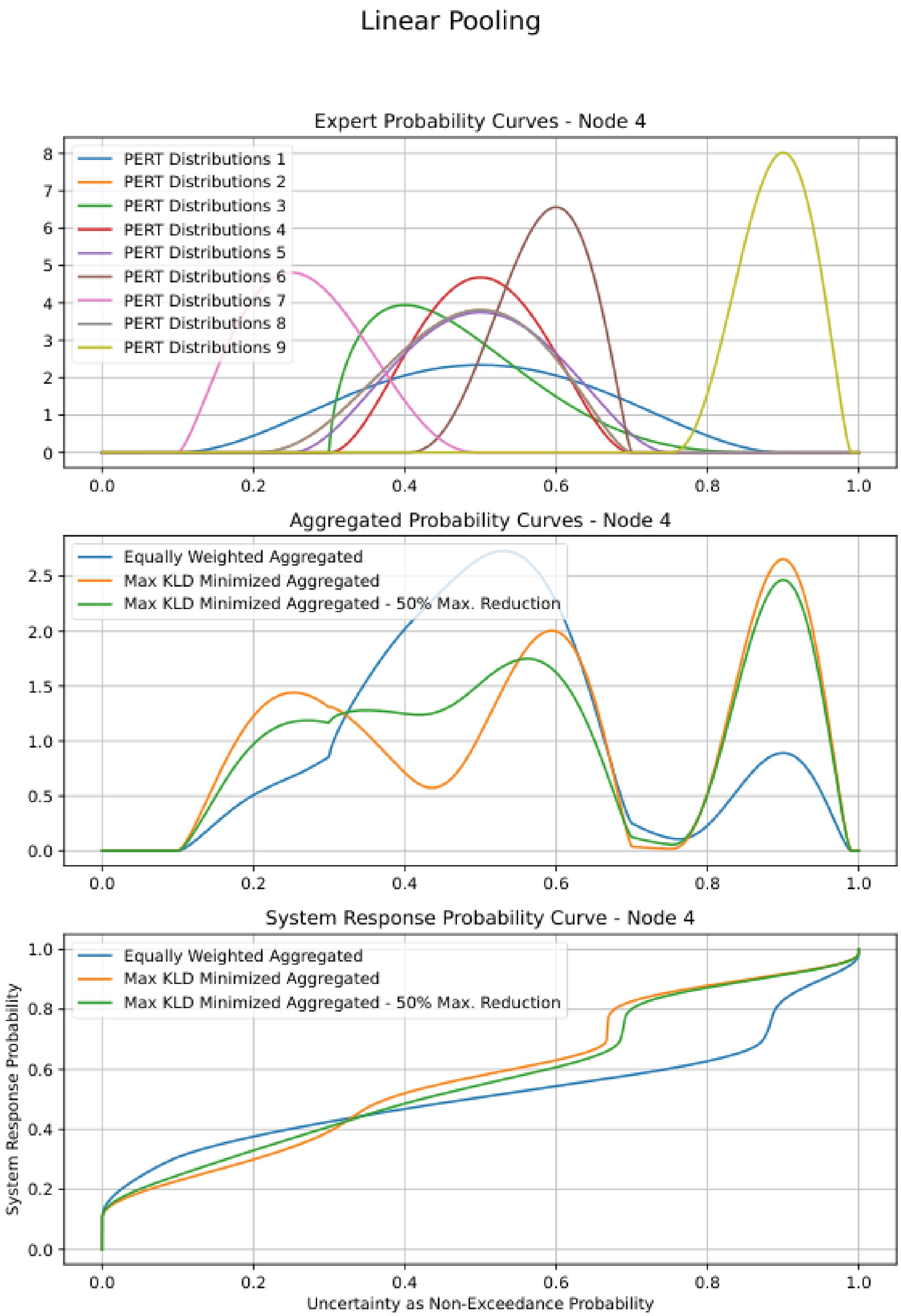

Figure B.16: Linear Pooling - Node 4

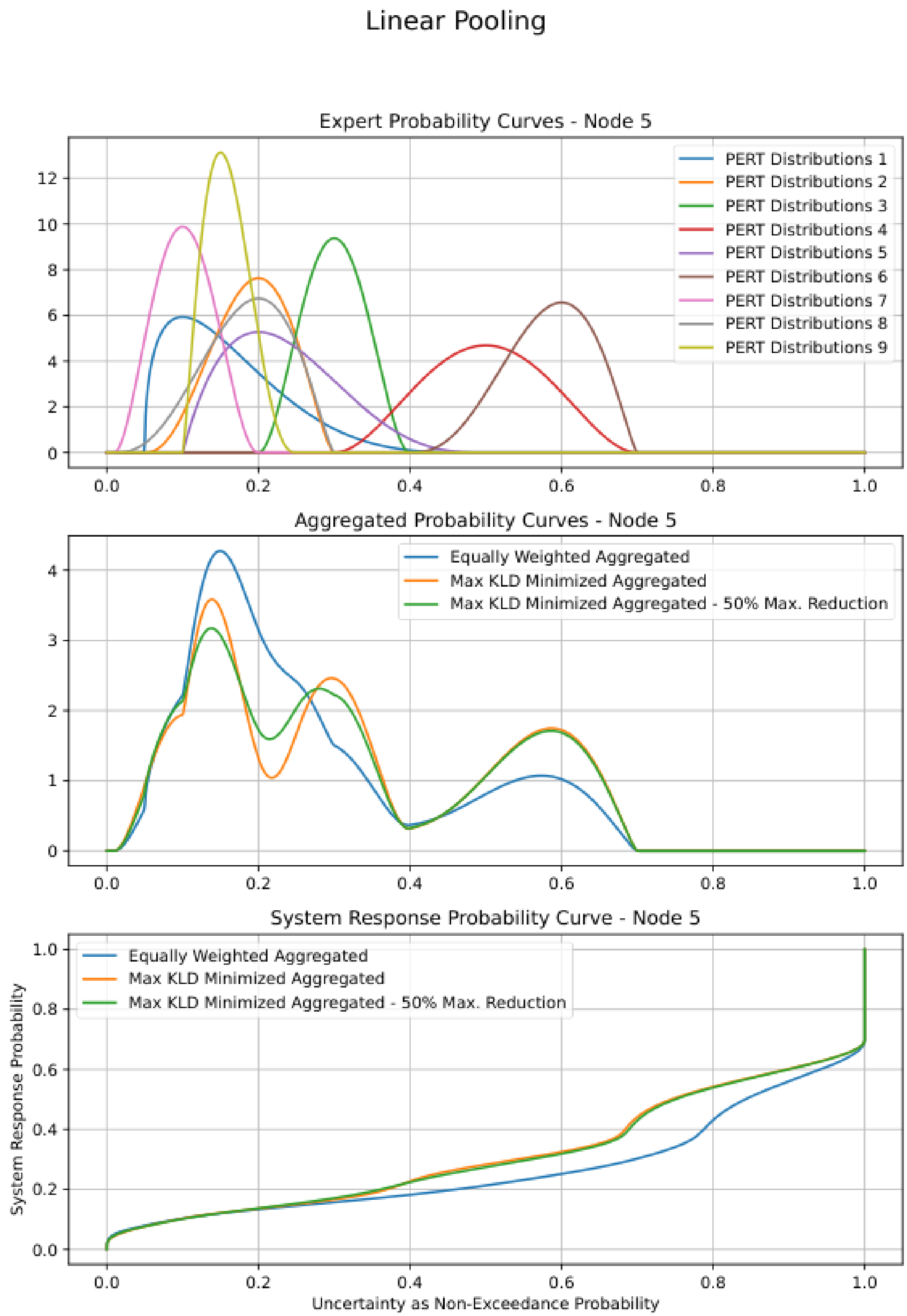

Figure B.17: Linear Pooling - Node 5

Linear Pooling

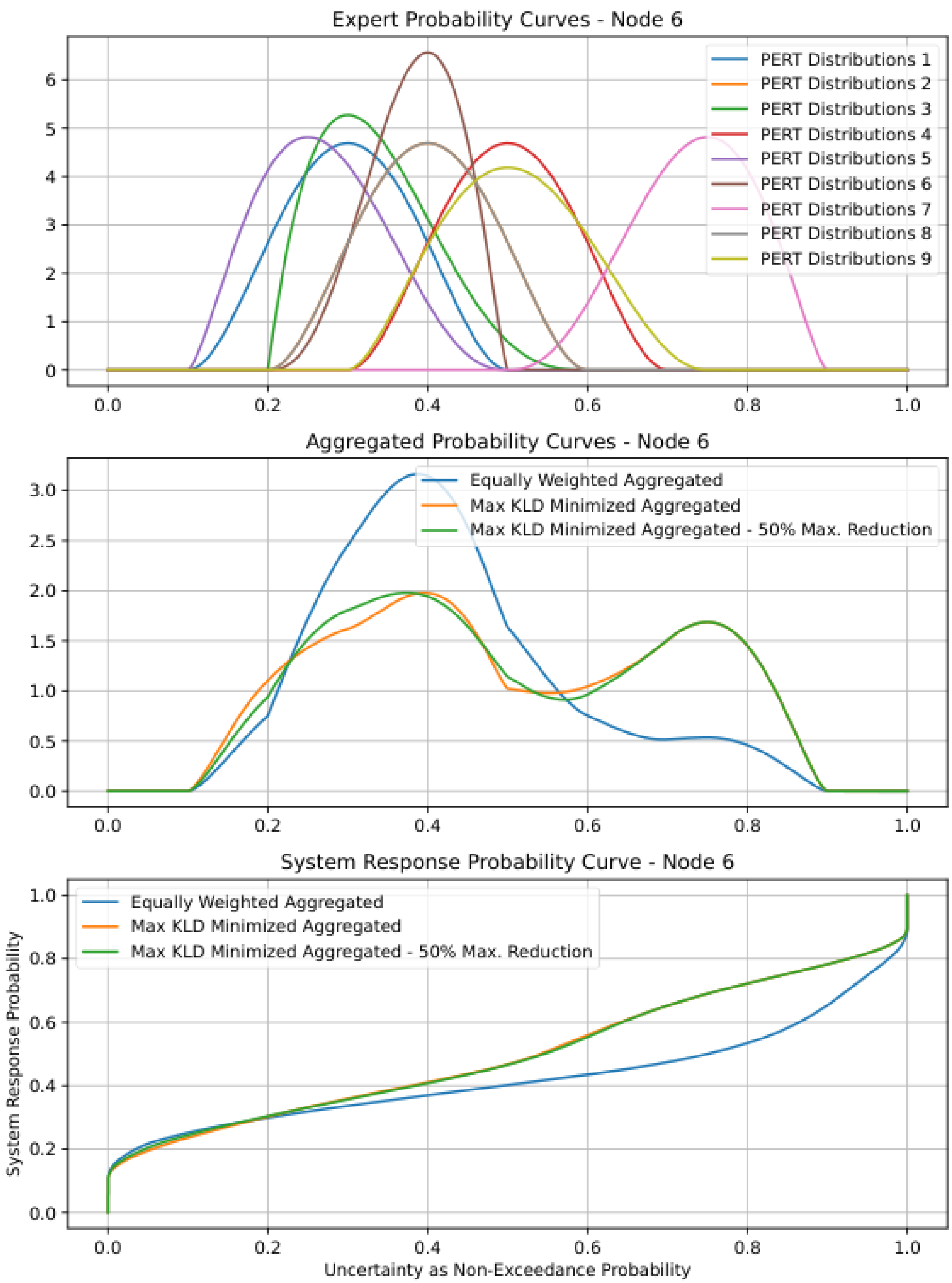

Figure B.18: Linear Pooling - Node 6

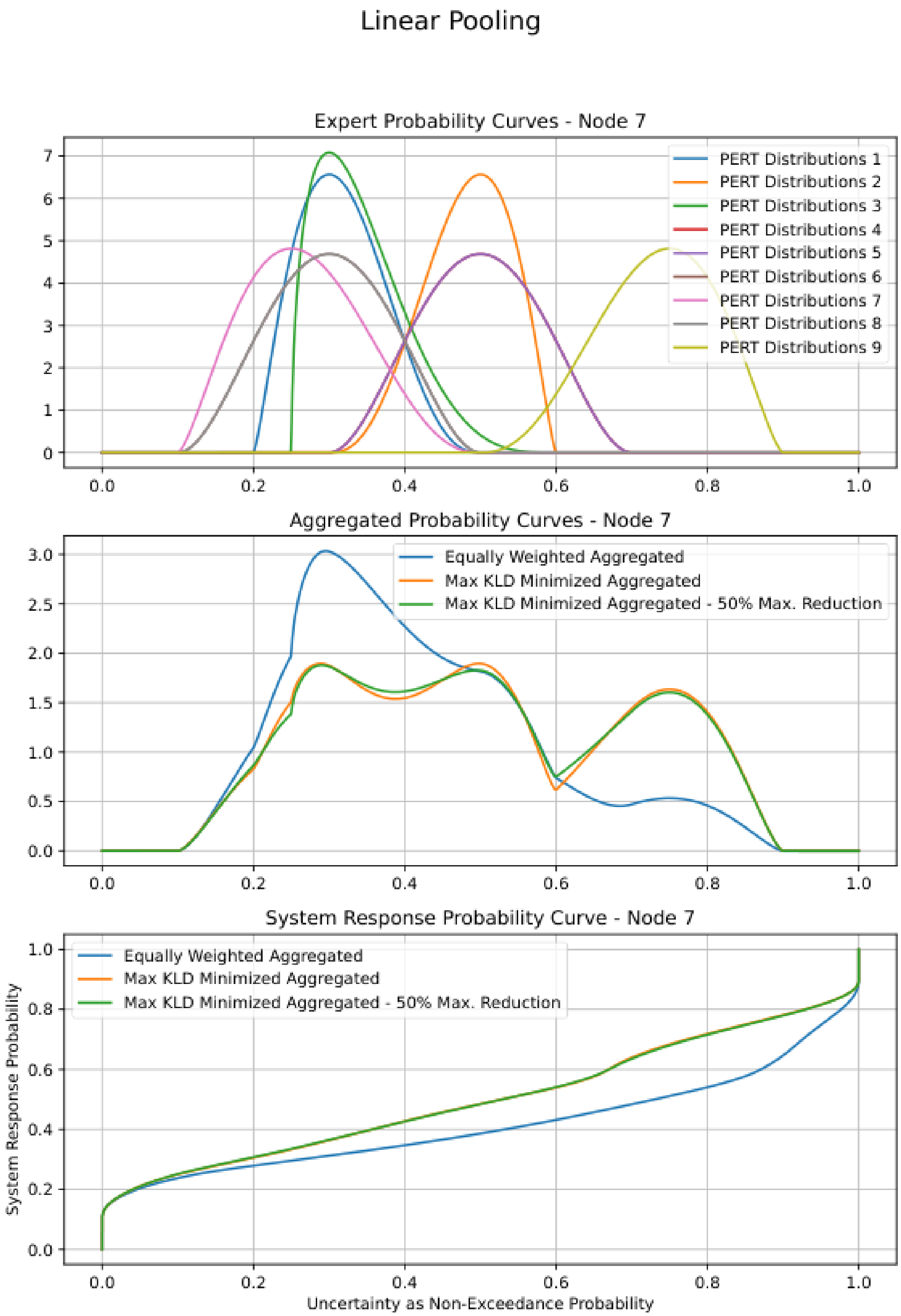

Figure B.19: Linear Pooling - Node 7

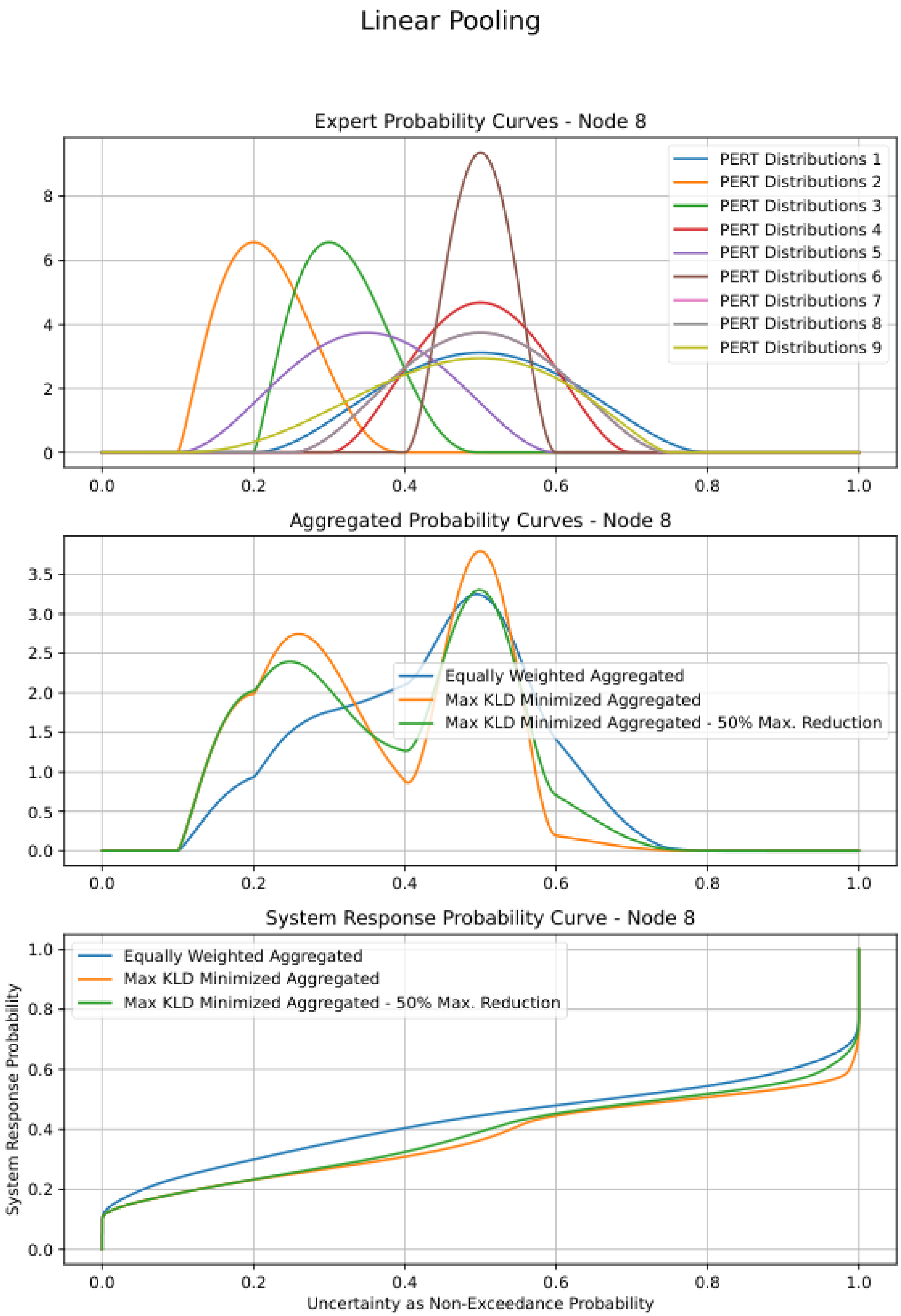

Figure B.20: Linear Pooling - Node 8

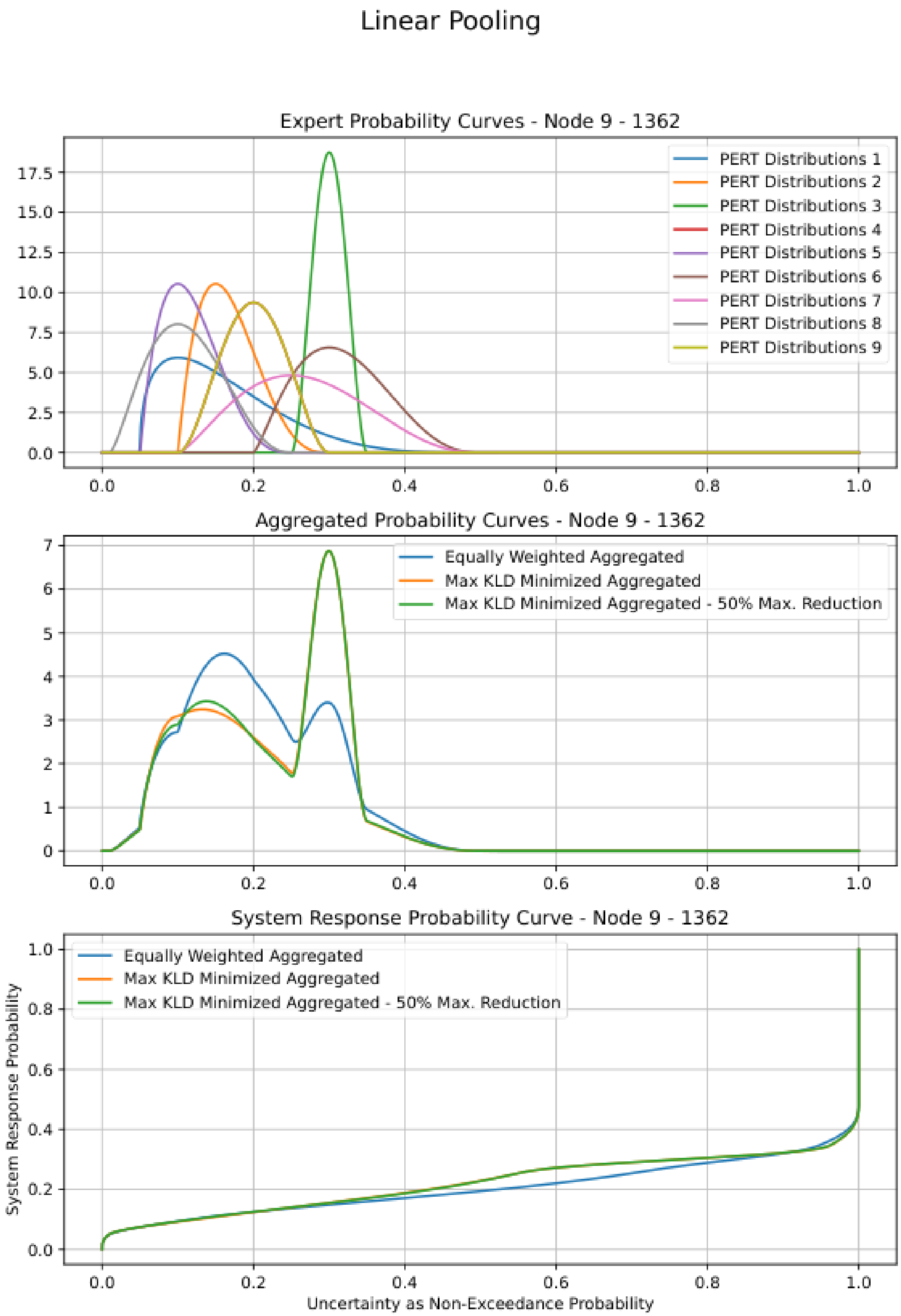

Figure B.21: Linear Pooling - Node 9

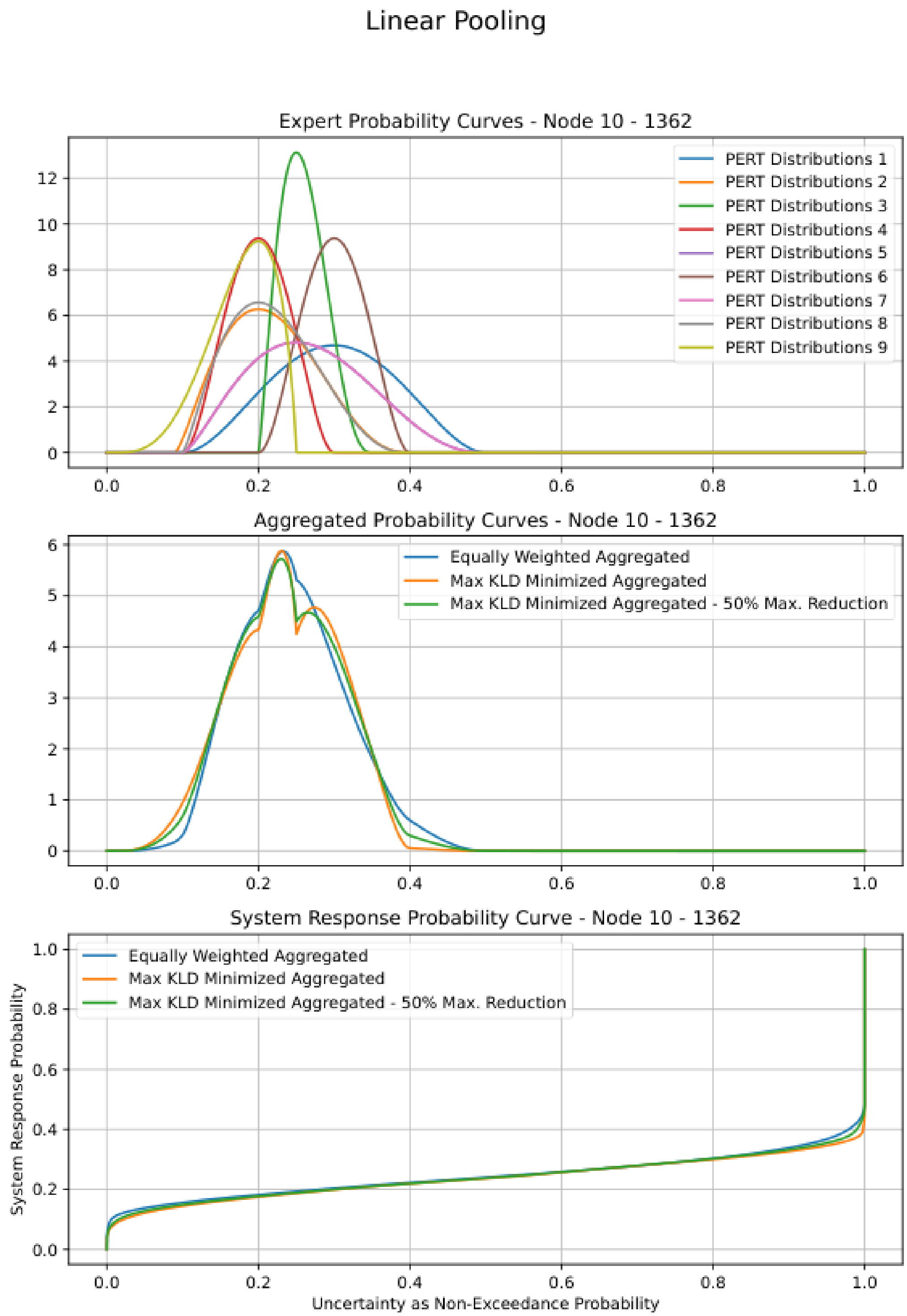

Figure B.22: Linear Pooling - Node 10

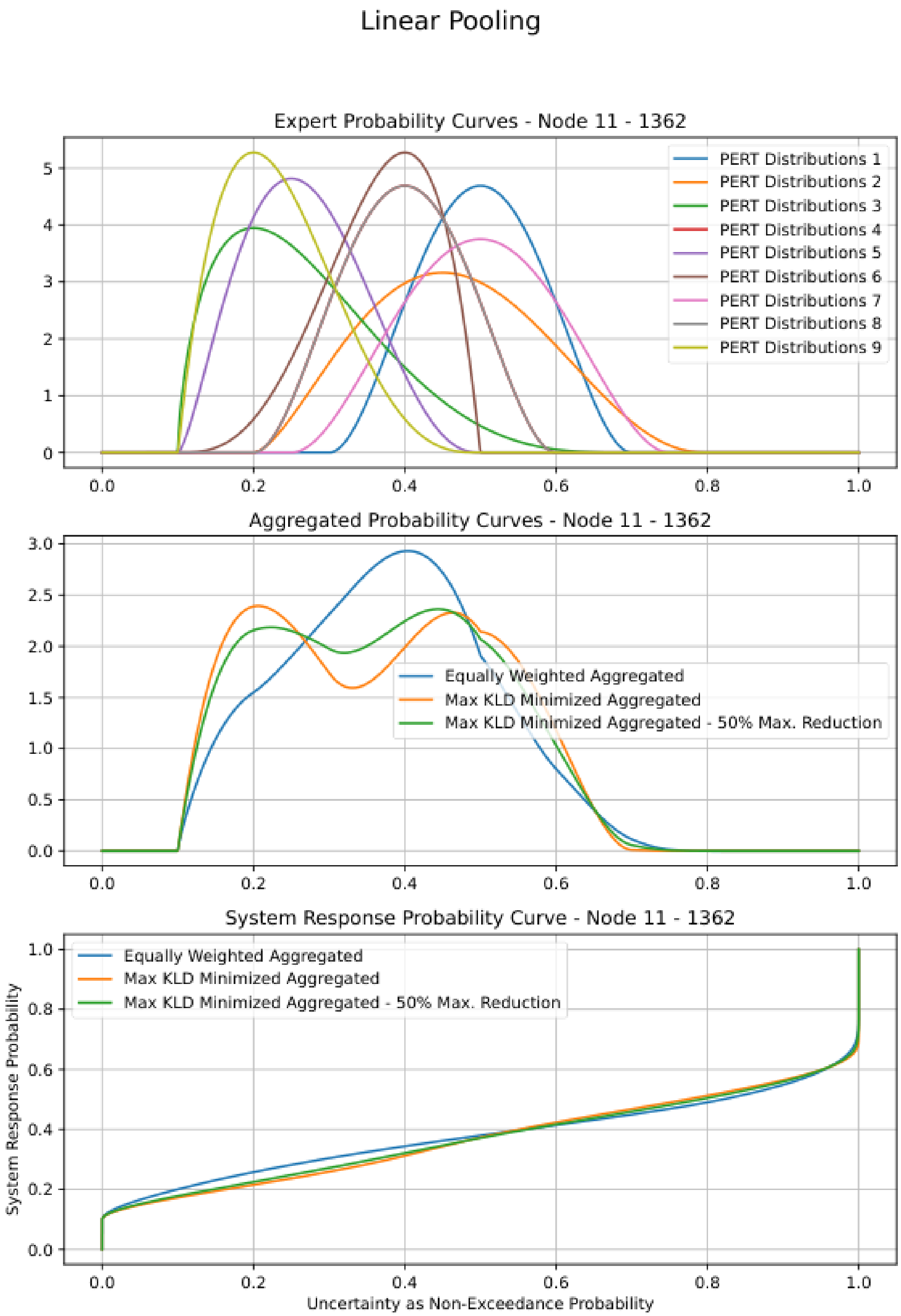

Figure B.23: Linear Pooling - Node 11

Linear Pooling

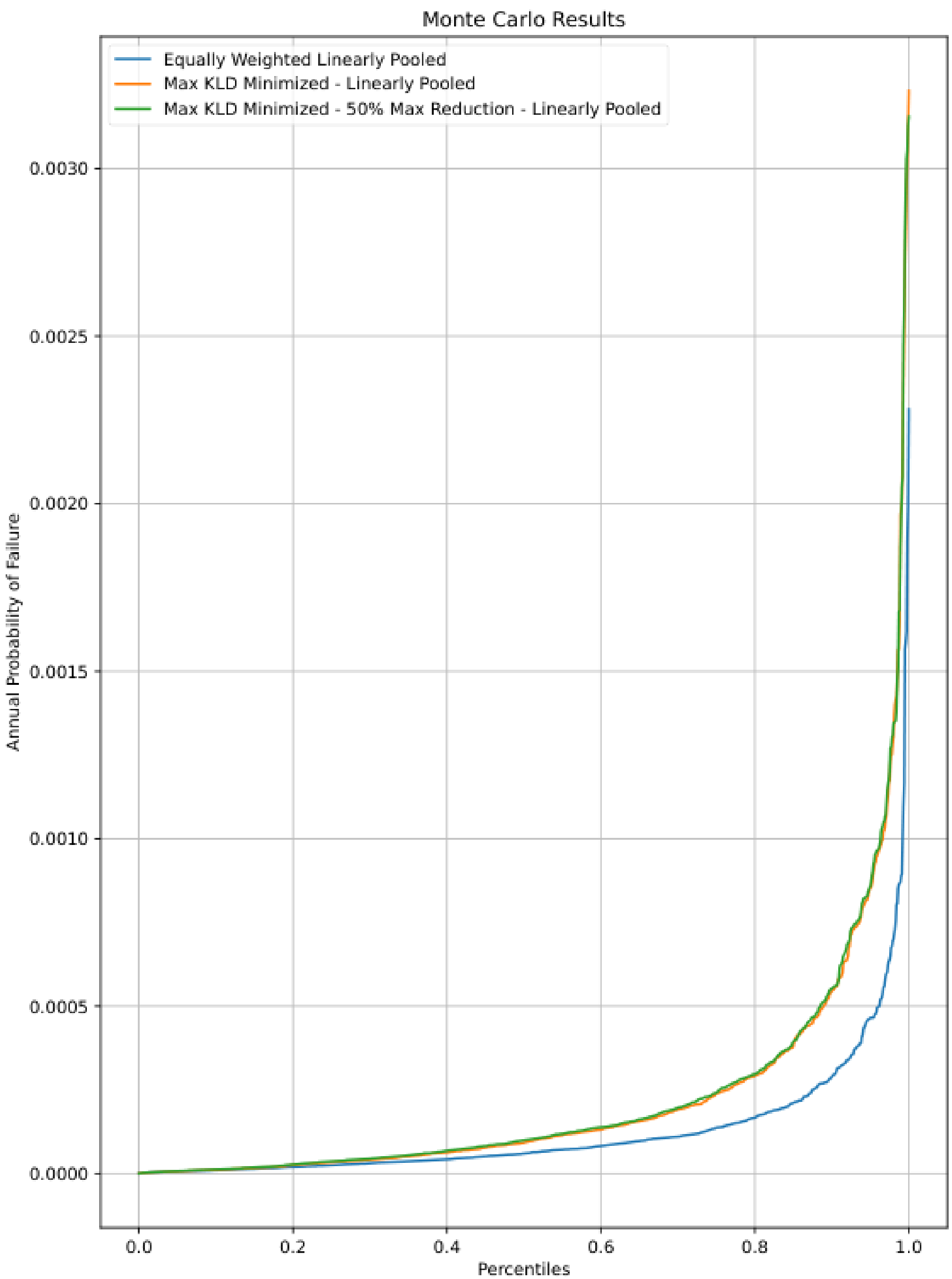

Figure B.24: Linear Pooling - Monte Carlo Results

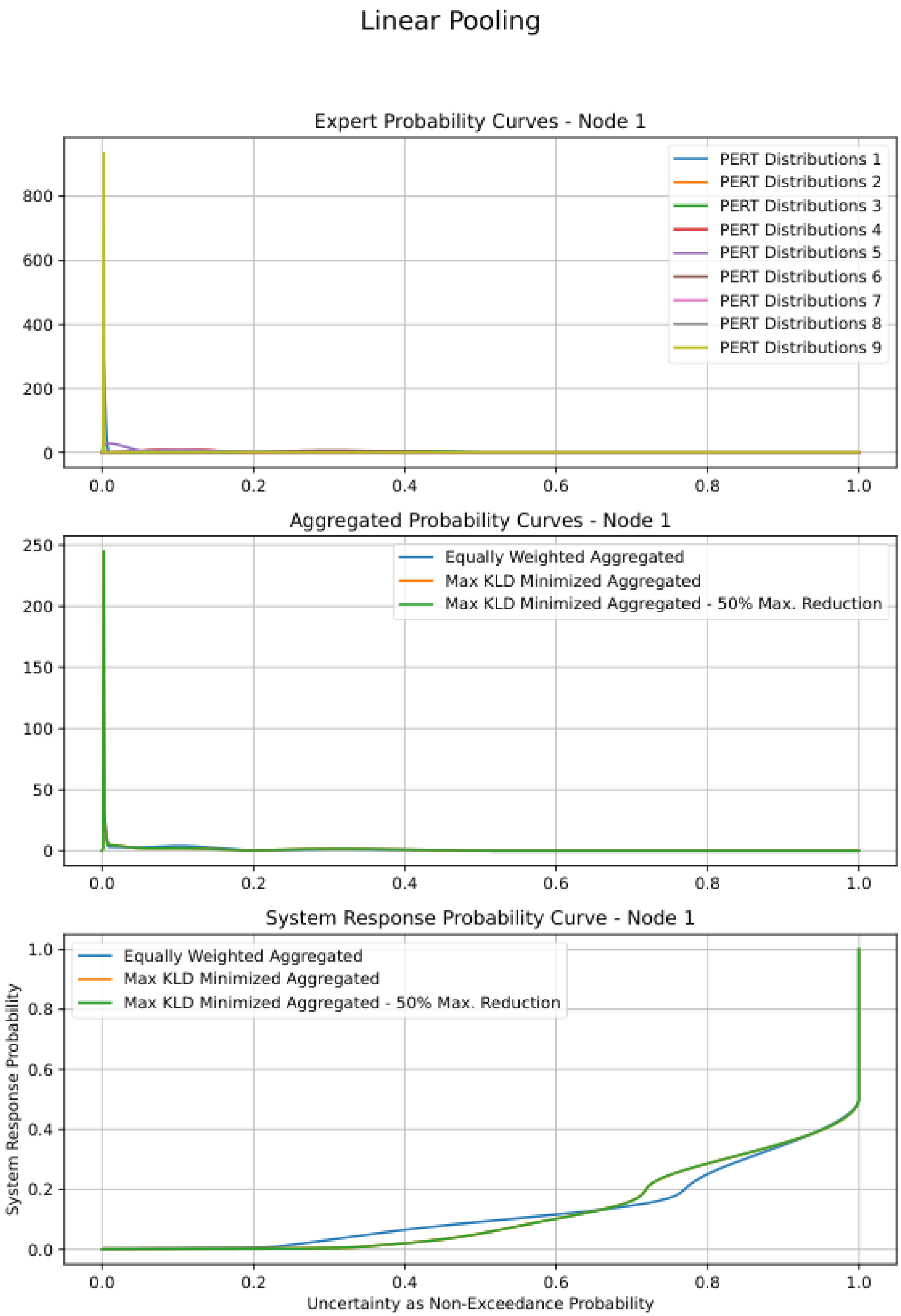

Figure B.25: Linear Pooling (Inverse) - Node 1

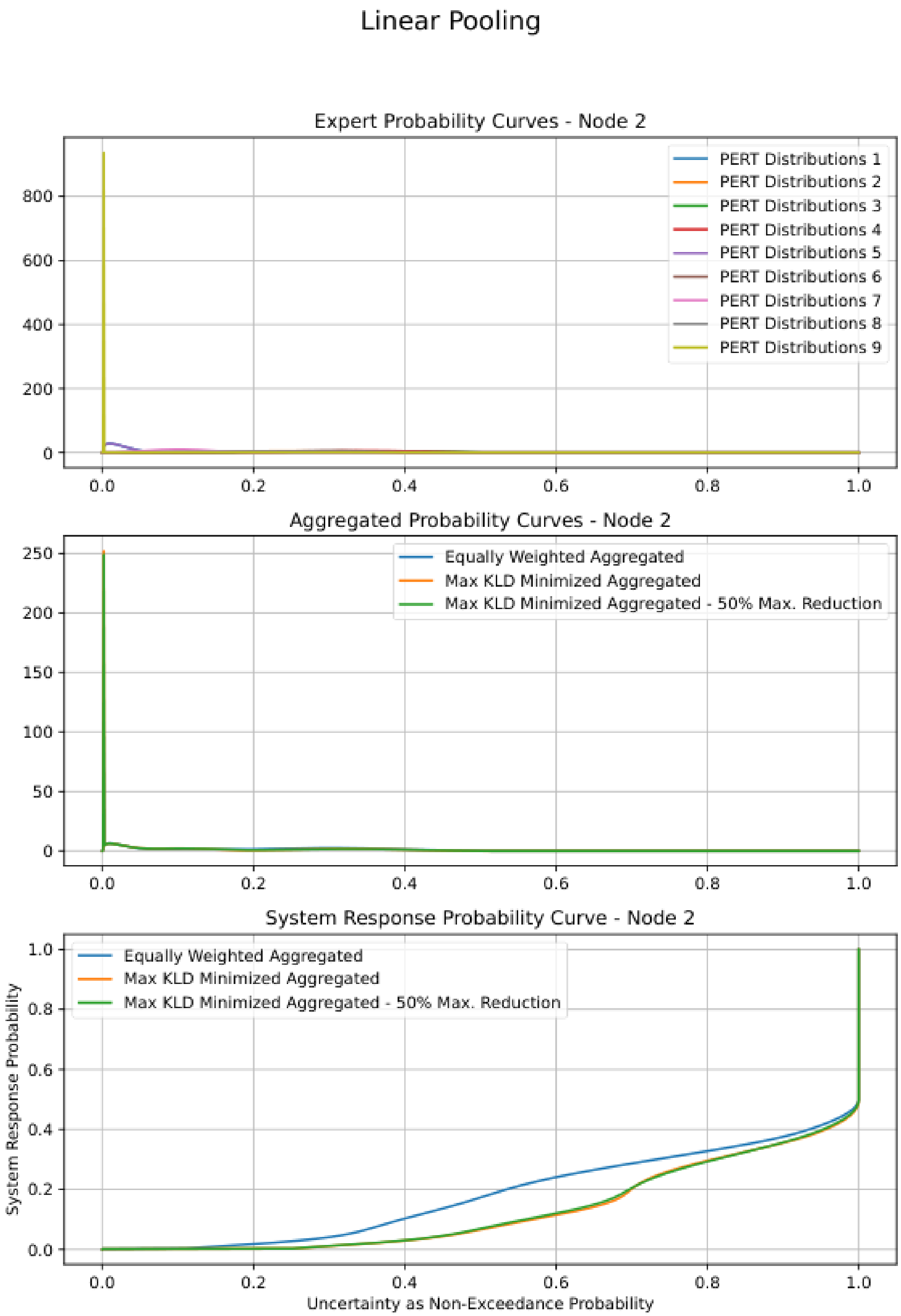

Figure B.26: Linear Pooling (Inverse) - Node 2

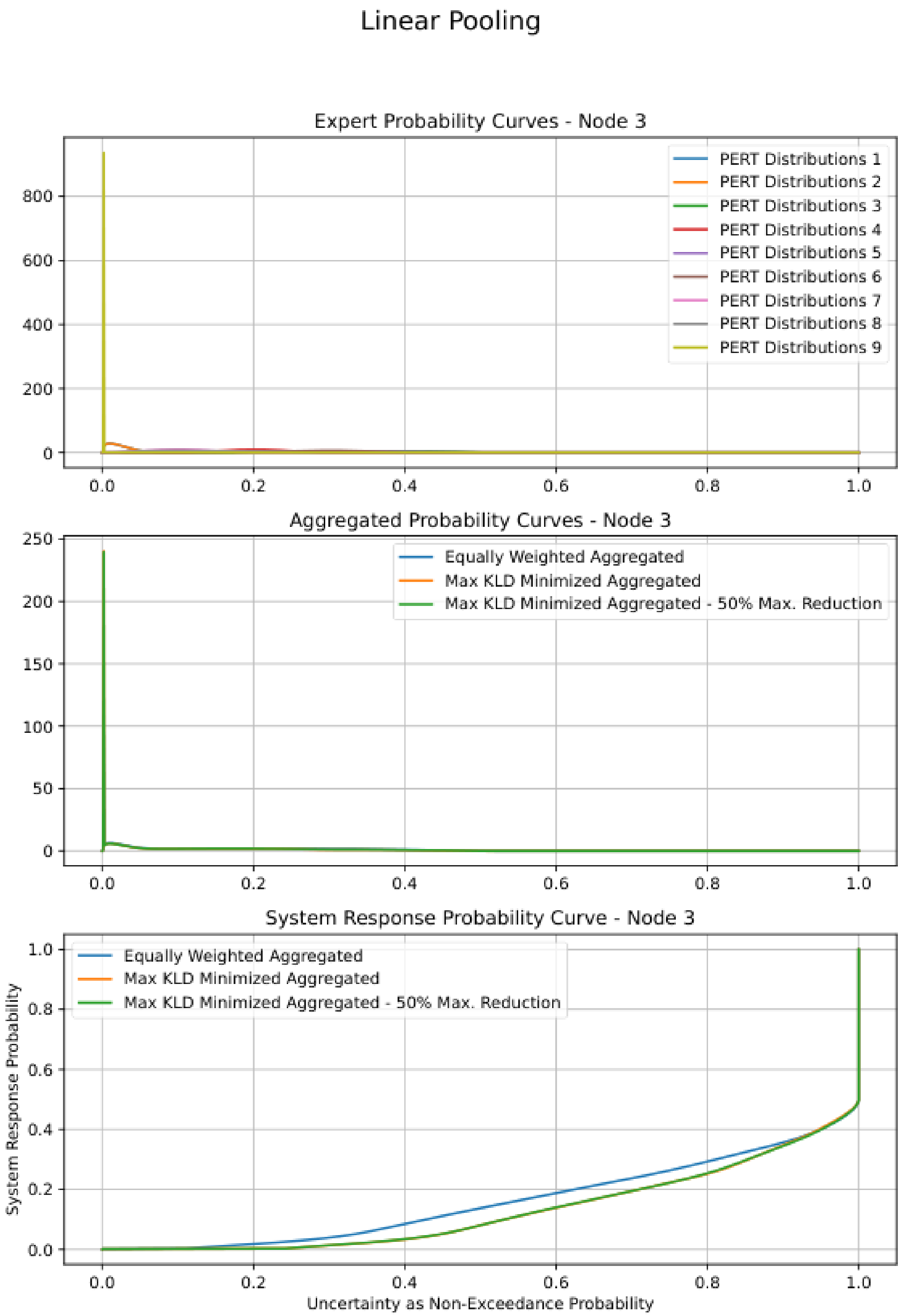

Figure B.27: Linear Pooling (Inverse) - Node 3

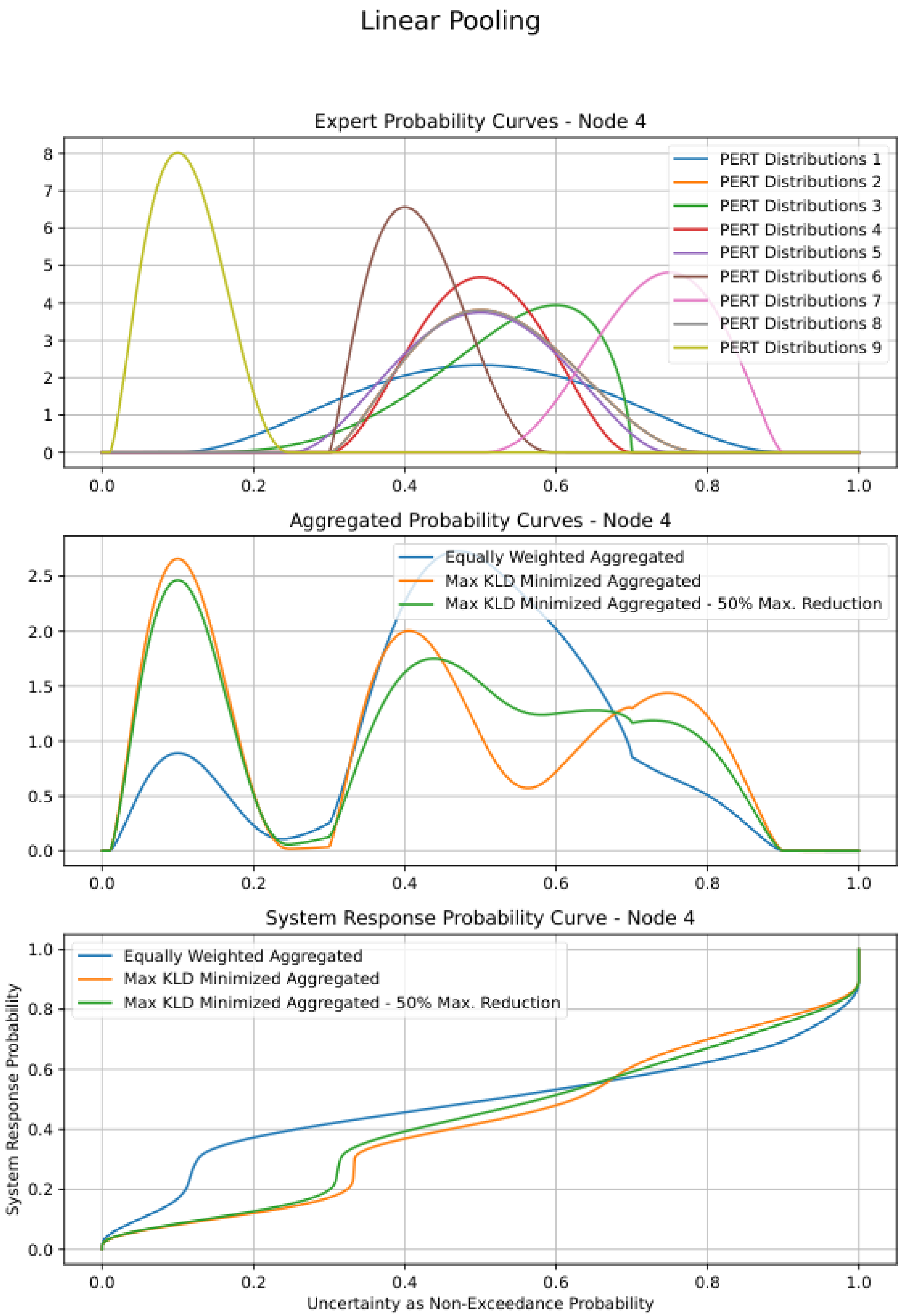

Figure B.28: Linear Pooling (Inverse) - Node 4

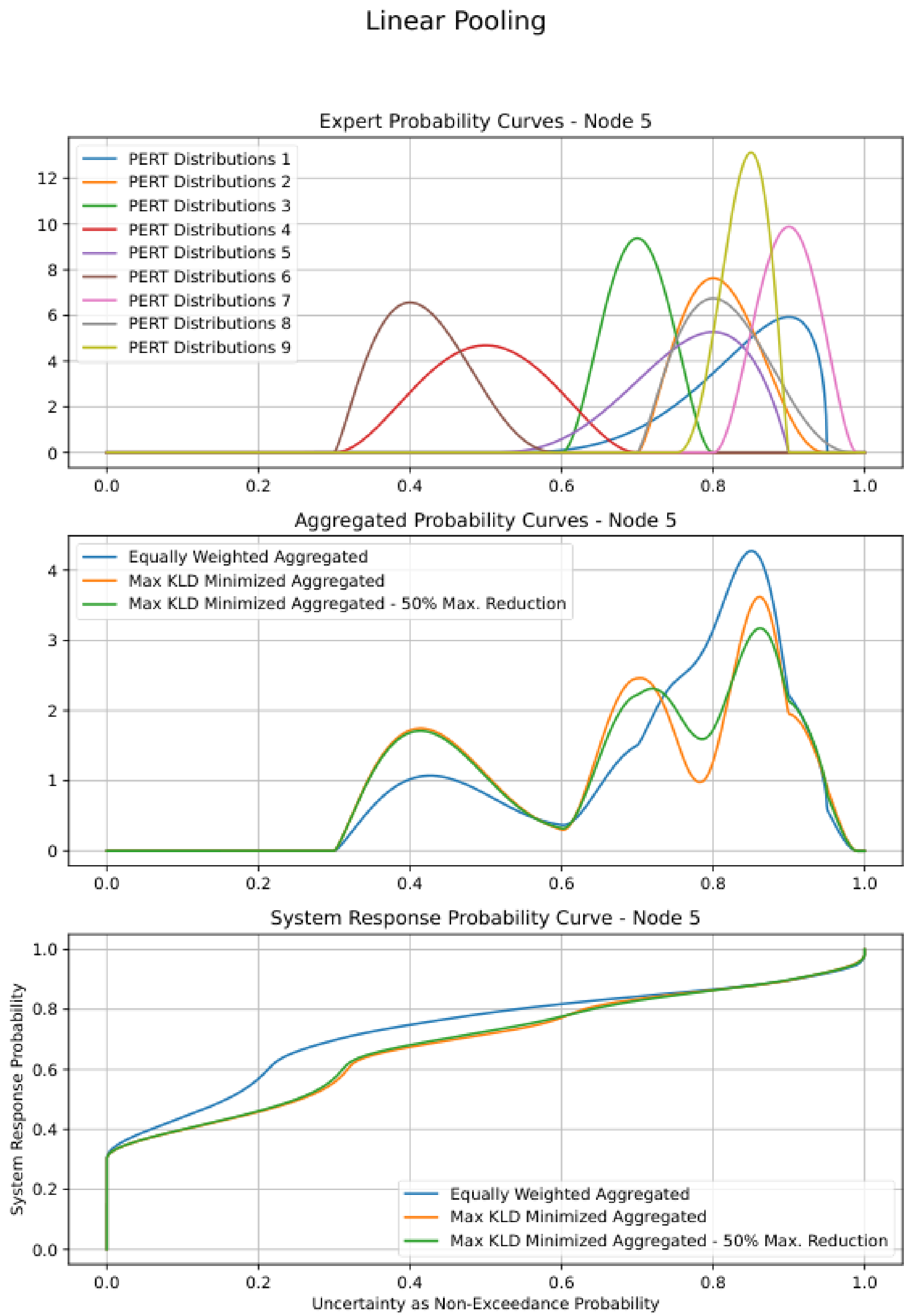

Figure B.29: Linear Pooling (Inverse) - Node 5

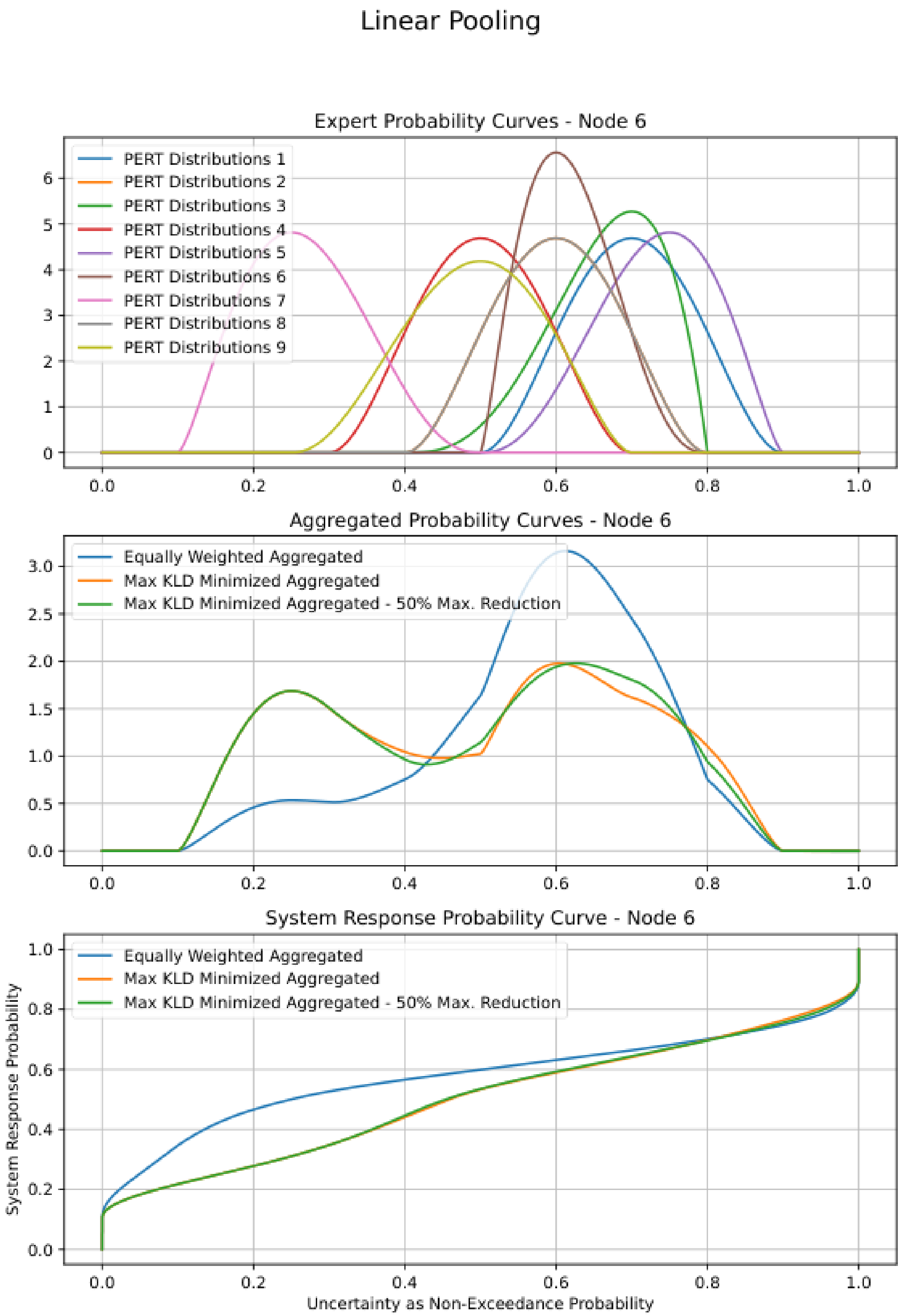

Figure B.30: Linear Pooling (Inverse) - Node 6

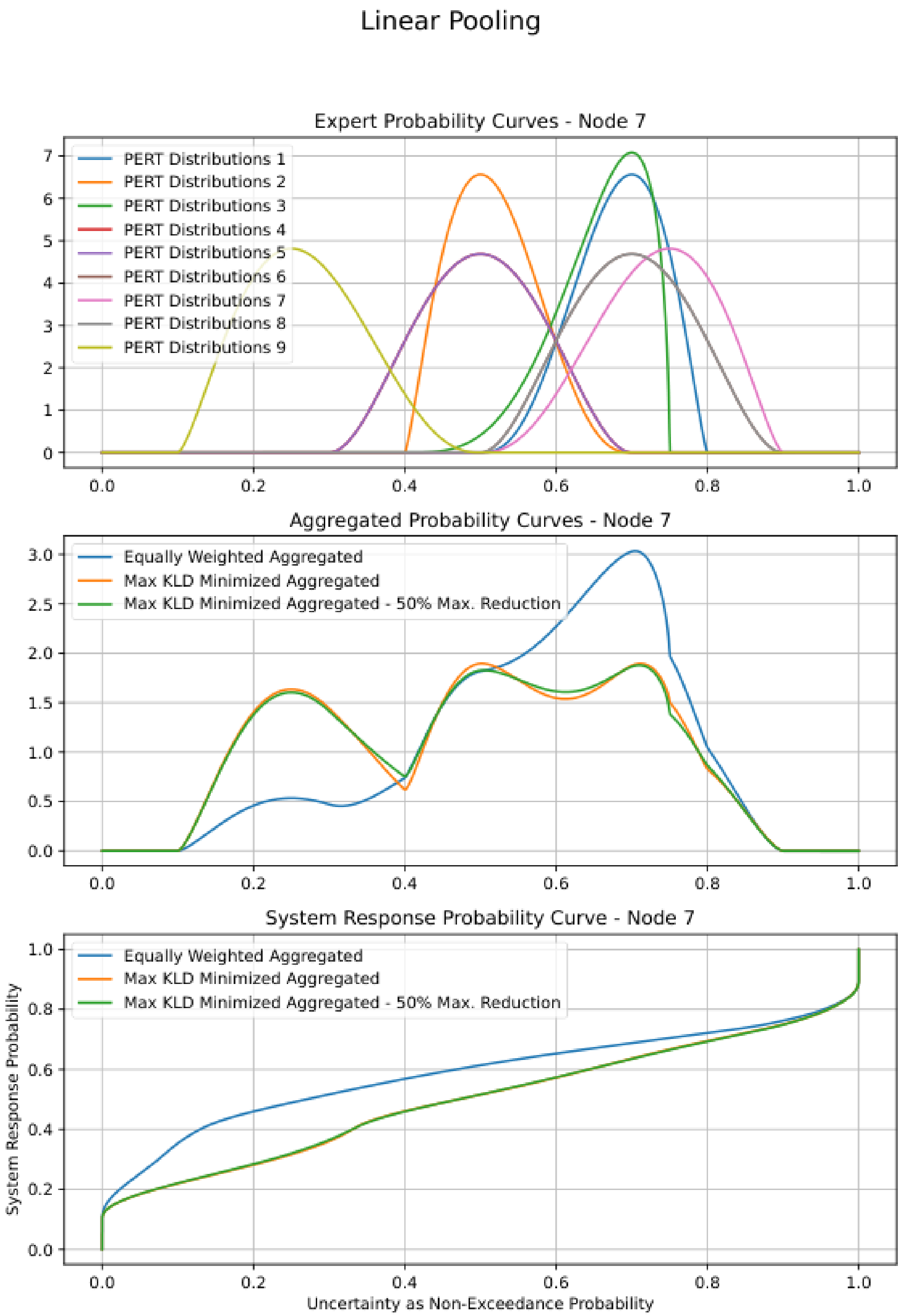

Figure B.31: Linear Pooling (Inverse) - Node 7

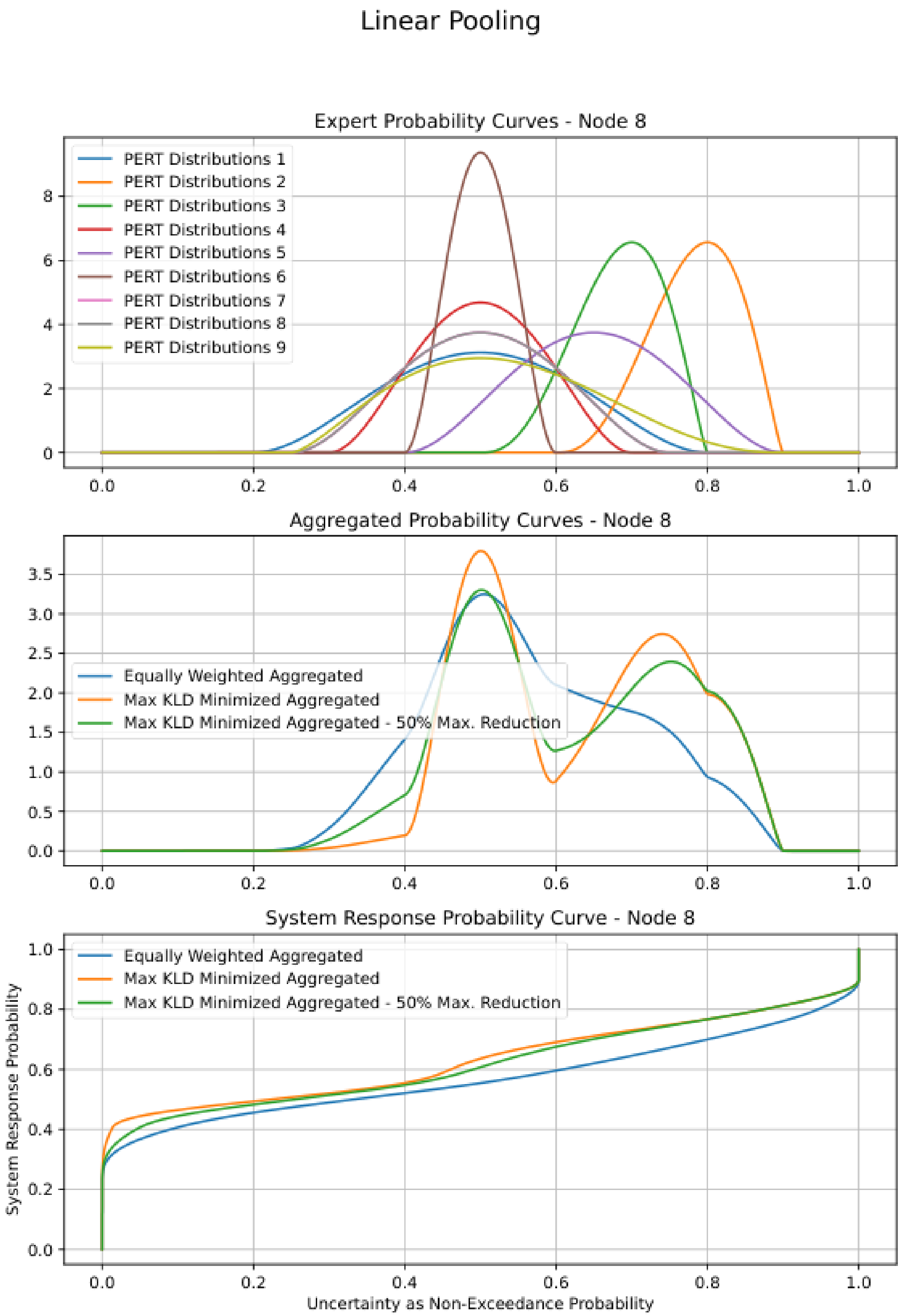

Figure B.32: Linear Pooling (Inverse) - Node 8

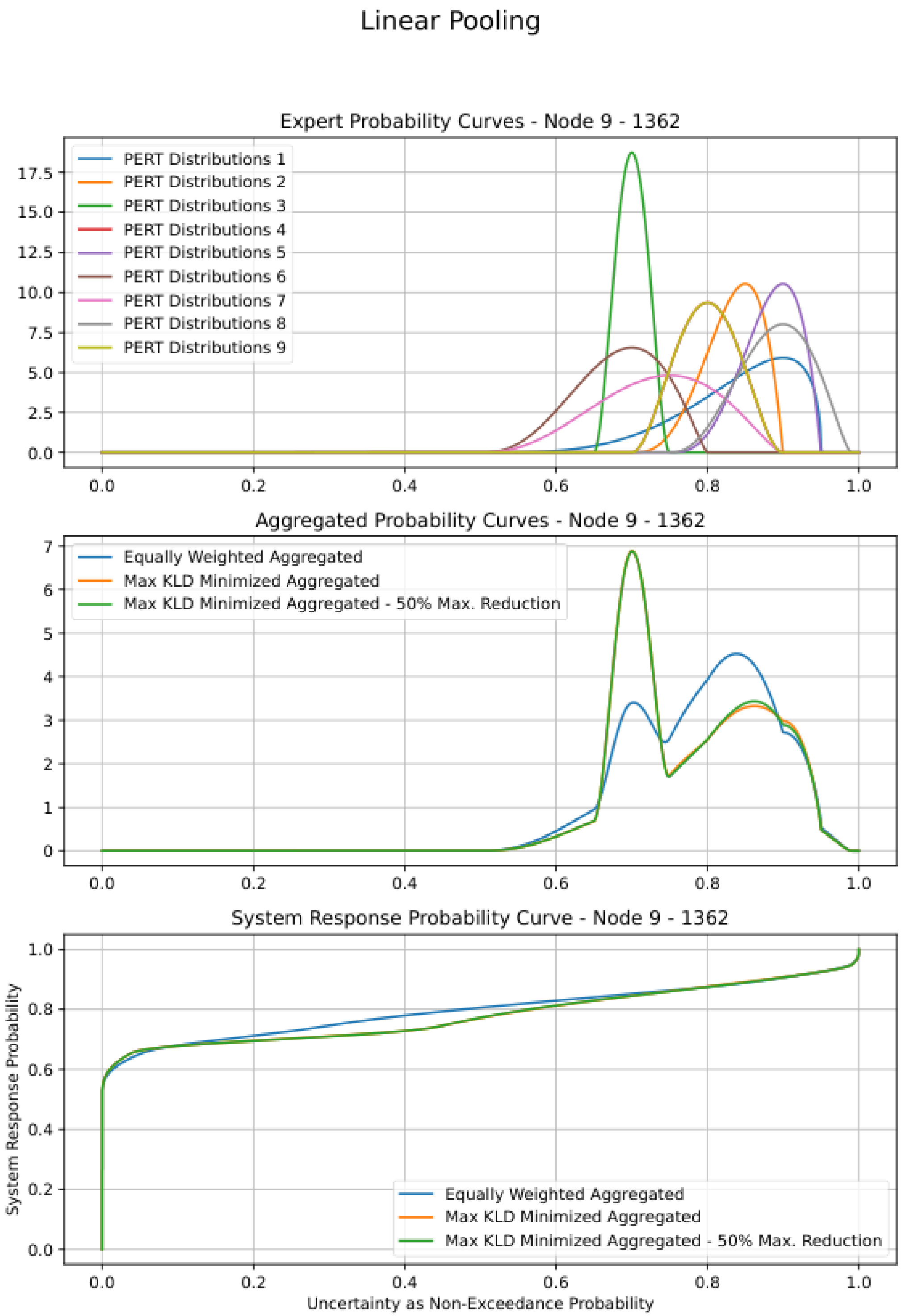

Figure B.33: Linear Pooling (Inverse) - Node 9

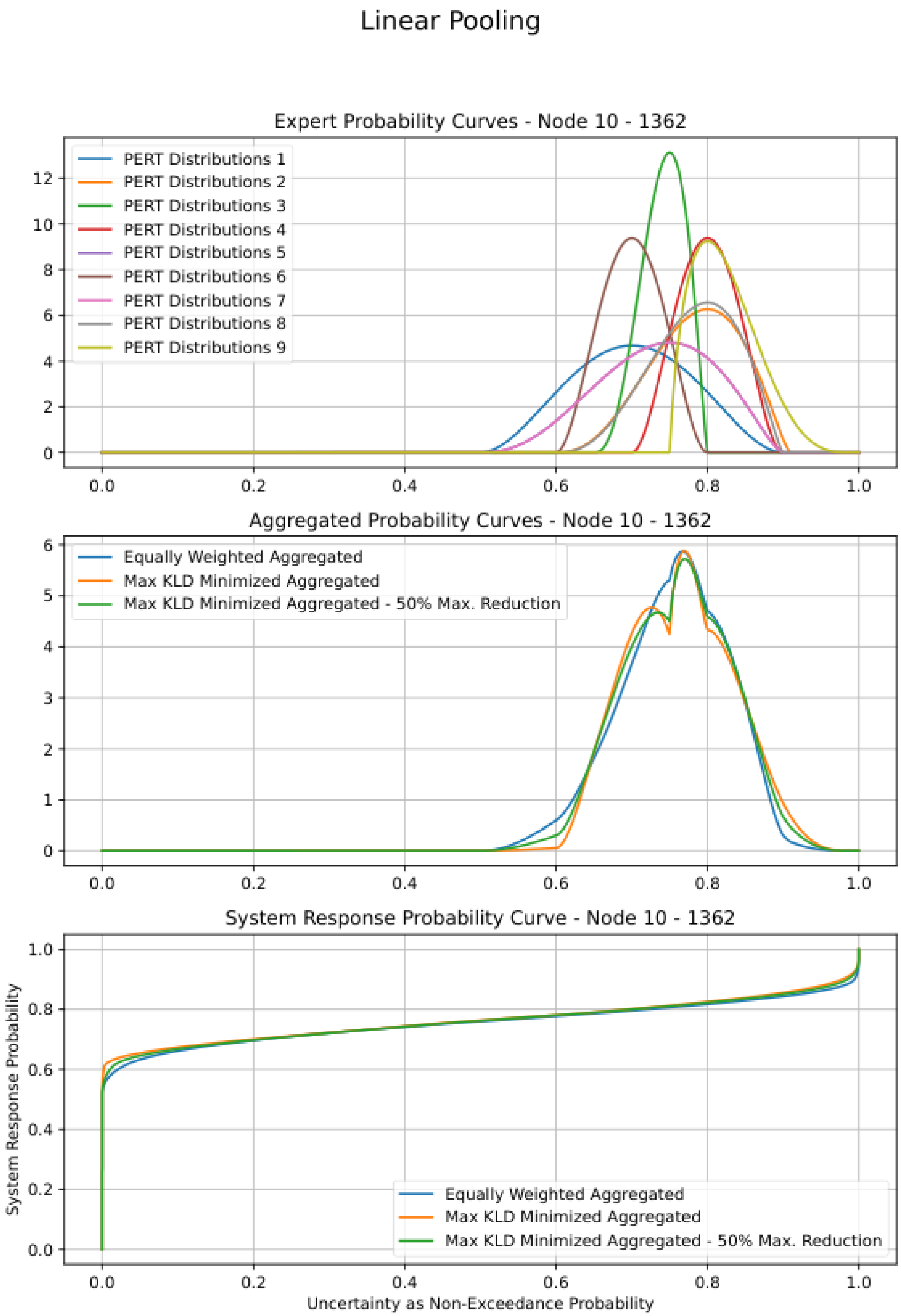

Figure B.34: Linear Pooling (Inverse) - Node 10

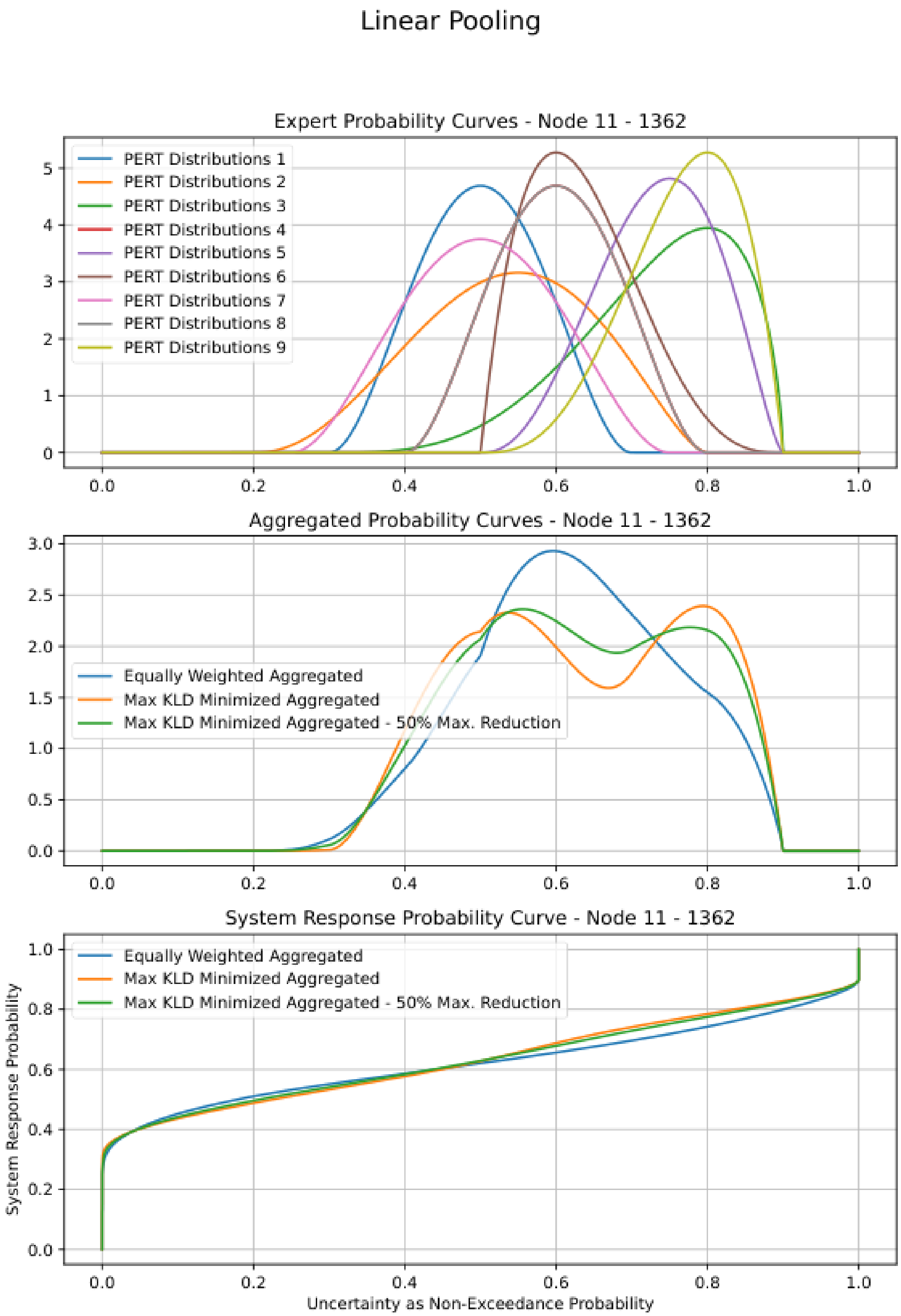

Figure B.35: Linear Pooling (Inverse) - Node 11

Linear Pooling

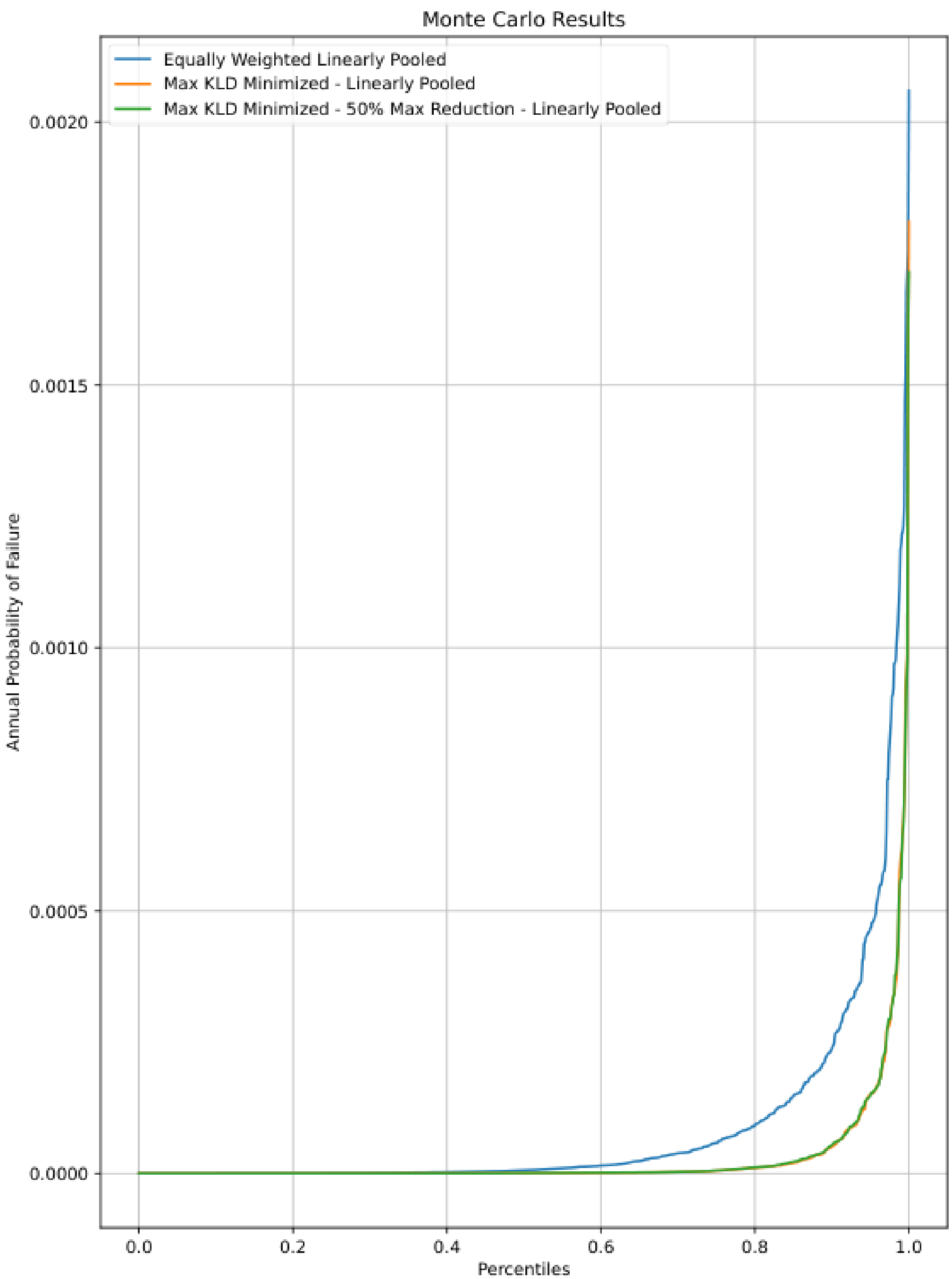

Figure B.36: Linear Pooling (Inverse) - Monte Carlo Results

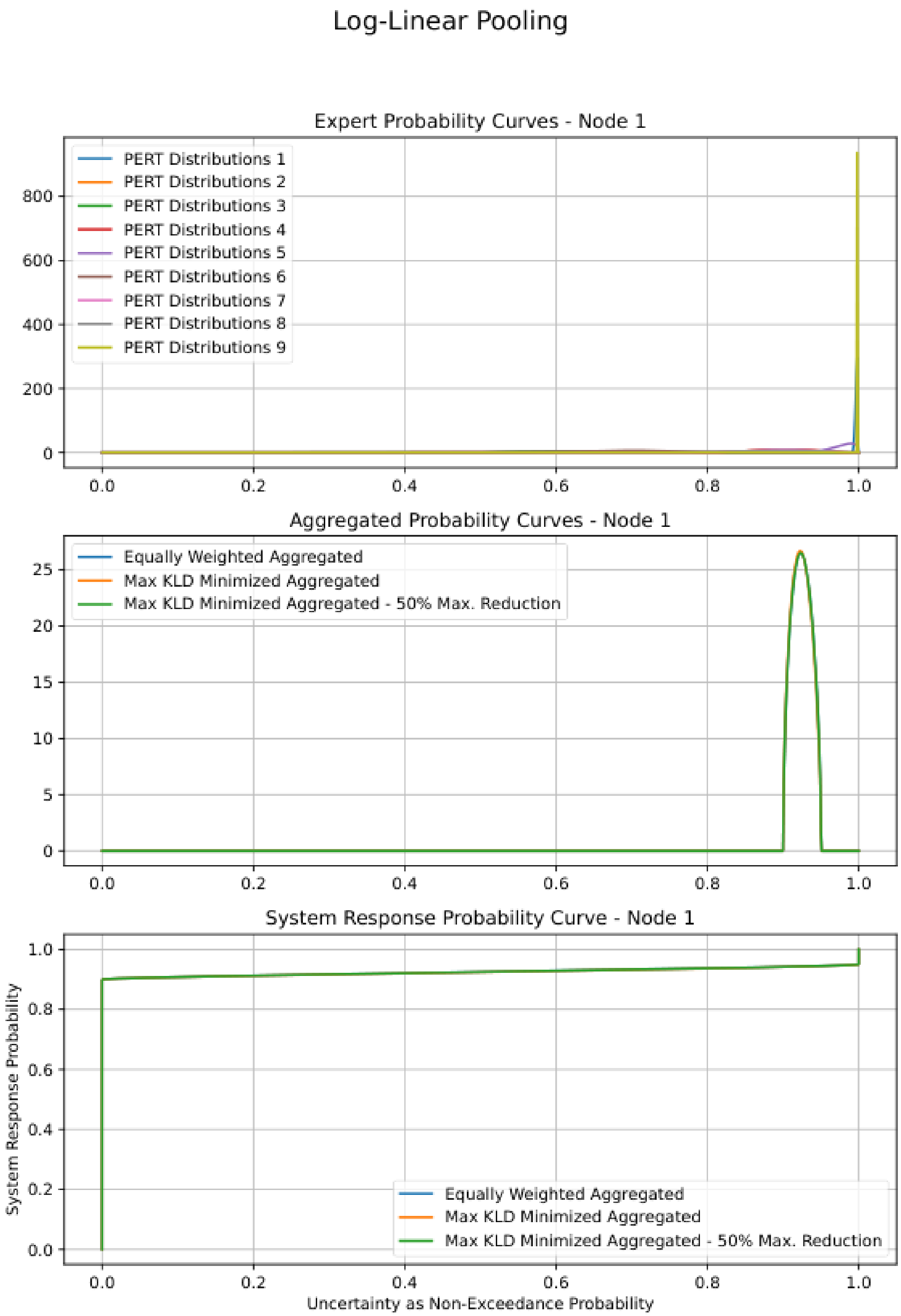

Figure B.37: Log-Linear Pooling - Node 1

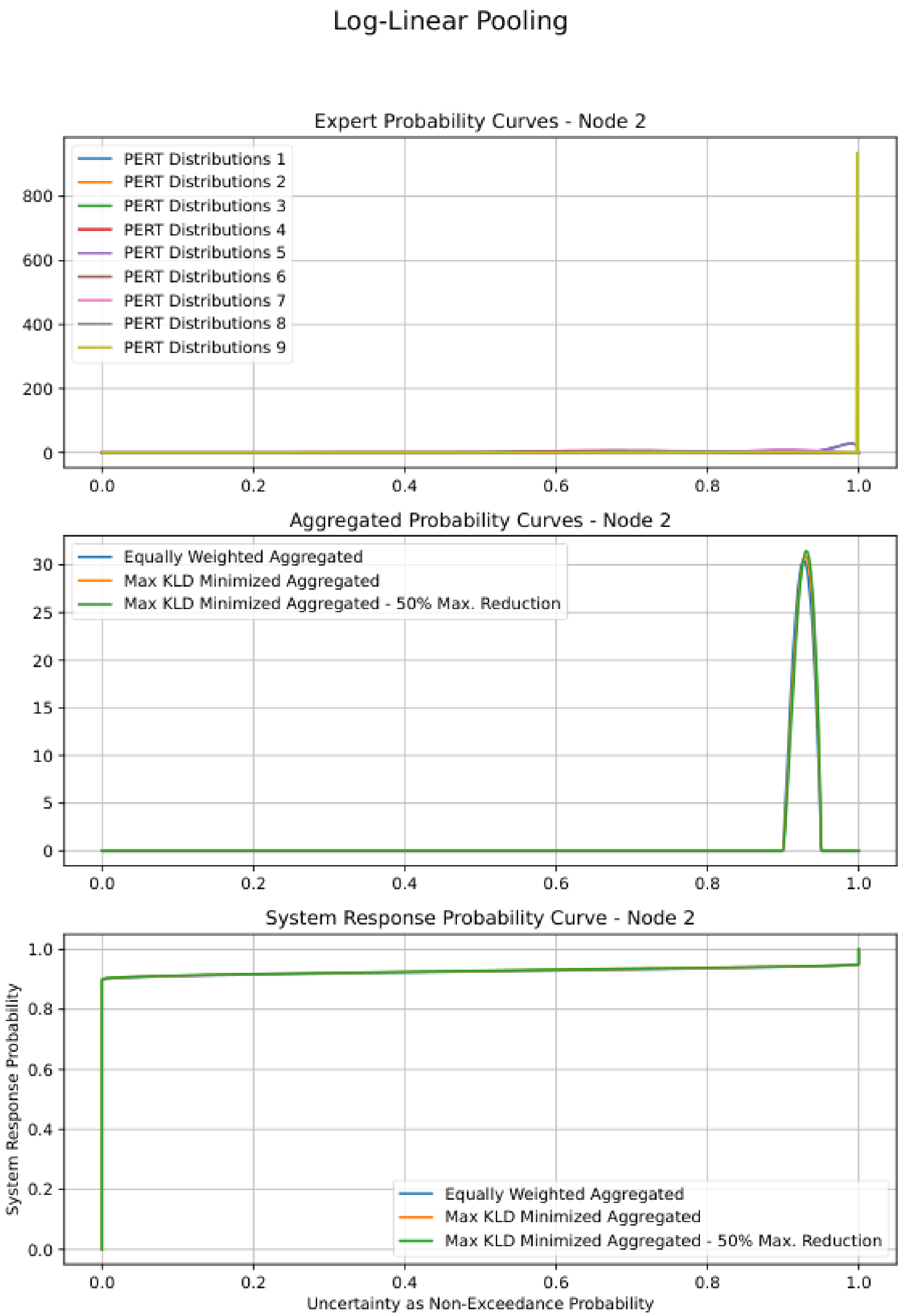

Figure B.38: Log-Linear Pooling - Node 2

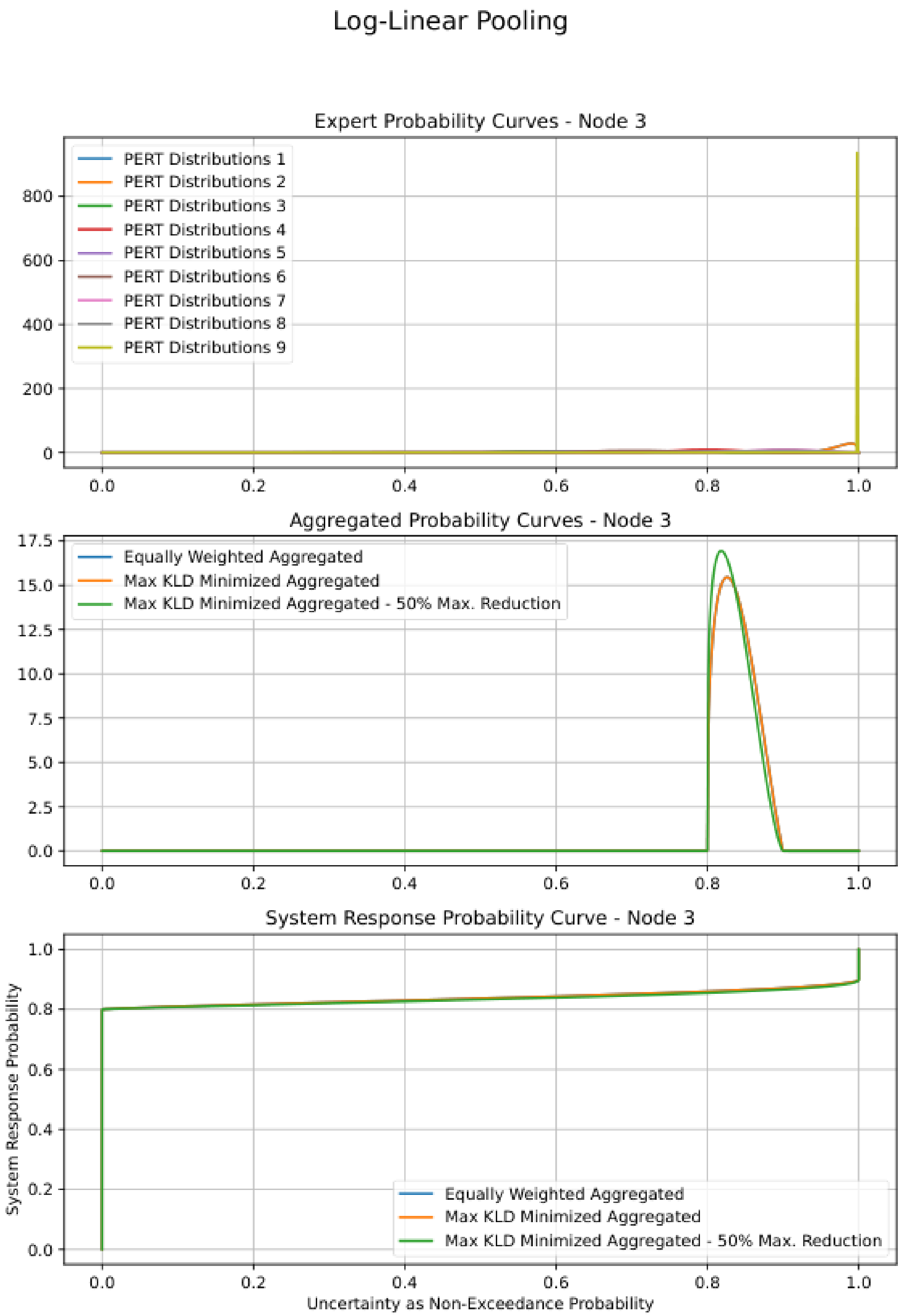

Figure B.39: Log-Linear Pooling - Node 3

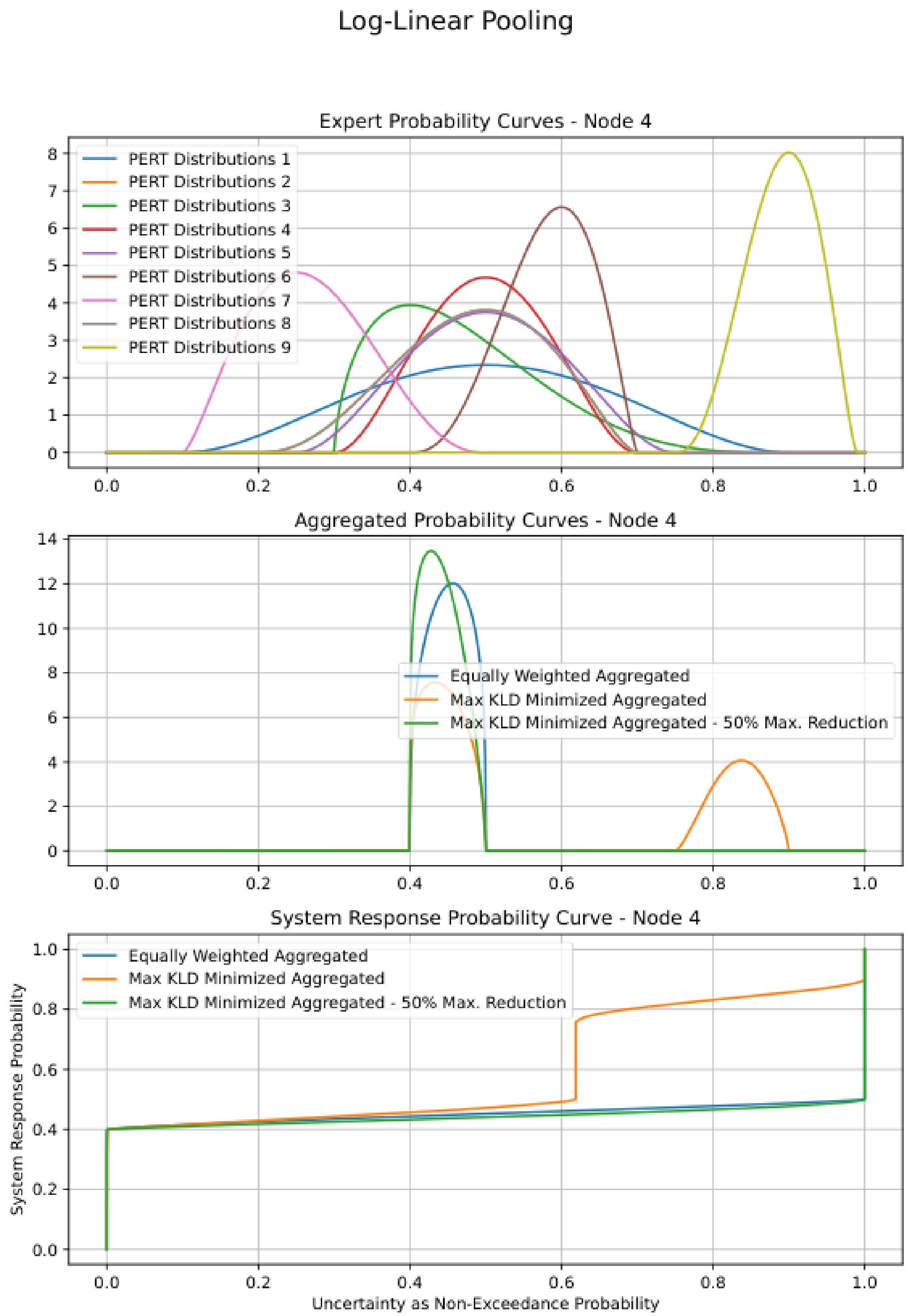

Figure B.40: Log-Linear Pooling - Node 4

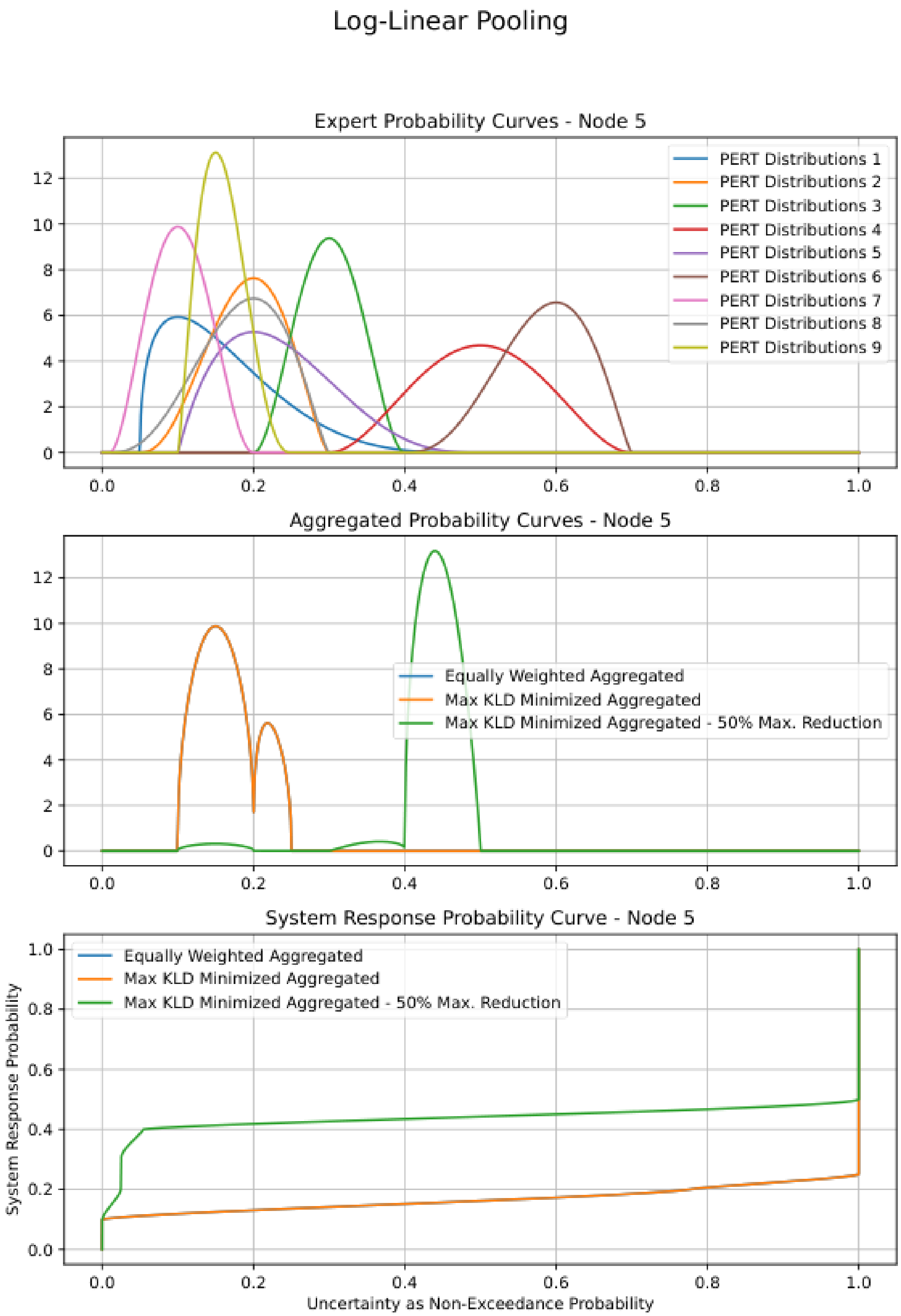

Figure B.41: Log-Linear Pooling - Node 5

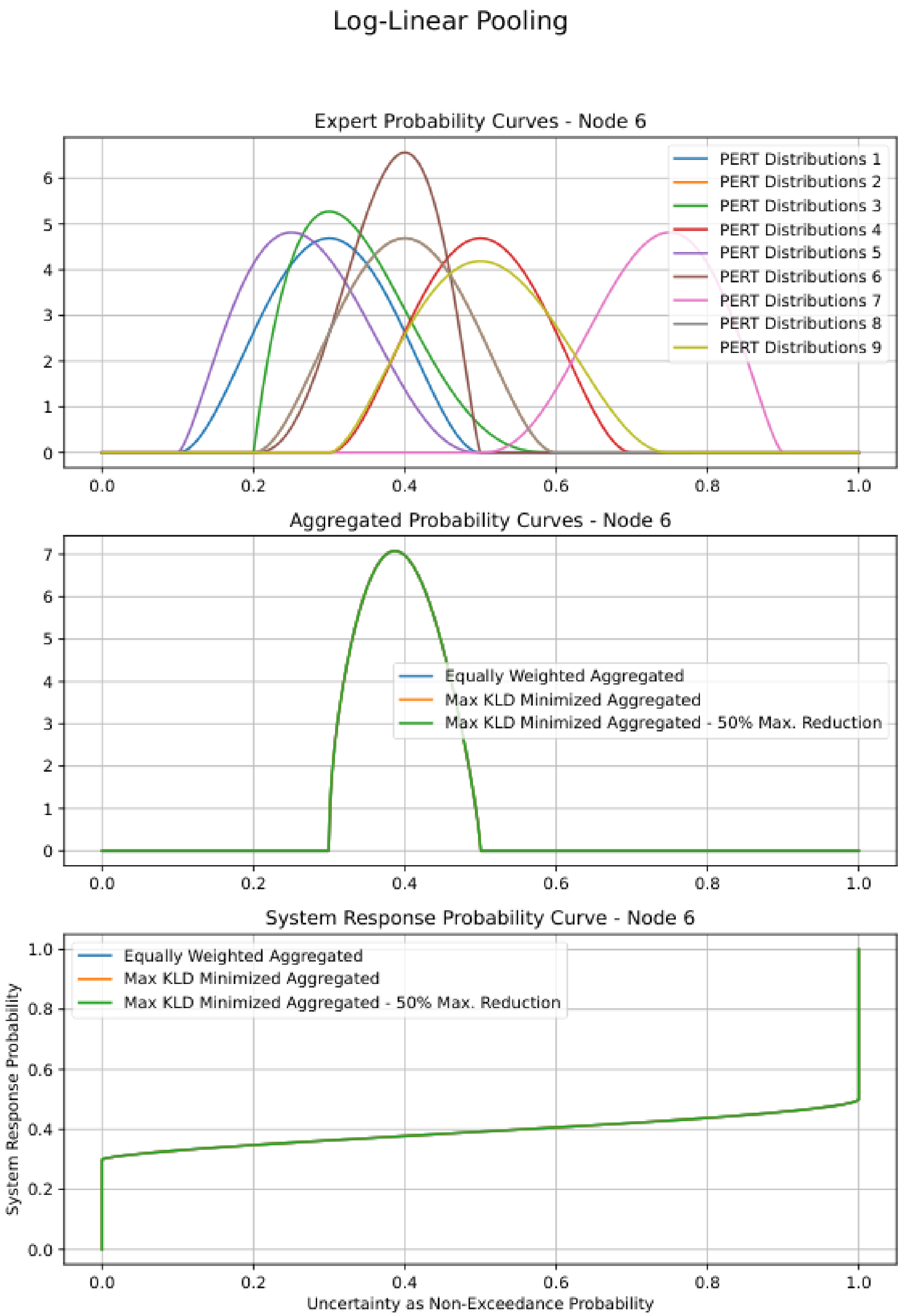

Figure B.42: Log-Linear Pooling - Node 6

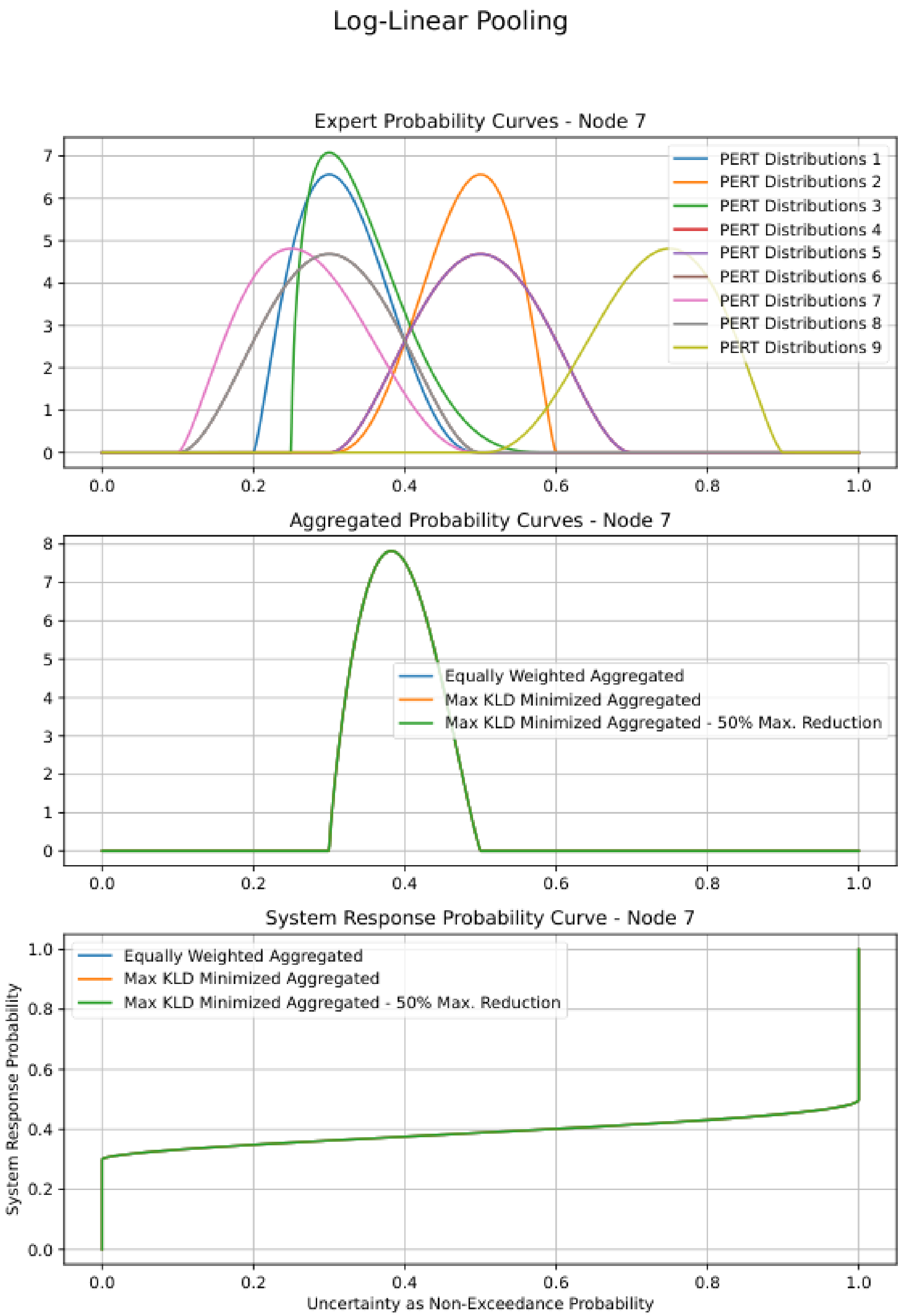

Figure B.43: Log-Linear Pooling - Node 7

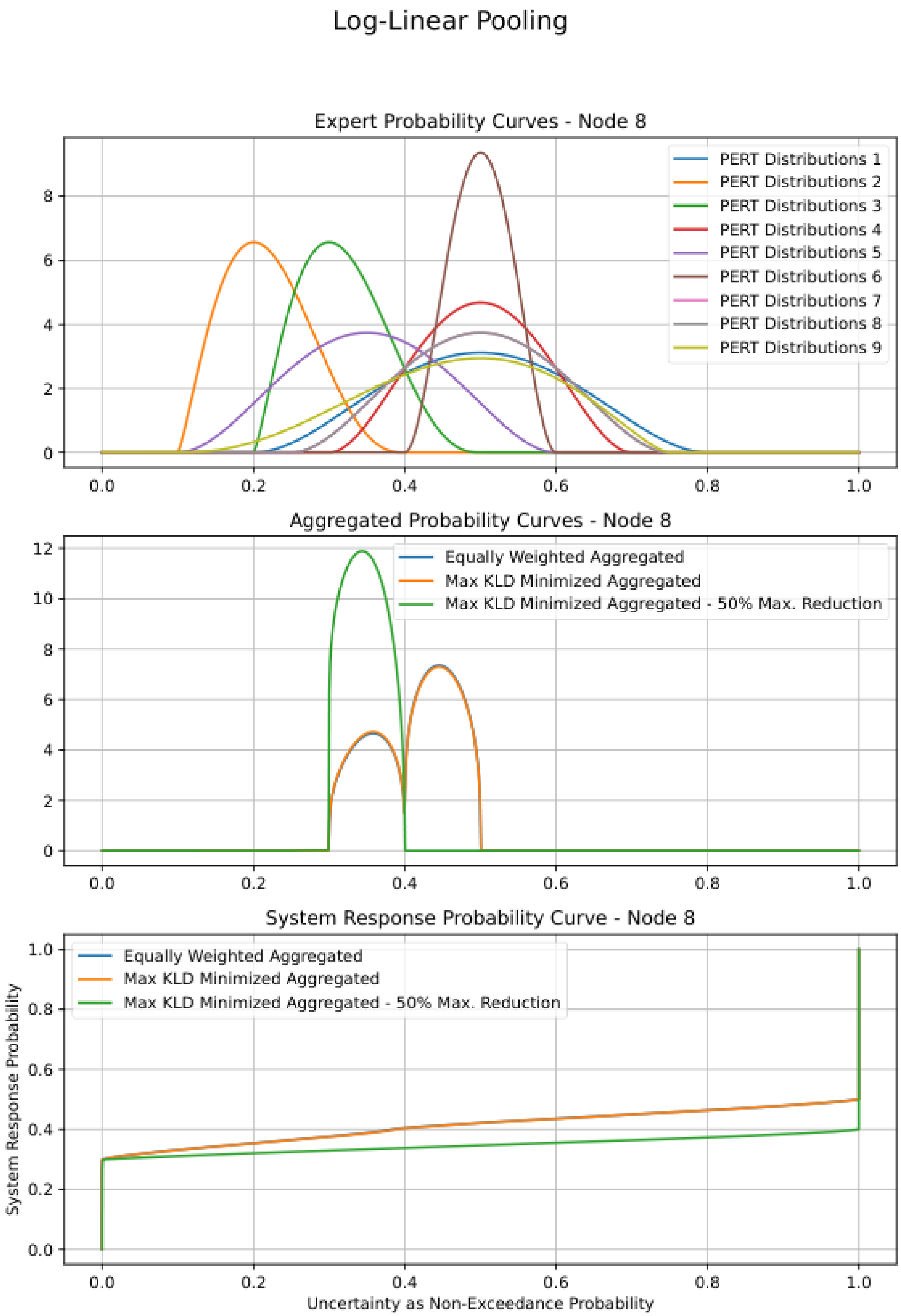

Figure B.44: Log-Linear Pooling - Node 8

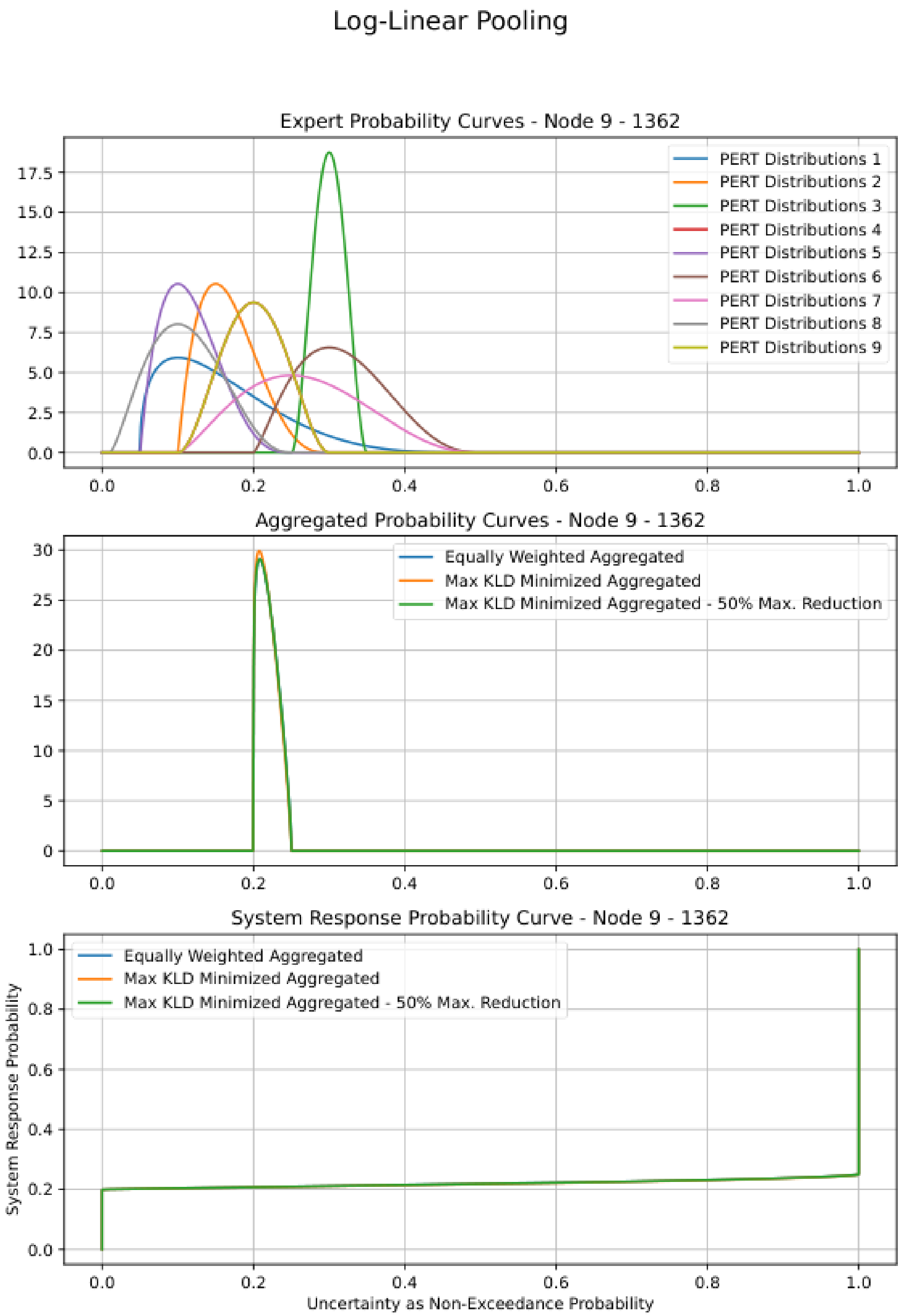

Figure B.45: Log-Linear Pooling - Node 9

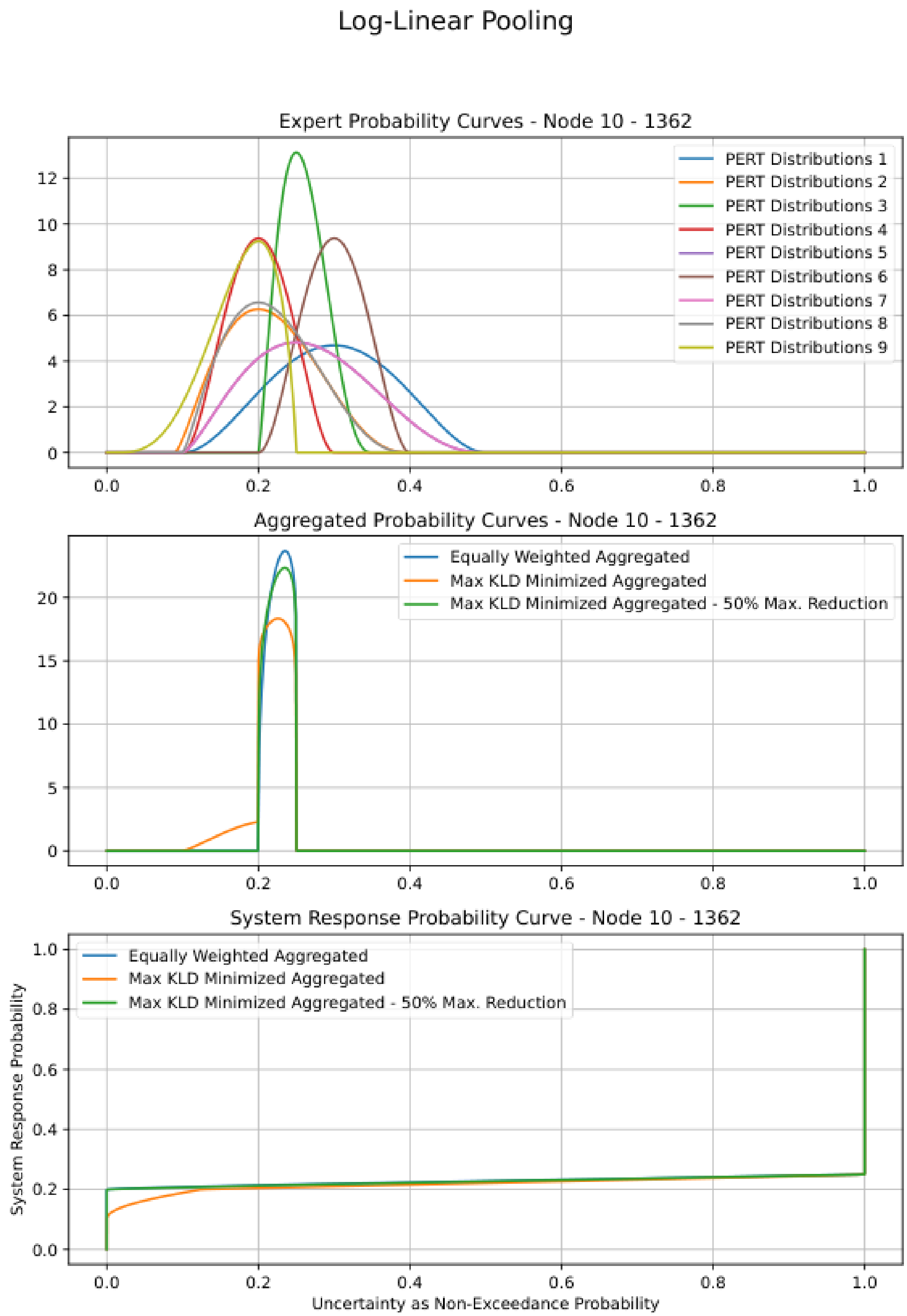

Figure B.46: Log-Linear Pooling - Node 10

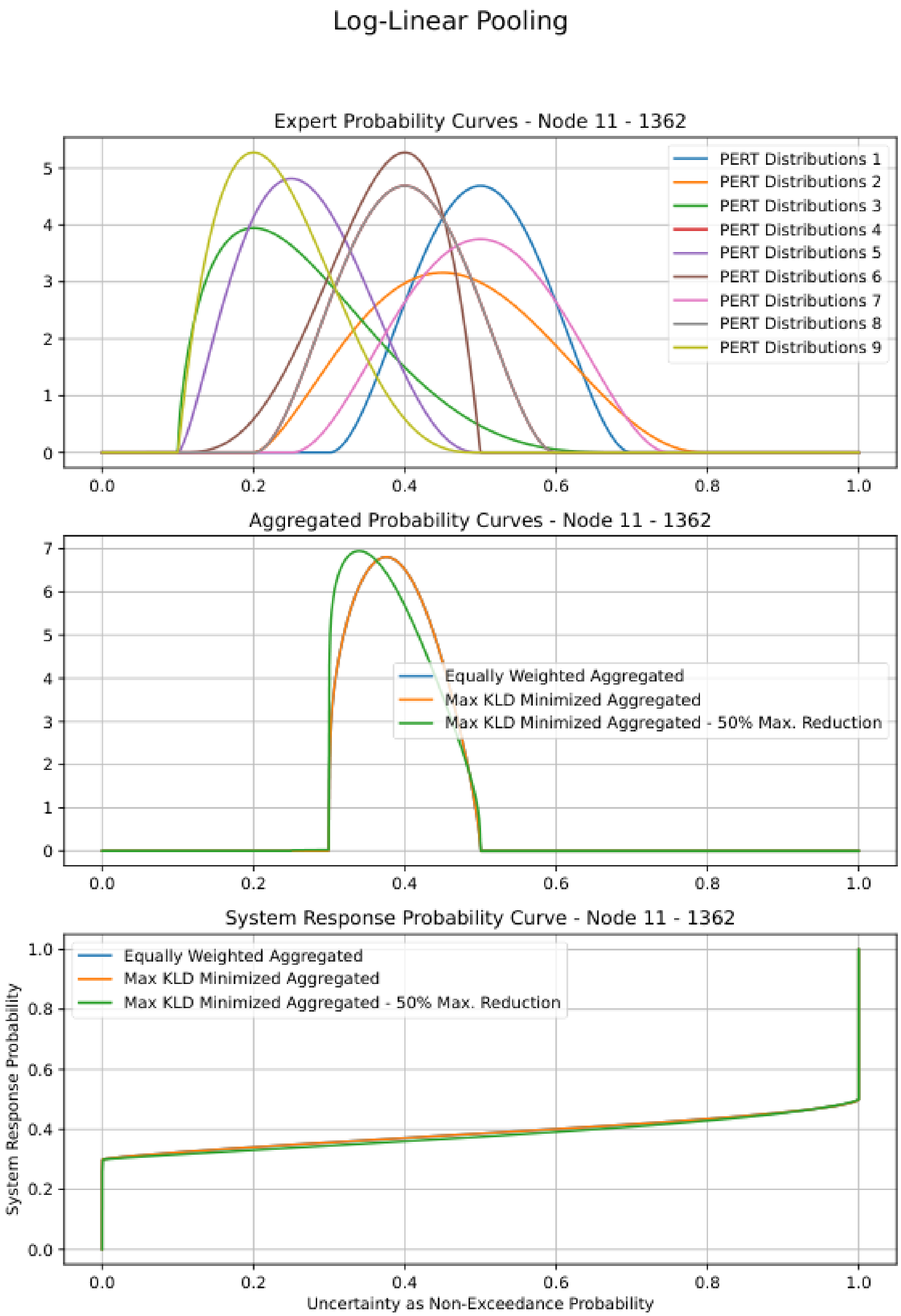

Figure B.47: Log-Linear Pooling - Node 11

Log-Linear Pooling

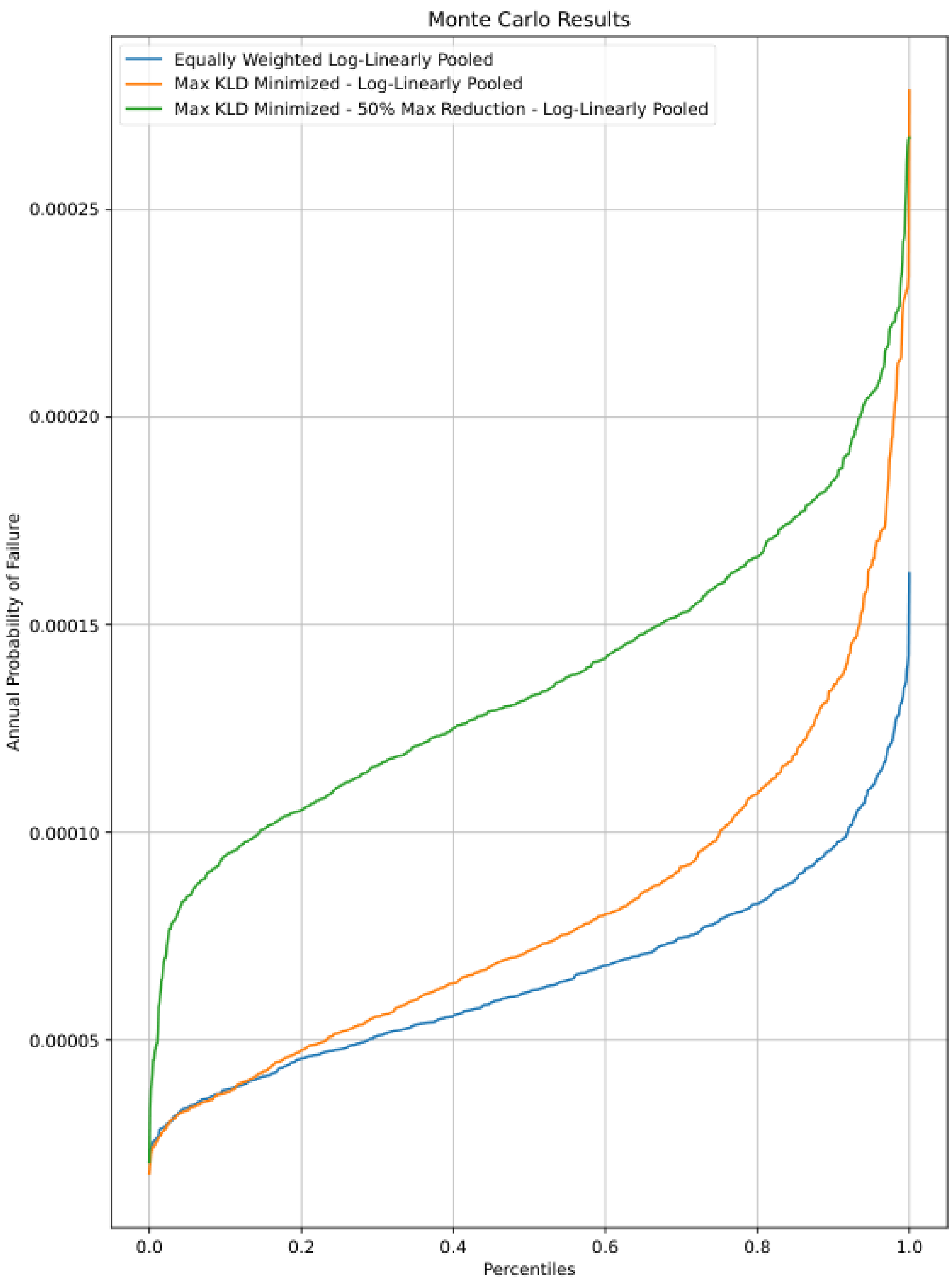

Figure B.48: Log-Linear Pooling - Monte Carlo Results

Vita

Joshua Adam Thomas is a Senior Program manager at the Tennessee Valley Authority and a professional engineer licensed in the state of Tennessee. He holds a Bachelor of Science degree in Civil and Environmental Engineering with a concentration in Structural Engineering from Tennessee Technological University and a Master of Science degree in Engineering Management from the University of Tennessee, Chattanooga. He is also a PhD student in the Industrial and Systems Engineering department at the University of Tennessee, Knoxville.