Evidence-Based Expert Judgment in Flood Risk

Rongen, G.W.F.

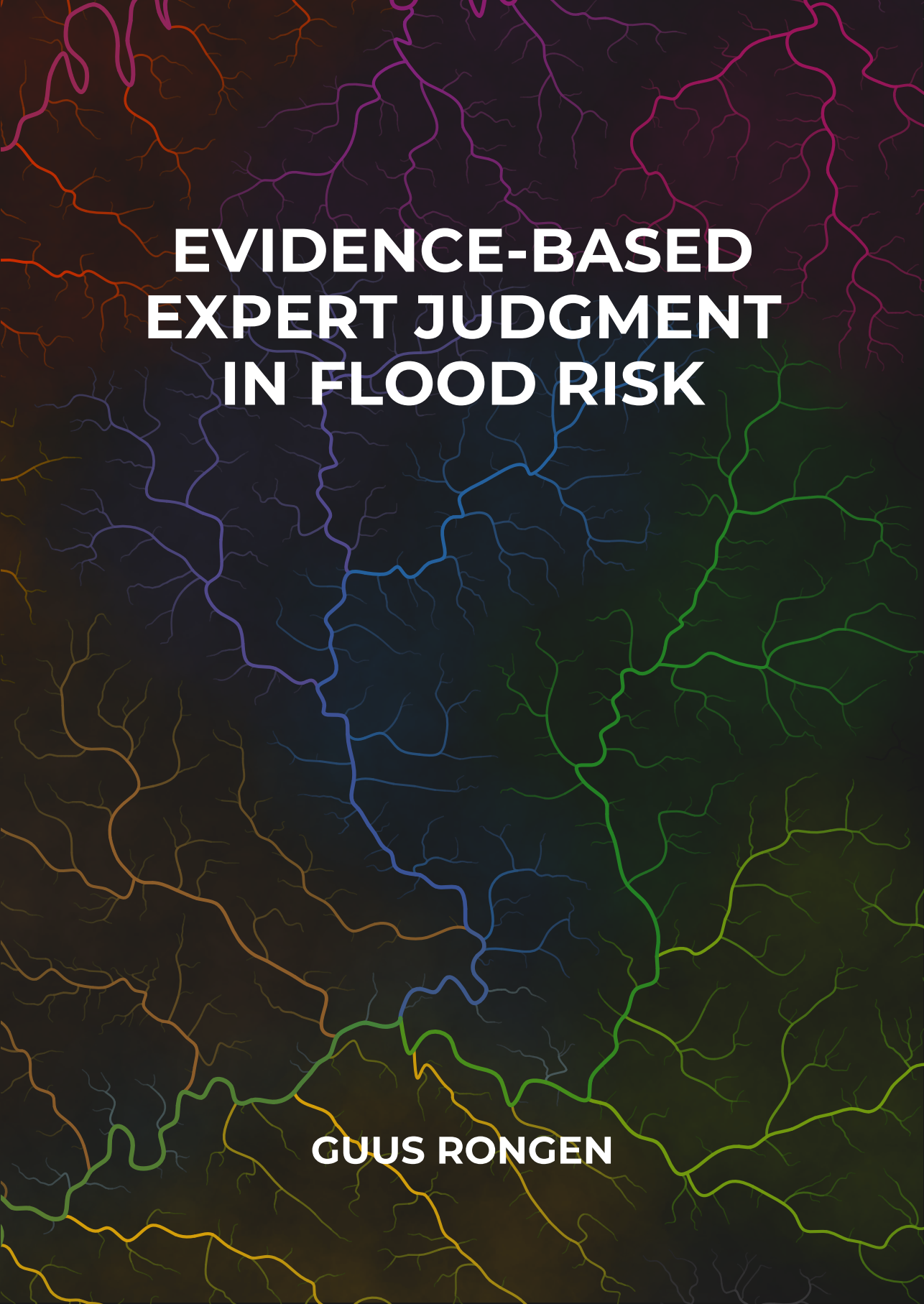
DOI
10.4233/uuid:8babc3a3-cbbb-4b01-a9c4-07e07ad03fd6

Publication date
2024

Document Version
Final published version

Citation (APA)
Rongen, G. W. F. (2024). *Evidence-Based Expert Judgment in Flood Risk*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:8babc3a3-cbbb-4b01-a9c4-07e07ad03fd6

EVIDENCE-BASED EXPERT JUDGMENT IN FLOOD RISK



GUUS RONGEN

Evidence-Based Expert Judgment in Flood Risk

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 14 november 2024 om 10:00 uur

door

Guus Willem Franciscus RONGEN

civiel ingenieur, Technische Universiteit Delft, Nederland,
geboren te Geleen, Nederland.

Dit proefschrift is goedgekeurd door de

promotor: dr. ir. O. Morales-Nápoles
promotor: prof. dr. ir. M. Kok

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Dr. ir. O. Morales-Nápoles	Technische Universiteit Delft
Prof. dr. ir. M. Kok	Technische Universiteit Delft
Onafhankelijke leden:	
Prof. dr. ir. S.N. Jonkman	Technische Universiteit Delft
Prof. dr. ir. P.H.A.J.M. van Gelder	Technische Universiteit Delft
Prof. dr. T. Bedford	University of Strathclyde
Ir. R.E. Jorissen	Rijkswaterstaat
Prof. dr. ir. M. Bakker	Technische Universiteit Delft, reservelid
Overige leden:	
Dr. ir. G.F. Nane	Technische Universiteit Delft

This work is funded by the TKI project EMU-FD.

Keywords: structured expert judgment, probabilistic modeling, flood risk, hydrology, Bayesian statistics, dependence modeling

Printed by: Gildeprint – The Netherlands

ISBN 978-94-6384-663-9

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Every belief that we hold is a leap of faith,
but not all leaps of faith are created equal.

Stephen West

Contents

Summary

Flood risk assessments heavily rely on physics-based models. These models contain significant uncertainties that need to be quantified to adequately assess failure probabilities of flood defenses. Assumptions about these uncertainties remain often hidden within the modeling process, even though they are critical to the outcomes. Structured expert judgment offers a method to explicitly quantify these uncertainties with probabilities. One specific method for this that is used throughout this thesis is the Classical Model, also known as Cooke's Method. It uses performance-based weights to combine estimates from a group of experts into a single estimate. This method is explicitly focused on estimating uncertainty. It recognizes that unknown variables are inherently uncertain, and consequently scores experts based on their proficiency in estimating uncertainty.

Flood risk, especially in the Dutch context, involves events with small probabilities (e.g., 10^{-3} to 10^{-5} every year) but large consequences. The rarity of these events makes it difficult to quantify them on an empirical basis. Their recurrence intervals typically exceed human lifespans and the length of most observational records. The severe consequences, however, make it crucial to get a grasp on the extremity of such events as best as possible. Additionally, the dependence between events or contributing factors must be considered as well. Statistical dependence has a prominent role in hydrological extremes leading to flooding, whether they relate to hydraulic conditions such as water levels and waves, spatial aspects like rain across different catchment, or temporal factors such as consecutive storms.

The objective of this thesis is to explore how expert judgment can contribute to estimation of more credible failure probabilities for engineered flood defense systems. This involves obtaining more credible estimates of the frequency of hydrological extremes such as floods, and subsequently assessing how this affects dike failure probability, while considering that these are rare, non-experienced, events involving probabilistic dependence. This exploration is structured around four research questions concerning 1) the measures and distributions used to evaluate expert performance in the Classical Model, 2) the types of uncertainties that are best estimated by experts, 3) the application of expert judgment as priors in a Bayesian framework to reduce uncertainty in hydrological extremes, and 4) the experts' capability to estimate statistical dependence.

The first research question evaluates the performance of various statistical tests and distributions within the Classical Model for structured expert judgment. The research evaluates two distributions to represent expert estimates, under the hypothesis that a smooth, Metalog, distribution would better represent these than the typically used piecewise uniform distribution. However, the findings indicate that neither distribution adequately represents expert estimates. Thus, for a more accurate representation of the dis-

tributions that expert have in mind, it is recommended to elicit more percentiles from the experts. Furthermore, the research investigates alternative measures of statistical accuracy to the chi-square test that is traditionally used to compare an expert's estimates to known outcomes (realizations). The chi-square test generates a p-value that for a large part determines the weight of an expert's estimates. Four alternatives, the Kolmogorov-Smirnov, Cramer-von Mises, Anderson Darling tests, and the Continuously Ranked Probability Score, are explored for their suitability for evaluating expert performance. This shows that none of these tests significantly outperforms all others, and that each test displays different biases and sensitivities. The most notable difference between tests is the sensitivity to overconfidence, which is the tendency of experts to give too precise answers. This is the most predominant bias in structured expert judgment studies. Tests that are less sensitive to overconfidence assign higher weights to a larger part of the expert pool.

The second research question investigates the accuracy with which experts estimate uncertainties for different types of variables, specifically within dike failure assessments along the Dutch Rhine River. This analysis compares estimates for system-level failure probability (e.g., at which discharge do you think at least one dike along the river fails?) to detailed dike section failure probability (e.g., at what water level does this specific section fail due to piping?) The findings show that experts provide more credible estimates for system-level failure probabilities than for detailed dike sections. For the latter, the experts that scored best according to the Classical Model (i.e., gave a better representation of uncertainty) did not have the technical knowledge to produce accurate failure probabilities for all dike failure mechanisms. Scaling these single dike section results to a system level failure probability yielded non-credible probabilities, showing that estimates for system-level failure probabilities provide more accurate results. While this does not disqualify structured expert judgment for detailed dike safety assessment, it does show the limitations of using the Classical Model where failure probabilities are highly sensitive to small variations in tail probabilities of estimated dike parameters.

To address the third question, the thesis describes the use of structured expert judgment to reduce uncertainty in hydrological extremes. When a statistical distribution is fitted to such discharges based on only measurements, the uncertainty in the tail is typically very large. To reduce this, hydrologists estimated rare and extreme discharges for the river Meuse's tributaries. Without knowledge of observed historical discharges, we consider the experts' estimates of the extremes to be prior information. By evaluating the experts' performance for more frequently occurring (and therefore more verifiable) discharges, performance-based weights are derived to combine the experts' estimates. This estimate is then used together with measured discharges within a Bayesian framework, leading to a reduction in statistical uncertainty through the experts' estimates, particularly in the range of extreme discharges. This demonstrates the effectiveness of combining expert judgment in statistical models using Bayesian sampling techniques, particularly to achieve more credible outcomes.

The fourth and last research question explores the quantification of statistical dependence through expert judgment. In the same case study for the river Meuse, the dependence between the tributary discharges was quantified using a Non-parametric Bayesian

Network. This network of nodes (representing tributary discharges) and edges (representing their dependence) simplifies the task of quantifying a correlation matrix that serves as input for a dependence model. This case study demonstrates that experts can effectively quantify dependence between nine tributaries' discharges. Evaluated by the d-calibration score, all experts performed better than the significance threshold. This was not the case when they estimated the extreme discharges under the previous research question. Whether this conclusion holds for more complex dependence patterns, especially when considering different variables, remains a topic for future research. It does however demonstrate the potential of expert judgment to quantify hydrological dependence, which can be a key factor affecting the severity of flooding.

Through the four research questions, this study demonstrated that structured expert judgment can effectively quantify the small probabilities and dependence that play an important role in flood risk assessments, particularly for engineered flood defense systems with high protection standards. However, the quality of these outcomes strongly depends on the types of uncertainties that are estimated, and the methods used to elicit or process them. Through a mix of theoretical research, software development, and applications, this thesis aims to guide future practitioners and researchers by having showed 1) which methods perform well and which did not, 2) which findings require further investigation, and 3) which tools are available to support expert elicitation. We encourage future researchers to keep exploring this field, particularly integrating expert judgment with high-level (statistical) models. In the search for more credible failure probabilities, this is where experts and models can benefit from each other the most.

Samenvatting

Risicoanalyses van overstromingen zijn veelal gebaseerd op modellen die zijn gebaseerd op fysica. Deze modellen bevatten aanzienlijke onzekerheden die moeten worden gekwantificeerd om faalkansen van waterkeringen nauwkeurig in te kunnen schatten. Aannames over deze onzekerheden zijn vaak onduidelijk, ook al hebben ze een grote invloed op de uitkomsten. Expertschattingen bieden een methode om deze onzekerheden expliciet te kwantificeren met kansen. Dit proefschrift past specifiek "the Classical Model" (vertaald: het Klassieke Model) toe. Deze methode staat ook wel bekend als Cooke's Method. In deze aanpak wordt de deskundigheid van een expert op een specifiek onderwerp bepaald, waarna een gewogen combinatie van de groep experts wordt gemaakt. Deze methode is gericht op het schatten van onzekerheid. Ze erkent dat onbekende variabelen inherent onzeker zijn en scoort experts op basis van hun vaardigheid in het schatten van deze onzekerheid.

Overstromingsrisico, vooral in de Nederlandse context, betreft gebeurtenissen met kleine kansen op voorkomen (bijvoorbeeld 10^{-3} tot 10^{-5} per jaar) maar met grote gevolgen. De zeldzaamheid van deze gebeurtenissen maakt het moeilijk om de frequentie empirisch te bepalen. Hun terugkeertijden overschrijden veelal de duur van een menselijk leven en de lengte van de meeste meetreeksen. Vanwege de serieuze gevolgen van een overstroming is het echter cruciaal om een inschatting te krijgen van de frequentie en extremiteit van dergelijke gebeurtenissen. Bovendien moet de afhankelijkheid tussen gebeurtenissen ook in overweging worden genomen. Statistische afhankelijkheid speelt een prominente rol in hydrologische extremen die tot overstromingen leiden, of ze nu verband houden met hydrodynamische omstandigheden zoals waterstanden en golven, ruimtelijke aspecten zoals regenval over verschillende stroomgebieden, of temporele factoren zoals opeenvolgende (tweeling)stormen.

Het doel van dit proefschrift is te onderzoeken hoe expertschattingen kunnen bijdragen aan het schatten van geloofwaardigere faalkansen voor waterkeringen. Dit omvat het verkrijgen van schattingen van de frequentie van hydrologische extremen (zoals overstromingen) en de beoordeling hoe deze de faalkans van dijken beïnvloedt, rekening houdend met het feit dat dit zeldzame, niet eerder meegemaakte gebeurtenissen zijn die probabilistische afhankelijkheid met zich meebrengen. Het onderzoek is gestructureerd rond vier onderzoeksvragen met betrekking tot 1) de statistische methoden en kansverdelingen die worden gebruikt om de prestaties van experts te bepalen in the Classical Model, 2) de soorten onzekerheden die het best door experts kunnen worden geschat, 3) de toepassing van expertschattingen als a priori verdelingen in een Bayesiaanse aanpak om onzekerheid in hydrologische extremen te verminderen, en 4) het vermogen van experts om statistische afhankelijkheid te schatten.

De eerste onderzoeksvraag betreft de geschiktheid van verschillende statistische tests

en verdelingen binnen the Classical Model voor expertschattingen. Het onderzoek evalueert twee kansverdelingen, onder de hypothese dat een gladde Metalog-verdeling een betere weergave is van expertschattingen dan de normaliter gebruikte trapsgewijze uniforme verdeling. De resultaten geven echter aan dat geen van beide verdelingen de schattingen van experts adequaat vertegenwoordigt. Voor een nauwkeurigere weergave van de verdelingen die experts in gedachten hebben, is het daarom aan te bevelen om hen meer percentielen te bevragen. Verder worden in het onderzoek alternatieve toetsen voor het beoordelen van experts onderzocht. Normaliter wordt hiervoor de chikwadraattoets gebruikt. Deze toets genereert een p-waarde die voor een belangrijk deel het gewicht van een experts schattingen bepaalt. Vier alternatieven, de Kolmogorov-Smirnov, Cramer-von Mises, en Anderson Darling toets, en de Continuously Ranked Probability Score, worden beoordeeld op hun geschiktheid als expertgewicht. Het onderzoek toont aan dat geen van deze toetsen significant beter presteert dan alle anderen, en dat elke toets verschillende biases en gevoeligheden vertoont. Het meest opvallende verschil tussen de toetsen is de gevoeligheid voor overvloedig zelfvertrouwen; de neiging van experts om te precieze schattingen te geven. Dit is de meest voorkomende bias in studies naar expertschattingen. Toetsen die minder gevoelig zijn voor overvloedig zelfvertrouwen kennen hogere gewichten toe aan een groter deel van de expertgroep.

De tweede onderzoeksvraag onderzoekt de nauwkeurigheid waarmee experts onzekerheden voor verschillende soorten variabelen schatten, specifiek binnen dijkbeoordelingen langs de Nederlandse takken van de Rijn. Deze analyse vergelijkt schattingen voor systeemniveau faalkans (bijv. bij welke afvoer denkt u dat ten minste één dijk langs de rivier faalt?) met gedetailleerde faalkans van dijksecties (bijv. bij welk waterniveau faalt deze specifieke sectie door piping?) De bevindingen tonen aan dat experts geloofwaardigere schattingen geven voor faalkansen op systeemniveau dan voor gedetailleerde dijksecties. Voor de laatste hadden de experts die volgens the Classical Model het beste scoorden (d.w.z. een betere weergave van onzekerheid gaven) niet de technische kennis om nauwkeurige faalkansen voor alle dijkfaalmechanismen te produceren. Het opschalen van deze resultaten van een enkele dijksectie naar een systeemniveau faalkans leverde vervolgens ongeloofwaardige kansen op. De schattingen voor faalkansen op systeemniveau daarentegen waren nauwkeuriger. Hoewel dit het gebruik van expertschattingen volgens the Classical Model niet uitsluit voor gedetailleerde dijkveiligheidsbeoordelingen, toont het wel de beperkingen van het gebruik van deze methode waar faalkansen zeer gevoelig zijn voor kleine variaties in staartkansen van geschatte dijkparameters.

Om de derde vraag te beantwoorden, beschrijft het proefschrift het gebruik van expertschattingen om onzekerheid in hydrologische extremen te verkleinen. Wanneer een statistische verdeling wordt gefit aan dergelijke afvoeren op basis van alleen metingen, is de onzekerheid in de staart meestal zeer groot. Om dit te verminderen schatten hydrologen zeldzame en extreme afvoeren in een casestudie voor de zijrivieren van de Maas. Zonder kennis van waargenomen historische afvoeren beschouwen we hun schattingen als a priori informatie. Door de prestaties van de experts te evalueren op basis van vaker voorkomende (en dus beter verifieerbare) afvoeren, worden expertgewichten afgeleid om hun schattingen te combineren. Deze schatting wordt vervolgens samen met gemeten afvoeren toegepast volgens een Bayesiaanse aanpak met als effect dat, met name in

het extreme bereik, de statistische onzekerheid beperkt wordt door de expertschattingen. Dit toont de effectiviteit aan van het combineren van expertschattingen en statistische modellen met behulp van Bayesiaanse technieken om geloofwaardigere kansschattingen te verkrijgen.

De vierde en laatste onderzoeksvraag onderzoekt het kwantificeren van statistische afhankelijkheid met expertschattingen. In dezelfde casestudie voor de Maas werd de afhankelijkheid tussen de zijafvoeren gekwantificeerd met een niet-parametrisch Bayesiaans netwerk. Dit netwerk van knopen (die afvoeren in zijrivieren vertegenwoordigen) en kanten (die hun afhankelijkheid vertegenwoordigen) vereenvoudigt de taak van het kwantificeren van een correlatiematrix die als invoer dient voor een afhankelijkheidsmodel. Deze casestudie toont aan dat experts effectief de afhankelijkheid tussen de afvoeren van negen zijrivieren kunnen kwantificeren. Evaluatie met de "d-calibration score" toont aan dat alle experts beter presteerden dan de significantiedrempel. Dit was niet het geval toen zij extreme afvoeren schatten onder de vorige onderzoeksvraag. Of deze conclusie standhoudt voor complexere afhankelijkheden, vooral bij combinaties van verschillende typen variabelen, blijft een onderwerp voor toekomstig onderzoek. Het toont echter wel het potentieel van expertschattingen voor het kwantificeren van hydrologische afhankelijkheid, een potentieel belangrijke factor voor de ernst van overstromingen.

Aan de hand van de vier onderzoeksvragen toont deze studie aan dat expertschattingen effectief kunnen worden gebruikt om de kleine kansen en afhankelijkheid te kwantificeren die een belangrijke rol spelen in overstromingsrisico en dijkfalen, met name voor waterkeringen met een hoog beschermingsniveau. De kwaliteit van deze uitkomsten hangt echter sterk af van de soorten variabelen waarvoor onzekerheden worden geschat en de methoden die worden gebruikt om ze te verkrijgen of te verwerken. Door een mix van theoretisch onderzoek, softwareontwikkeling en toepassingen, poogt dit proefschrift toekomstige beoefenaars en onderzoekers te helpen door te laten zien 1) welke methoden goed werken en welke niet, 2) welke bevindingen verder onderzoek vereisen, en 3) welke hulpmiddelen beschikbaar zijn om expertschattingen te ondersteunen. We moedigen toekomstige onderzoekers aan om dit onderzoeksveld verder te verkennen, met name door expertschattingen te integreren met globale (statistische) modellen. In de zoektocht naar geloofwaardigere faalkansschattingen is dit het punt waar experts en modellen elkaar het best kunnen versterken.

1

Introduction

Humankind has always had the tendency to live close to water. Water is indispensable for both individuals and societies. It provides a source of food and drinking water, opportunities for transport and trade, but also exposure to the risk of flooding. This applies to most of the world to a varying degree. Rentschler et al. (2022) calculated that 23% of the world population is directly exposed to floods that happen on average once per 100 years, most of them living in low- or middle-income countries. The Netherlands are, in terms of flood risk, a particularly dire case, with the world's highest relative exposure to flood risk (58.7% of the population). Much of its area was created through land reclamation and is therefore relatively flat and at similar elevations to mean sea level. Consequently, 60% of the country's land surface is prone to flooding, either from the sea or rivers (Kok et al., 2017). To drain water from this area is challenging, hence the reliance on extensive water infrastructure to keep dry feet. The windmill, a Dutch national symbol, was originally used for 'pumping' water from lower to higher drains. In light of climate change, sea-levels have risen over the past century and are projected to rise further during the next century (IPCC, 2023). Increased rainfall intensities and peak river flows combined with ongoing urbanization will make flood risk, if anything, a more relevant topic in the near future.

1.1. Dutch flood risk framework

Protecting against floods from rivers or seas is a task that benefits from collaboration. It takes fewer flood defenses to protect a larger area (with twice the dike length a four times larger area can be protected). More often, collaboration is a necessity, as the task of protecting against floods is simply too large to be done by a single person. This, in combination with the just described geographical properties of the Netherlands, has led to the establishment of so-called water authorities. These organizational structures have emerged over the past centuries to limit the threads of flooding together. This has cre-

ated a situation in which flood defenses protect substantial areas that would otherwise flood regularly.

These defenses have high levels of safety. This implies a high probability that a flood defense is safe and, conversely, a very small probability that it might fail. These probabilities are expressed as a failure probability in a year, and typically range from 0.01 for areas where floods would have relatively limited impacts to 0.00001 (once per 100,000) for densely populated areas that would suffer severely from a flood. Another way to express them is as return periods or average recurrence intervals: In a very long time series, a 0.01 probability in a year corresponds to an average of 100 years between subsequent floods exceeding a specific size. These are periods that exceed most humans' lifetimes as well as the duration of most measured data records. This complicates estimating the extremity of such events on empirical basis, introducing a need for statistics and modeling to describe such rare events.

Flood risk is typically defined by the product of flood probability and flood consequences. The flood probability has two components, load, and resistance. Loads are the water levels or waves exerting forces on the flood defenses. The magnitude of the loads is determined by meteorological conditions, hydrology, and river or coastal hydraulics. The resistance is the structural or geotechnical strength of the flood defenses. The consequences are the impact of a flood on society, such as the mortality of both individuals, groups of people, and economic losses. For more background on the Dutch flood risk framework, refer to (Kok et al., 2017).

1.2. Models for failure probabilities and consequences

The easiest way to assess the probability of a flood event is by counting historical observations. However, because flood events are rare in the Netherlands and we only have roughly a hundred years of representative data, this method cannot be relied upon for accurately estimating flood probabilities. Instead, models and simulations are needed to estimate failure probabilities and consequences. Moreover, in the Netherlands, safety levels are defined by law. This encourages the use of standardized methods and models to determine flood defense safety. Within the Dutch flood risk framework, different models are used for various parts of the flood probability calculations. The principles on which these models are based differ. For example:

- The hydraulics of water flow (e.g., water flowing through a river) can largely be modeled from first principles. This means that the model reflects physical laws, such as conservation of mass or energy. The finer details, such as the interaction between the water and the riverbed, are however modeled with a "lumped" roughness parameter.
- Flood defense failure is an interaction between water and the flood defense. Flood defenses are often geotechnical structures. Different model types are used to determine flood defense failures. For example, the Sellmeijer piping model (Sellmeijer, 1988), which is often used in this case, is based on the pressure gradient caused by a physical quantity such as the difference in water level across a dike. Detailed

processes, such as the erosion of the soil, are however based on empirical relations derived from research.

- Hydrological modeling from rainfall to river discharge is an example of a process with many unknowns. Even for something as visible as runoff resulting from rainfall, physical processes are a challenge to describe. This only gets more difficult for subsurface flow. Hydrological modeling is therefore an example where the physical process is modeled with significant simplification or is based on empirical relationship.

Models do not need to be based on physical processes or a causal chain of events. For example, linear regression (drawing a relational line through a cloud of points) can describe the relationship between two variables without explicitly stating how they are connected, and which causes which. Joint probability distributions extend such a model by expressing the relation between two (or more) variables including the probabilistic uncertainty (how far the points are from the line).

Models used for flood risk assessment are generally physics-based models, which are usually deterministic and have no explicit representation of uncertainty. Given the hydraulic loads, the model will return a single answer (e.g., failure or not, or a specific water level) and not a probability that a certain value is exceeded. However, a probabilistic model outcome can still be achieved by defining the uncertainty in the input parameters and propagating this through the model. Two approaches can be distinguished for doing this. Firstly, the probabilistic approach. In this approach all combinations of (uncertain) model input parameters and their dependence, and the outcome of each combination, are integrated in the failure probability assessment. While the approach is in theory most accurate, it can be a challenging task to correctly resolve all uncertainties. The second approach is a simplification of this, called the “semi-probabilistic” approach, in which a design value of a model uncertainty parameter is calibrated beforehand using a (fully) probabilistic approach. Such design values contain a safety factor and can, for example, be derived from the most likely combination of model parameters that led to failure in a fully probabilistic assessment. While this approach is much more straightforward, semi-probabilistic design values are (or should be) more conservative, as they could be applied in varying conditions where failure can be sensitive to different parameters. The Dutch flood risk approach combines semi-probabilistic and probabilistic elements and distinguishes itself from more widely used approaches by putting more emphasis on (fully) probabilistic assessments. This increases the importance of deriving uncertainties on a case-by-case base.

1.3. Uncertainties and their impact

To understand the types of uncertainties involved in flood risk, consider a dike along a river that fails due to overflow. The hydrology of the upstream catchment determines the amount of water flowing through the river during extremes and is expressed with discharge statistics. Typical sources of uncertainty in these are (Fig. 1.1 a): 1) the need for extrapolation from relative short time series of historical observations, 2) a non-stationary

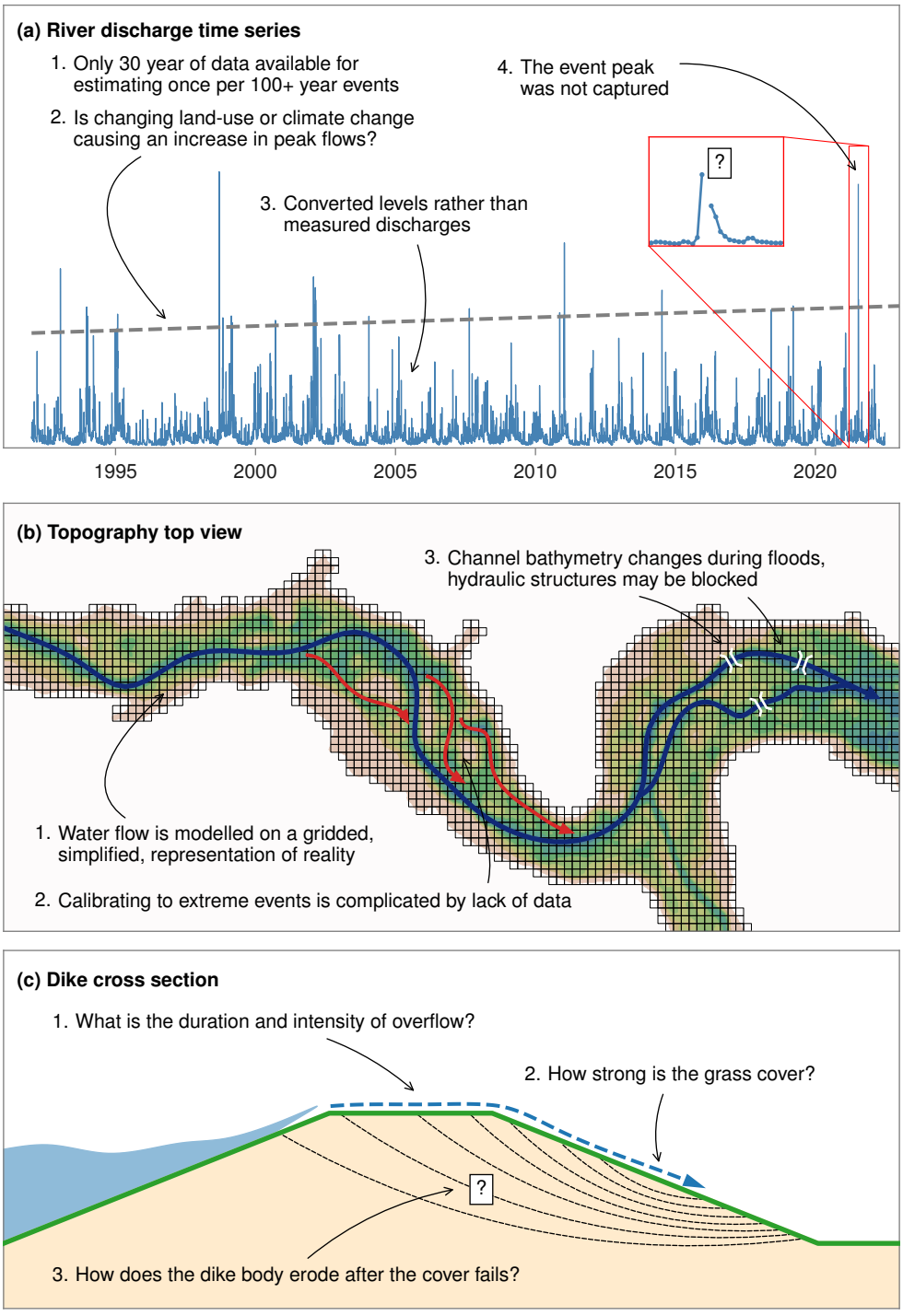


Figure 1.1: Illustration of different sources of uncertainty in hydrology, hydraulic modeling, and flood defense failure.

system, for example a changing catchment or climate, 3) missing observations (floods often exceed or destroy measuring devices), and 4) inaccuracies in stage-discharge relation that is used to transform the measured water level to a discharge. Subsequently, these (transformed) discharges are converted to water levels in the river using a hydrodynamic model. Typical sources of uncertainty in this are (Fig. 1.1 b): 1) a lack of detail in the model, 2) the absence of suitable calibration events, or 3) a changing river during the event. Finally, uncertainties in overflow failure could be caused by (Fig. 1.1 c): 1) the duration and intensity of overflow, 2) uncertainty about the quality of the grass cover and its ability to withstand overflow, and 3) the erosion of the dike body after grass cover failure. Many of these uncertainties can be interrelated. For example, a changing catchment may not only affect runoff but also affect the stage-discharge relation. Or perhaps the quality of the grass-cover is partly related to the same aspect that affects erosion speed of the subsoil.

All these uncertainties can be related to a physical parameter like peak discharge, local water level, or critical flow velocity. These are considered random variables, variables that do not take on a single value but follow a probability distribution. The peak discharge is an example of a random variable that serves as input for the subsequent flood models. It stems from a non-stationary stochastic process and is known as an aleatoric (random) uncertainty. The uncertainty in using historical data (counting occurrences) is an aleatoric uncertainty because we never have infinite data (all possible occurrences). Another example of uncertainty is the uncertainty that results from lack of knowledge. These could in theory be reduced by taking more measurements or using a modeling with a finer resolution. Such uncertainties are called epistemic (knowledge-related) uncertainties and are often related to model parameters rather than boundary conditions. There are other uncertainties that do not fall within one of these categories. For example, uncertainties introduced by errors, human actors, or ethical and legal constraints (French, 2023). Epistemic and aleatoric uncertainties are however the most common uncertainties in modeling physical processes.

So how do we handle these uncertainties? The preferred way of quantifying uncertainties is through data. However, empirical data are often unavailable for model parameters with large spatial variation or parameters that are not visible to the naked eye (e.g., grain sizes or permeability). In absence of field data, a modeler will likely use default values and parameter values available in literature. If these uncertainties do not represent the actual situation, the modeled failure probability will be inaccurate. If the failure probability is very large, for example one in 5 years, and no failure has been observed over the last 30 years, the result can be invalidated with relative certainty. However, if the results would be a one in 30 years probability (i.e., a 63% chance of at least one failure over the last 30 years), the data are not conclusive and it comes down to the modeler's perception to discover a potential error or inaccurate assumption, which might then be traced back to the input. Assuming that experienced practitioners are well able to identify such errors, perhaps they can also use that knowledge to estimate the failure probability, model parameters, or model outcomes directly? This is where expert judgment comes into play, a main topic of this research.

1.4. Structured expert judgment in flood risk

Expert Judgment involves a large variety of techniques to gather estimates for unknown variables from experts. This can be an informal estimate from your experienced colleague but also a formal process in which probability assessments are elicited and validated. Structured expert judgment (SEJ) falls in the last category. It formalizes the elicitation of expert judgments in such a way that the results can be treated as scientific data. Structured expert judgment is used to quantify variables that are difficult to quantify using other methods, for example because field measurements or experiments are infeasible, too costly, or unethical. The Classical Model, also known as Cooke's method, is one method of SEJ and plays a central role in this thesis. Section 2.1 explains the Classical Model in detail. To already get an idea of the method, these are the steps involved in an expert elicitation following the Classical Model:

1. The person or team in need of uncertainty estimates prepares a questionnaire with questions regarding the variables to quantify. Extra questions (usually 10 to 20) are added to this for the analyst already knows the answer.
2. A group of experts (usually 5 to 15) is gathered to participate in the elicitation.
3. The session takes place during one or a few days, preferably in-person. The participants will be introduced to the Classical Model as well as the topic, after which they will fill out the questionnaire individually.
4. After this, the experts will be weighted based on their performance in the extra questions with the known answer. These performance-based weights are used to combine the expert distributions for the questions of interest.

Flood risk is a field in which causal, physics-based, models dominate. Compared to this, expert judgment has a subjective connotation. After all, it does not ensure that an answer is based on physical principles. However, the alternative of relying on a model to generate an answer can similarly mask hidden assumptions. Models require inputs and schematization, which are mostly based on an expert's interpretation or judgment. Structured expert judgment does not aim to replace or exclude the use of physics-based models, it is usually done in conjunction with models. It is a method to quantify the experience from practitioners, with or without models, in a scientific manner.

Several studies have applied structured expert judgment in flood risk, considering either the hydrological or dike failure aspects (e.g., Cooke & Slijkhuis, 2002; Hathout et al., 2019; Kindermann et al., 2020; Slijkhuis et al., 1998). More applications of SEJ are found in the wider fields of reliability analysis, nuclear safety, public health, aviation, climate, and natural hazards (Cooke & Goossens, 2008). Guidelines for practical applications of SEJ (not exclusively the Classical Model) in flood risk are available in the United States (Ayyub, 2000), and in the Netherlands (RWS-WVL, 2020). Despite this, structured expert judgment plays a minor role in flood risk, particularly in addressing the specific challenges of this field, such as large uncertainty, rare events, and dependence between risk factors. This thesis aims to address this gap by conducting research and applications on expert judgment in flood risk applications, thereby focusing these specific aspects.

1.5. Research objective

This thesis' objective is to research how structured expert judgment can enrich the model-based approach typically found in flood risk. Two characteristics that distinguish this field from (some of the) other fields in which expert judgment is applied, are:

- Flood risk, especially in the Dutch context, is characterized by very small probabilities (10^{-3} to 10^{-5} per year). These correspond to events that have likely not occurred in the lifetime of a practitioner and are likely not on record.
- Extreme flood risk events involve many uncertainties. It is not unlikely that a failure is caused by a combination of "outliers" rather than a single variable being extreme. This makes the dependence between different uncertainties a relevant aspect to consider.

These two characteristics, small probabilities, and the dependence between them, are not well covered in existing expert judgment research and literature, while their impact on flood risk can be substantial in flood risk assessments. The objective of this research is to consider how expert judgment can contribute to more credible failure probabilities for engineered flood defense systems, in the context of rare non-experienced events and their probabilistic dependence. This objective is addressed through four research questions:

1. What measure should be used to score experts, to obtain the best results in structured expert judgment studies following the Classical Model?
2. Which variables' uncertainties are most accurately estimated by experts?
3. How can expert judgment reduce uncertainty in the tails of probability distributions?
4. How do experts perform in estimating probabilistic dependence?

1.6. Approach and outline

Each of Chapters 4, 5, 6 and 7, is based on a publication or manuscript that treats one of these questions. Sections 1.6.1 to 1.6.4 give a brief description of the followed approaches. The conclusions (Chapter 8) present the learnings from these chapters in the context of the four questions, thereby addressing the main research objective.

Because not every reader might be familiar with the concepts and methods used in this thesis, Chapter 2 provides an accessible explanation of these, being 1) structured expert judgment, 2) statistical dependence, and 3) Bayesian statistics. Additionally, two software programs were developed and extensively used throughout the research. Chapter 3 presents these. The first, ANDURYL, is used for eliciting univariate uncertainty. The second, Matlatzinca, is a specialized software to quantify dependence using Non-Parametric Bayesian Networks. This software is publicly available for future users to reduce the burden of processing elicitation results.

In summary, the research has three elements: applications, theory, and software. The research presented in this thesis involves applications of structured expert judgment through the second and third question. The first and fourth question involve research on the theoretical side of the structured expert judgment. Additionally, software that supports expert elicitation was developed during the research. Figure 1.2 shows how the different chapters and research questions tie together.

1.6.1. Methods for evaluating expert estimates in the Classical Model

When applying the Classical Model for structured expert judgment, experts are asked to express their estimates typically through three or five percentiles. Together, these uncertainty estimates form a probability distribution for the variables. For part of the questions, the outcome is known to the researcher. The Classical Model evaluates the expert's accuracy by examining in between which estimated percentiles the answer is located. Doing this for all questions with known outcomes results in a set of ratios (observed over expected outcomes per interval) that can be evaluated using the statistical test based on the chi-square distribution. The resulting p-value expresses the probability that an expert is accurate and consequently contributes to the expert's weight accordingly.

If a probability distribution is assumed to connect the estimated percentiles, the outcomes can be transformed into a set of quantiles rather than quantile intervals. This creates possibilities for using different statistical tests as measure of statistical accuracy. Chapter 4 demonstrates this, by calculating quantiles using the normally used piecewise uniform distribution and the smooth Metalog distribution. These were then evaluated using statistical tests such as the Kolmogorov-Smirnov-, Cramer-von Mises- and Anderson Darling-test and the recently published Continuously Ranked Probability Score. This reveals biases in the original chi-square-based test as well as the newly evaluated tests, and shows which distribution best represents expert estimates. Through this, research question 1 is answered.

1.6.2. Different types of uncertainty assessments for dike safety

To find out how the Classical Model performs for dike failure assessment, dike system failure is estimated for the branches of the Dutch Rhine River. Part of this research, described in Chapter 5, involves learning which type of variable's uncertainty is most accurately estimated by experts in this flood risk context. To find out, the failure probability of dikes along the Dutch River Rhine is estimated using two approaches. First, experts estimate the full conditional failure probability, i.e., the discharge at which at least one dike in the system would fail. Combining this with discharge statistics results in the failure probability. In the second approach, failure is assessed for detailed dike sections. These estimates are compared to model results to determine the model bias as perceived by the experts. These estimates and biases are then combined with existing model-results to obtain, again, a system-wide failure probability. Comparing the two approaches shows 1) how conservative experts think models are, 2) how safe they think the dikes currently are, and 3) how consistent their detailed estimates are with the high-level estimates and which of the two are more credible. From this last point, research question 2 is answered.

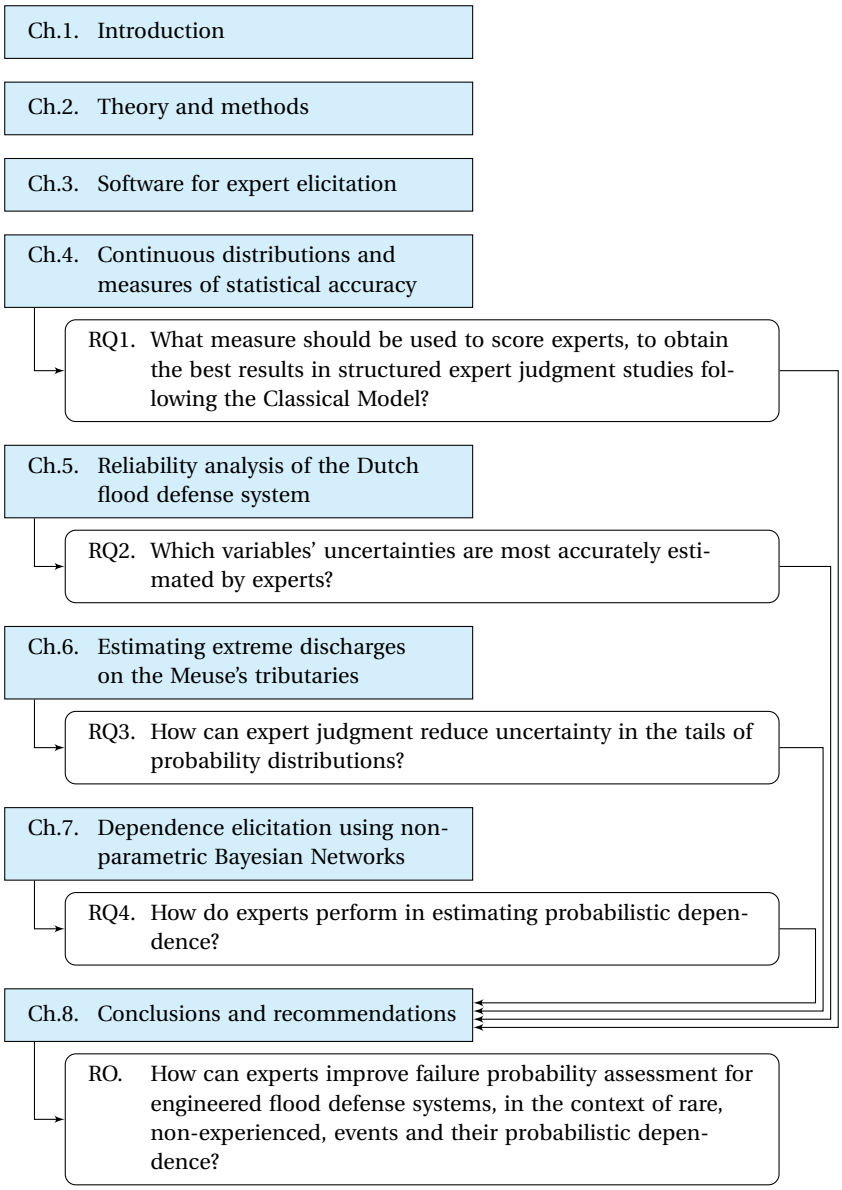


Figure 1.2: Relationship between chapters, research questions, and research objective.

1.6.3. Limiting uncertainty in extreme discharges with expert judgment

Determining the frequency of extreme hydrological events is an important task in flood risk, as it is the main driver of potential floods. Often no more than a few decades of observations are available. Based on this an estimate of, for example, a discharge that is exceeded on average once per 1000-year needs to be made, which leads to extrapolating with great uncertainty. Chapter 6 describes a case study to assess if expert judgment can contribute to reducing the uncertainty for such rare hydrological events. Seven hydrologists estimated extreme discharges for the river Meuse's larger tributaries. Based on a relatively frequent 10-year discharge, the expert's performance is evaluated using the Classical Model. The resulting weight is applied to their 1000-year estimates, resulting in a single aggregated estimate per tributary. These estimates are used as prior distributions for an extreme value distribution of discharges and combined with observed discharges in a Bayesian framework. This means a model is fitted not just to the observations, but also to an expert's perception of what the outcome should be. The results show how expert judgment performs in limiting uncertainty, answering research question 3. Additionally, it shows the potential of fitting models to expert estimates to limit their uncertainty or incorporate prior information.

1.6.4. Estimating dependence and aggregating scores

Dependence between hydraulic, hydrological, or strength-related variables is another difficult to quantify element in flood risk assessments. When considering the dependence between an increasing number of variables, the number of observations needed to cover the full range of combinations increases as well. In the same study that considered hydrological extremes, the dependence between the Meuse's tributary extremes is assessed by the seven hydrologists as well. Chapter 7 describes how the experts perform for estimating dependence, and based on this, what scoring rule should be used to combine experts' dependence estimates. The experts are tasked with estimating correlations between different tributary discharges during a flood event. To aid them in quantifying a correlation matrix, they use a Non-parametric Bayesian Network (NPBN), a network of nodes and edges that expresses the dependence between the node variables. To aggregate the estimates into a single dependence estimate, the d-calibration score is used. This is a performance-based weight calculated from the agreement between observed and estimated correlations. While scoring and weighting univariate uncertainty (i.e., the Classical Model) has been extensively researched, scoring dependence is relatively unexplored. Therefore, several characteristics of the d-calibration score are explored as well. The results demonstrate how NPBNs can assist experts in quantifying large correlation matrices and explores how the different experts' estimates can be aggregated. Through this, research question 4 is answered.

1.7. Code and data availability

The data and code underlying this thesis are openly available under the GNU GPL 3.0 license at: https://doi.org/10.4121/a6333b17-bab2-476f-a636-61244b5c6f9e

2

Theory and methods

2.1. Structured expert judgment

This research applies the Classical Model to elicit information from experts. The Classical Model, also known as Cooke's Method, is a structured approach to elicit uncertainty for unknown quantities. It formalizes expert judgment in such a way that the results can be treated as scientific data. The method combines expert judgments based on empirical control questions with the aim to find a single combined estimate for the variables of interest (i.e., a rational consensus). The Classical Model is typically used when alternative approaches for quantifying uncertain variables are infeasible or unsatisfying (e.g., due to costs or ethical considerations). The method is extensively described in (Cooke, 1991) while applications are discussed in, for example, (Cooke & Goossens, 2008). This section gives an accessible explanation to provide the reader with the necessary background to help understand the remainder of this thesis.

2.1.1. Quantifying and evaluating uncertainty

In the Classical Model, a group of mostly five to fifteen participants, which are often researchers or practitioners in the field of interest, provides uncertainty estimates for a set of questions. The participants meet up, preferably in person, for the expert elicitation. If results from an elicitation are published, the experts' names are (with consent) listed, but the results are anonymized such that they cannot be related to the individual participants. The participants of the expert studies are consistently called "experts". However, the goal of the Classical Model is to assess their ability of estimating uncertainty for the topic of interest, regardless of their professional reputation.

Expertise is evaluated using a number of so-called seed questions for which the outcome is known by the person doing the elicitation (the analyst or problem owner). the performance of the participants is evaluated based on the estimates for the seed questions.

This is used to assign weights to each expert, which are then applied to the different estimates for the variables of interest. In other words, seed questions are used to determine the participant's expertise in estimating uncertainty for the topic of interest. Seed questions are ideally sourced from similar studies or cases and are as close as possible to the variables of interest. Typically, about ten to twenty seed questions are answered. The other type of questions, which concern the variables of interest, are called target questions. For these, the outcome is unknown to both researchers and participants.

Because the goal is to elicit uncertainty, experts estimate percentiles rather than a single point value. Mostly these are the 5th, 50th, and 95th percentiles. Sometimes the 25th and 75th are added to these original three to elicit more detailed distributions. However, the analyse can choose a different set as they like. For example, in the research described in Chapter 5, the 1st and 25th were added to the original three to obtain more detail in the lower distribution tail.

To illustrate the method, consider the following example question that might come from a hypothetical study on crop yields under climate change: "How many oranges did an orange tree on orchard X yield, on average, during season Y?". Consider two experts and their estimates displayed in Fig. 2.1. For the 5th, 50th and 95th percentile, Expert A estimated 150, 350, and 500, Expert B 260, 320, and 390. The 5th, 50th, and 95th percentiles in combination with the item bounds (more on those below) create four quantile intervals represented by a probability vector with $p = (0.05, 0.45, 0.45, 0.05)$. If these probability masses are spread uniformly over the quantile intervals, the probability distributions in Fig. 2.1 emerge.

Assuming the outcome is 250, expert A's estimate captures it in between the 5th and 50th percentile. According to A's estimate, the probability that the outcome is in between 150 and 350 is 45%, which contributes to the likeliness that Expert A has estimated the correct distribution. On the other hand, Expert B's estimate "misses" the realization which is just below the 5th percentile. According to B's estimate, the probability that the outcome is below 260 is only 5%. Given that the realization is 250, expert B's estimate is likely worse. However, such conclusions cannot be drawn with confidence from just the one question. Therefore, the researcher could decide to elicit 19 more seed questions. From each of these questions, the realization can be categorized in a similar way over the quantile intervals. This results in a four-element vector $s(e)$, expressing the fraction of realizations within each of expert e's quantile intervals.

Figure 2.2 illustrates the possible s-vectors for both experts. The realizations from Fig. 2.1 are circled. The closer the expert's vector s is to p, the higher the *statistical accuracy*. Expert A, with $s(A) = (0.0, 0.40, 0.45, 0.15)$, has a higher statistical accuracy than Expert B with $s(B) = (0.25, 0.15, 0.40, 0.20)$. The details of this calculation are presented in Section 2.1.2. The expert weight in the Classical Model is mainly determined by the statistical accuracy. Expert B did however give a substantially smaller range for the example question; 90% of the probability mass in between 260 and 390, rather than in between 150 and 500. This is a more informative estimate. While not as influential as statistical accuracy, this informativeness, expressed with the information score, also contributes to the expert weight and is explained in Section 2.1.3.

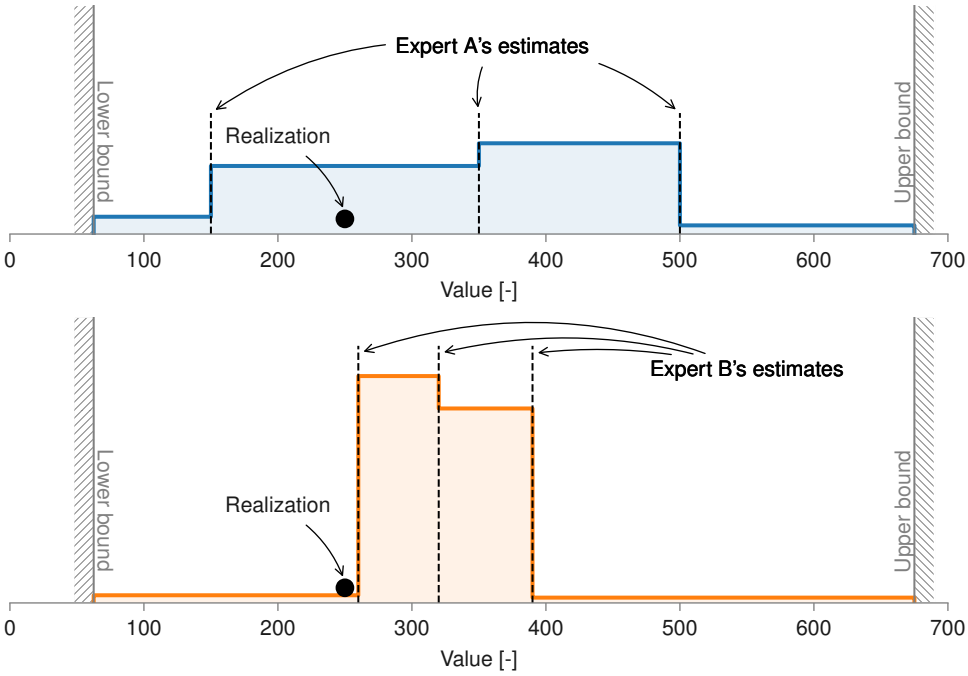


Figure 2.1: Two experts' hypothetical estimates with probability density uniformly distributed in between the estimates and the bounds.

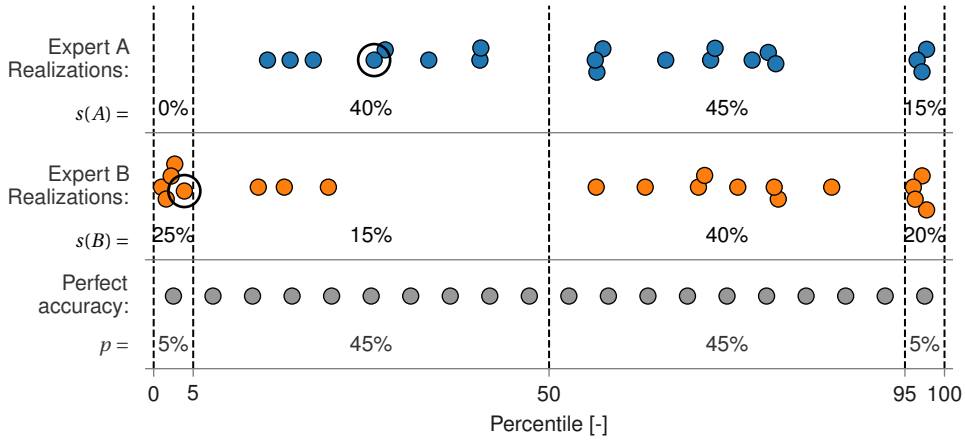


Figure 2.2: Realizations for 20 items in relation to the quantiles estimated by two experts (top and middle row), and the distribution required for perfect statistical accuracy (bottom row). The black circles indicates the realization illustrated in Fig. 2.1.

The Classical Model is a method to elicit uncertainties. The observed outcomes to the seed questions are considered to be realizations from a probability distribution (which is why the outcome is called a "realization"). Even in the case of the example question, which is very specific and only has one outcome, this outcome is still considered a realization from a distribution. For a similar orchard in a similar season, we would expect a different outcome.

2.1.2. Statistical accuracy

The ability of the expert to estimate uncertainty is expressed with the statistical accuracy (SA). This measure is calculated by comparing each inter-quantile probability p_i to $s_i(e)$. If an expert's uncertainty estimates are accurate and the sample is very large, $s(e)$ approaches p, and the interquantile ratios s_i/p_i are close to 1.0. In that case, the quantity

$$2NI(s(e)|p) = 2N\sum_{i=1}^{4} s_i \log(s_i/p_i), \tag{2.1}$$

in which, $I(s(e)|p)$ is the relative information and N the number of items, is asymptotically χ^2 distributed with three degrees of freedom[1]. This means that if an expert is statistically accurate, we know which distribution $2NI(s(e)|p)$ will follow (the chi-square distribution). In other words, by calculating this score for an expert's estimates we are calculating the probability that the expert is statistically accurate (i.e., the probability that the realizations are indeed independently drawn from the expert's distributions). This calculated probability, or p-value, is used as measure of SA. In order to obtain a perfect statistical accuracy in the example from Fig. 2.2, the experts' estimates should result in overestimating one seed question (i.e., the observed outcome was below the 5th percentile), underestimating one question, and nine questions in both the [5%, 50%] and [50%, 95%] interval. This is illustrated by the dots on the lower row, which do not necessarily need to be evenly spaced in between the quantiles. Given the uncertain nature of the realizations, it is unlikely that the realizations $s(e)$ match p exactly. Even for a perfectly accurate expert, the possible p-values are uniformly distributed between 0 and 1 (like the p-value from any statistical test). However, less accurate experts tend to get p-values that are one or more orders of magnitude smaller, such that the p-value is still useful for distinguishing expert performance.

2.1.3. Information score

A high statistical accuracy does not automatically imply that an expert's estimate is informative. For example, in Fig. 2.1 Expert A captures the realization within the 5th and 50th percentile interval, while Expert B does not. This likely contributes to Expert A's SA, but their wide estimate does not provide much information. A second score is therefore introduced to reward informativeness.

[1] Asymptotically means that the relative information is chi-square distributed for an infinite number of questions. Since this is never the case, the calculated p-value is an approximation. This asymptotic property is further elaborated in Chapter 4

This score, called the information score, compares the degree of uncertainty in an expert's answer to other experts' answers. Percentile estimates that are close together (compared to the other participants) are more informative and get a higher information score. The information score is calculated by using the relative information between expert e's estimate for item i and the background probability density for the item: $I(f_{e,i}|g_i)$. This is illustrated in Fig. 2.3, which shows, again, the two experts' estimates for the question about the number of oranges. The larger the difference between the expert's probability density ($f_{e,i}$) and the background density (g_i), the higher the informativeness of the estimate. The disagreement between $f_{e,i}$ and g_i is illustrated with the hatched areas. To define the background probability density, a possible interval for the item needs to

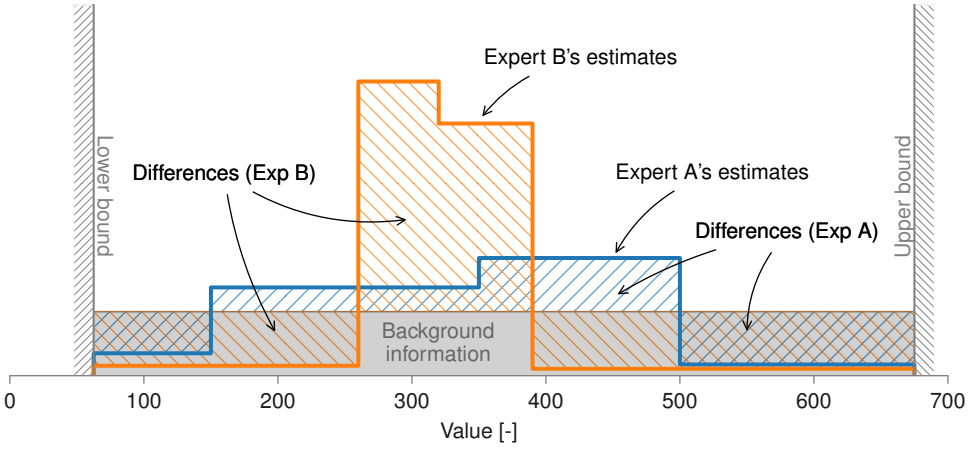


Figure 2.3: Conceptual illustration of the informativeness of the two experts' estimates for a single item, relative to the background information that is uniformly distributed between the item bounds.

be defined, between which the background probability is uniformly distributed. This interval is defined for every item, by combining all experts' quantile estimates $q_{e,i}$ and realization x_i. This gives the interval:

$$[L_i, U_i] = \left[\min\left(x_i, \min_{e=1}^{n}\left(q_{e,i}\right)\right), \max\left(x_i, \max_{e=1}^{n}\left(q_{e,i}\right)\right)\right] \tag{2.2}$$

Additionally, an overshoot k (typically 10%) is added to this, to add tails to the distributions and ensure a finite probability density outside each expert's 5th and 95th percentile (5% of the probability needs to be assigned outside both the 5th and 95th quantile). This changes the interval to

$$[L_i^*, U_i^*] = [L_i - k \cdot (U_i - L_i), U_i + k \cdot (U_i - L_i)]\,. \tag{2.3}$$

With this so-called intrinsic range $[L_i^*, U_i^*]$, a vector $Q = (L_i^*, q_{0.05}, q_{0.50}, q_{0.95}, U_i^*)$ for item i can be composed, comprising the expert's quantile estimates and the outer bounds. The relative information is calculated from this vector:

$$I\left(f_{e,i}|g(i)\right) = \log\left(U_i^* - L_i^*\right) + \sum_{j=1}^{4}\left(p_j \cdot \log\left(\frac{p_j}{Q_{j+1} - Qj}\right)\right) \tag{2.4}$$

Finally, to calculate the expert's overall information score, the average of $I(f_{e,i}|g(i))$ over all items i, is taken.

2.1.4. Decision makers

The Classical Model combines the different experts' estimates using rational consensus. This implies that a single estimate is composed from the different expert estimates using weights that are based on the experts' performance. The resulting combination is called the decision maker (DM).

The combined score of expert e, $w_\alpha(e)$, is calculated from the product of the statistical accuracy and the information score, which after normalization will be the expert weight:

$$w_\alpha(e) = 1_\alpha \times \text{statistical accuracy}(e) \times \text{information score}(e). \tag{2.5}$$

The statistical accuracy varies more steeply (e.g., from 10^{-5} to 0.5 between experts) than the information score (e.g., 0.5 to 2.5 between experts). This (intentionally) makes the statistical accuracy dominant in the expert weight. The information score only modulates between experts with a similar SA. Experts with an SA lower than α can be excluded from the pool by using a threshold, expressed by the 1_α in Eq. (2.5). This threshold is usually 5%, similar to the threshold mostly used in other statistical tests. This means that the probability of discarding an expert that is statistically accurate is 5%[2]. The experts contribute to the DM's ith item estimate by their normalized weight:

$$\text{DM}_\alpha(i) = \sum_e w_\alpha(e) f_{e,i} \Big/ \sum_e w_\alpha(e). \tag{2.6}$$

This is called the global weight (GL) DM, composed from weights calculated using the Classical Model. Alternatively, a few other DMs are available to choose from:

- A variation on the global weight DM is the item weight (IT) DM. This option varies experts' weights between items (questions) based on the expert's information score for that specific item. The rationale for this is that more informative estimates come from greater knowledge about specific items and should therefore be granted a larger weight in that item. Using item weights does not change the statistical accuracy so the effect is usually limited (as the SA dominates the weight).
- A third option is to assign experts the same weight, resulting in the equal weight (EQ) DM. This more inclusive option does not require eliciting seed variables but neither does it distinguish experts based on their performance, a key aspect of the Classical Model. Cooke et al. (2021) compared GL weights to EQ weights in an out-of-sample cross validation. They showed that using performance-based weights increased the informativeness of the decision maker estimates by assigning weight to a few experts, without compromising the DM's statistical accuracy (i.e., the performance of the DM in "quantifying" uncertainty).

[2] This is an approximation because the statistical accuracy is calculated through an asymptotic distribution, rather than an exact distribution

- Finally, weights can be assigned on different criteria than described above. In software, this is called the 'user' weight (US) DM.

Because the expert estimates are combined into a DM, the DM "makes" estimates as well and a statistical accuracy and information score can be calculated for it. This property is used to optimize and evaluate the DMs.

The Global and Item weight DMs are optimized by choosing a significance level α such that the resulting experts (with highest SA) have the largest performance-based weight $w(\alpha)$. This typically leaves one or two experts that contribute to the DM, removing the majority of the experts that had a low weight, increasing the informativeness of the decision maker.

Furthermore, a robustness analysis can be performed to assess the sensitivity of a study to the experts and items included in it. This involves removing one or more experts or items from the project and recalculating the decision makers. The resulting variation in SA and information score indicates the sensitivity of the project to specific experts or items. It is however only an in-sample check (we cannot check the sensitivity to experts or items that are not in the study). Figure 5.5 illustrates the results of a robustness analysis.

On a final note, Chapter 4 presents several alternatives for different parts of the Classical Model. For example, the stepped distributions in Fig. 2.1 are replaced by smooth distributions in order to test if these better represent the experts' estimates in between the estimated quantiles. Furthermore, a number of alternatives measures of statistical accuracy are considered, which project realizations on a continuous scale rather than within intervals (see Fig. 2.2).

2.2. Statistical dependence

Statistical dependence in flood risk is found in different forms. Some examples are: The spatial dependence between the strengths of adjacent dike sections with similar soil characteristics, the dependence between different parameters during a storm such as wind and waves, or the occurrence of a second big storm shortly after a first storm (temporal dependence). Which dependencies are relevant in which situation mainly depends on the way in which a phenomenon is described. For example, when modeling hydraulic loads for a flood defense, the correlation between water level and wave height needs to be considered because the combination of both can exacerbate failure. Alternatively, the meteorological conditions (wind speed, wind direction, barometric pressure) can be modeled, from which the water level and waves follow more naturally. However, this simply shifts the dependence that needs to be modeled and does not resolve it. Different approaches to modeling were described in Section 1.2. Modeling phenomena through statistical dependence is suitable when relationships between variables (their joint occurrence) are relevant but difficult to model effectively based on physical principles. This section describes different methods for modeling statistical dependence through flood risk related examples.

2.2.1. Scatter plots and conditional probabilities

Dependence can intuitively be visualized using a scatter plot, such as shown in Fig. 2.4. The left panel shows two variables with a weak correlation, the right two variables with a strong correlation. Dependence between two variables implies that information about one variable contains information about the other. For example, knowing that one variable is in the right hatched area (i.e., $P(X_1 > 0.85)$) increases the probability that the other variable is in the upper area, i.e., $P(X_2 > 0.70|X_1 > 0.85) > P(X_2 > 0.70)$. This is the case for both, but in the right scatter the "predictive power" of knowing $P(X_1 > 0.85)$ is much larger because of the strong correlation.

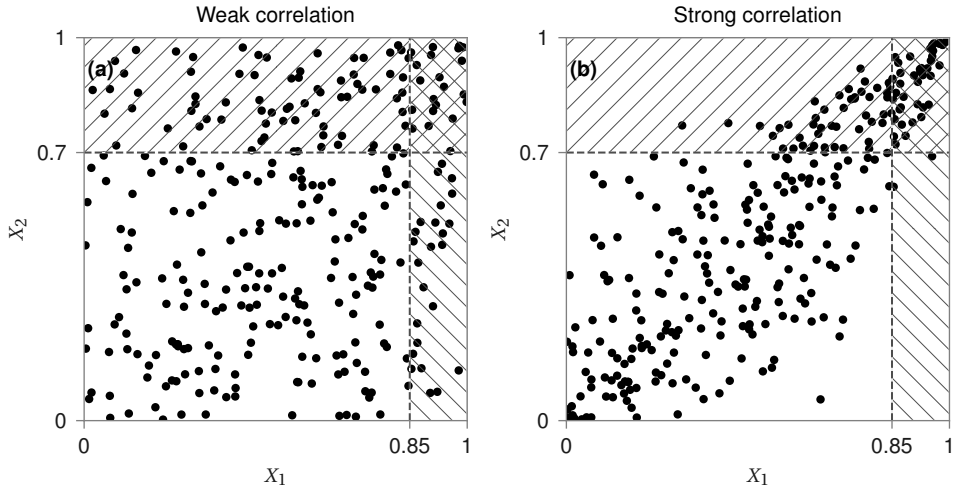


Figure 2.4: Scatter plots of weakly (a) and strongly (b) correlated variables X_1 and X_2. The hatched areas illustrate failure regions discussed in the text.

Sometimes, we are interested in determining the probability that two or more variables exceed a specific level, given their dependence. In the context of this thesis, this is mostly related to dike failure. Consider the scatters in Fig. 2.4 again. If a system failure happens when an event (x_1, x_2) is in one or both of the hatched areas, i.e., $P(X_2 > 0.70 \cup X_1 > 0.85)$, the strength of the dependence matters: While $P(X_1 > 0.85)$ and $P(X_2 > 0.70)$ are similar in both plots, the stronger dependence in the right plots means the total number of events in one of the areas is less. In other words, the failure probability is smaller because failures occur simultaneously instead of separately. This form of dependence is applied for example to combine failure probabilities of adjacent dike sections and is further elaborated in Section 2.2.2.

In other cases, we are interested in the shape of the scatter plot (i.e., the dependence over the full range of two or more variables). This is particularly relevant in probabilistic analysis where failure is caused by a combination of variables instead of the exceedance of specific threshold levels. The dependence of interest could be in between wave height and water level, permeability and grain size, or the rainfall in two different catchments. This application of dependence modeling is explained in Section 2.2.3.

2.2.2. Combining failure probabilities

In the Dutch flood risk methodology, a dike section is split into separate segments that are each represented by an individual transect. A flood risk assessment considers the failure probability of the complete system for which the transects' probabilities need to be combined into the system failure probability. Figure 2.5 shows what such a schematization may look like. The dikes on both sides of the river are split based on general orientation (relevant for wave attack) and soil characteristics (relevant for geotechnical failure).

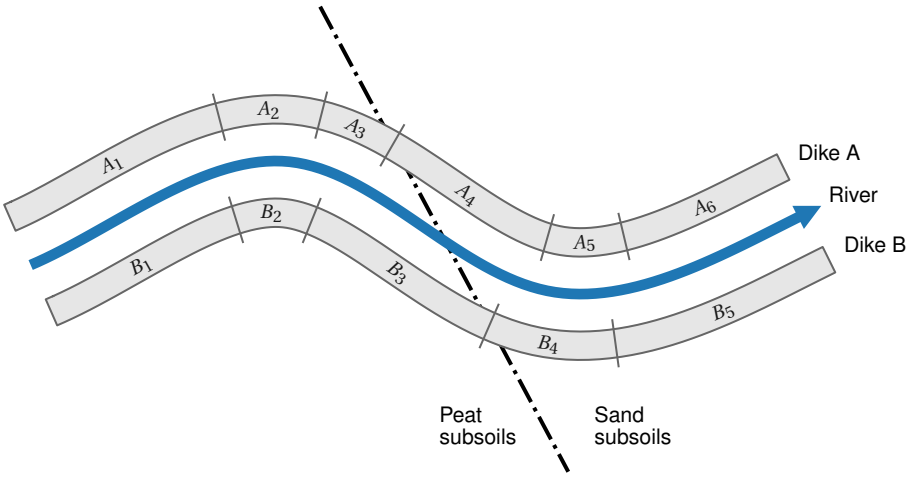


Figure 2.5: River with dike sections, split on direction and subsoil type

Most flood defenses behave like a serial system, meaning it will fail when one or more of the components fail (in contrast to a parallel system, where all components must fail for the system to fail). If the failure probabilities of the six individual components in dike A are considered independent, the total failure probability can be calculated with

$$P_{f,A} = 1 - \prod_{i=1}^{6} \left(1 - P_{f,A_i}\right), \tag{2.7}$$

in which $P_{f,A}$ is the system failure probability and P_{f,A_i} is section A_i's failure probability. In words, this means that the total failure probability is 1 minus the probability than none will fail. At the other end of the dependence spectrum is the assumption of full dependence, in which case the maximum failure probability determines the system failure probability, or

$$P_{f,A} = \max_{i=1}^{6} \left(P_{f,A_i}\right). \tag{2.8}$$

Independently combining failure probabilities in a serial system, results in the largest failure probability. It is the most conservative assumption, so, if possible, it is worth considering the dependence in some form. Describing the exact dependence between

possible failure of subsequent sections is difficult, given the many uncertain variables involved. Therefore, dependence in section failure probabilities is typically processed by making high level assumptions. A few examples are:

1. In the Dutch flood risk instrument, some dike failure mechanisms are assumed to fail more or less independently while for others the dependence is too strong to ignore. This depends on the spatial variation of parameters to which the failure mechanism is sensitive and consequently also on the length of the considered dike sections. One way to approach this is using the so-called length-effect, which adopts the minimum of two failure probabilities. The first is the independent combination of section failure probabilities, which is the most conservative approach for combining dike sections. A second failure probability is based on the section with the maximum failure probability and the number of times, N, that a representative length fits in the total dike trajectory that is considered: $N \cdot \max_{i=1}^{6}(P_{f,A_i})$. The final failure probability is the minimum of the two probabilities.

$$P_{f,A} = \min\left(N \cdot \max_{i=1}^{6}\left(P_{f,A_i}\right), 1 - \prod_{i=1}^{6}\left(1 - P_{f,A_i}\right)\right) \tag{2.9}$$

   In the case of similar failure probabilities for relatively short sections, this approach will give a less conservative outcome.

2. An alternative approach is to consider part of the variables to be dependent. For example, river discharges or storm conditions (wind speed and direction) are relatively uniform on a regional scale. Contrarily, strength related parameters such as subsoil characteristics are much more variable. In Chapter 5, an approach is presented that considers load to be dependent and strength to be independent. The failure probability can then be calculated by first calculating the failure probability conditional on the peak discharge k, $P_{f,A|k}$, for example by using Eq. (2.7) or Eq. (2.9). The total failure probability for several dike sections together is then calculated by integrating the conditional exceedance probabilities (i.e., fragility curves) with the probability density $f(k)$ of the peak discharge k:

$$P_{f,A} = \int_k f(k) P_{f,A|k} dk. \tag{2.10}$$

   When the dike sections are relatively sensitive to load, failure will only be possible once a certain load has been met. After this, the failure probability will quickly increase with the load. In such a situation, the approach ensures that failure will not take place until this minimum load has been met at, at least, one of the sections.

3. A third refinement is to consider the load reduction caused by an initial failure. If an upstream breach draws a significant part of the river's discharge, it will reduce downstream water levels. While this will not change the probability of at least one breach, since reduction only happens after the first breach, it will reduce the probabilities of additional failures and therefore the consequences. An approach that incorporates this is described in (Uemura et al., 2024). They define $P'_{f,A_2|k}$ as the

failure probability of section A_2, given peak discharge k, conditional to the upstream section A_1 not failing. Conditional to upstream failure and peak discharge, the failure probabilities can than be calculated as follows:

$$\begin{aligned} P'_{f,A_1|k} &= P_{f,A_1|k} \\ P'_{f,A_2|k} &= P_{f,A_2|k} \cdot \left(1 - P'_{f,A_1|k}\right) \\ P'_{f,A_3|k} &= P_{f,A_3|k} \cdot \left(1 - \left(P'_{f,A_1|k} + P'_{f,A_2|k}\right)\right) \\ P'_{f,A_n|k} &= P_{f,A_n|k} \cdot \left(1 - \sum_{i}^{N-1} \left(P'_{f,A_i|k}\right)\right) \end{aligned} \tag{2.11}$$

Because all the failure probabilities $P'_{f,A_i|k}$ are considered mutually exclusive (i.e., there is no overlapping area such as the double hatched area in Fig. 2.4), the total failure probability can be calculated as the sum of all the failure probabilities, conditional to peak discharge and upstream failure:

$$P_{f,A|k} = \sum_{i}^{N} P'_{f,A_i|k}. \tag{2.12}$$

Each of the above three examples present a different approach to account for dependence. The examples make an increasing number of assumptions. This results in more precise failure probabilities for the specific situations while simultaneously limiting the situations in which the approach can be applied.

On a final note, considering dependence in failure probabilities only has a substantial effect when the different failure probabilities are of a similar magnitude. If one of the sections is in a much poorer condition than the rest (e.g., has a 10 times higher failure probability), this weakest link will dominate the total failure probability. Consider for example Fig. 2.6, similar to Fig. 2.4 but now with one failure region 10 times large than the other. The strength of dependence between the variables now hardly affects the total number of events covered by at least one of the failure regions.

2.2.3. Quantifying dependence between continuous variables

The last section focused on combining exceedance probabilities with varying degrees of dependence. However, sometimes the pattern of dependence itself is of interest. The scatter in Fig. 2.6 could be observed waves and water levels, and we could for example want to define a model that, when sampled from, reproduces this scatter plot. Different statistical approaches are available to describe such a relation. In this thesis, the copula is applied to model the dependence, and the non-parametric Bayesian Network is used to elicit correlation strengths. Both are explained hereafter.

Copula

A copula is a multivariate cumulative distribution for which the marginals are uniform in the interval $[0, 1]$. The scatter plots in Fig. 2.6 are generated from Gumbel copulas. Figure 2.7 shows the probability densities of these copulas on the background of the scatter

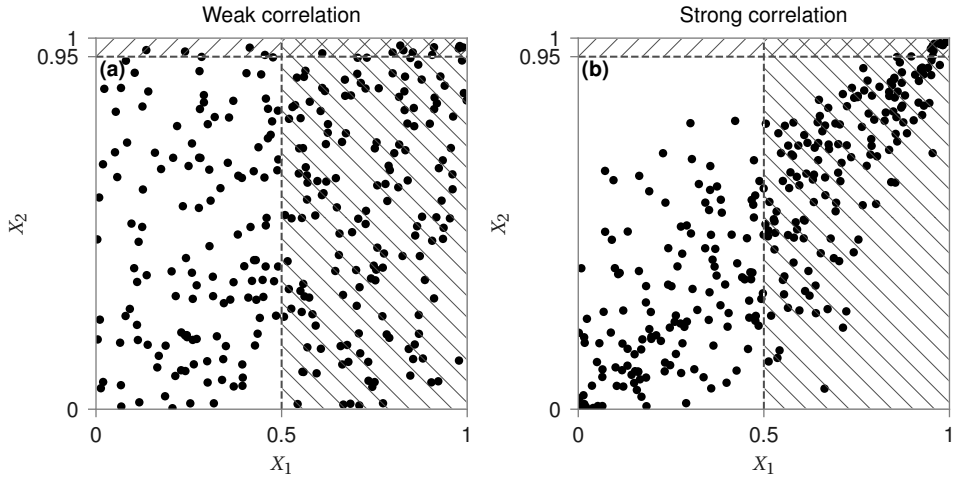


Figure 2.6: Scatter plots of weakly (a) and strongly (b) correlated variables X_1 and X_2, with a smaller effect of dependence due to the large difference in areas of failure.

plots. Because a copula's margins are uniform, they can be transformed to another distribution through that distribution's quantile function. Doing this for every marginal result in the joint probability distribution. An illustration of this is given in Fig. 2.8. Panel (a) shows the contour lines of the Gaussian copula's probability density function. Panel (b) and (c) show the marginal cumulative distribution functions. In this case, this is a Gumbel distribution for yearly maximum sea water level, and an exponential distribution for yearly maximum wind speed. Using the inverse of this cumulative distribution function (i.e., the quantile function), the $[0,1]$ interval is transformed to the physical space, resulting in the joint probability density function in panel (d).

Different types of copulas describe different dependence patterns across the unit square. Figure 2.8 showed the Gaussian copula, which is a bivariate or multivariate normal distribution, transformed to the $[0,1]$ interval using the normal distribution's cumulative distribution function (CDF). This copula can be used to model varying correlation strengths between different variable pairs, and conveniently facilitates the analytical calculation of conditional distributions. The Gaussian copula is however limited in its ability to differentiate dependence strength between parts of the distributions. This can be a disadvantage for modeling correlation between variables with strong dependence in the tails of the distributions; note that the correlation strength in Fig. 2.8 a (fairly strong) Pearson correlation coefficient of 0.75 was used. However, due to the marginal distributions, transforming to the physical scale results in a weaker correlation in the extreme domain than the frequent domain. Archimedean copulas are a class of copulas that can take many different forms and describe asymmetries in joint distributions. However, their ability to describe dependence between more than two variables is limited; they cannot model different dependence strengths for different pairs of margins of the joint distribution (e.g., model a stronger dependence between A and B than between B and C). The copulas in Fig. 2.7 are Gumbel copulas. These have a stronger correlation in the upper

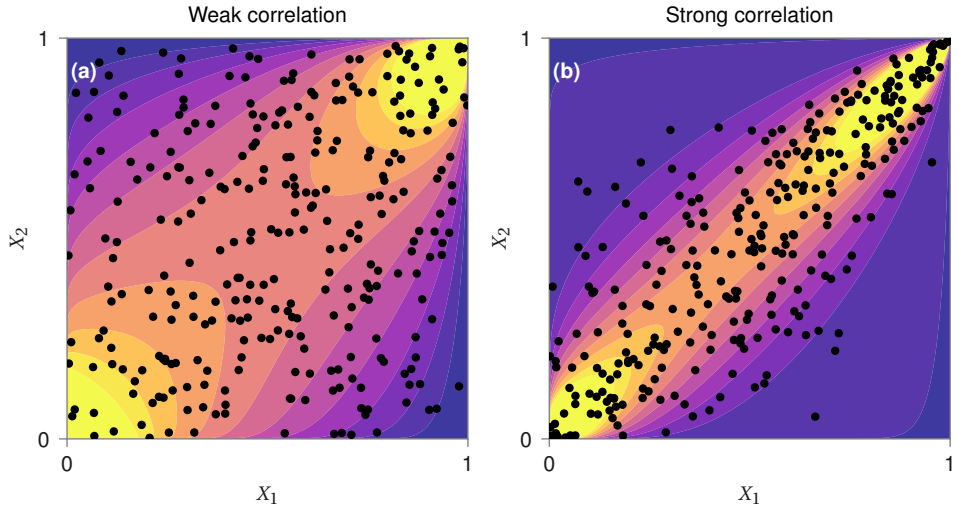


Figure 2.7: Scatter plots of weakly (a) and strongly (b) correlated variables X_1 and X_2, with their underlying Gumbel copulas. Light colored areas indicates high probability density, dark colored low.

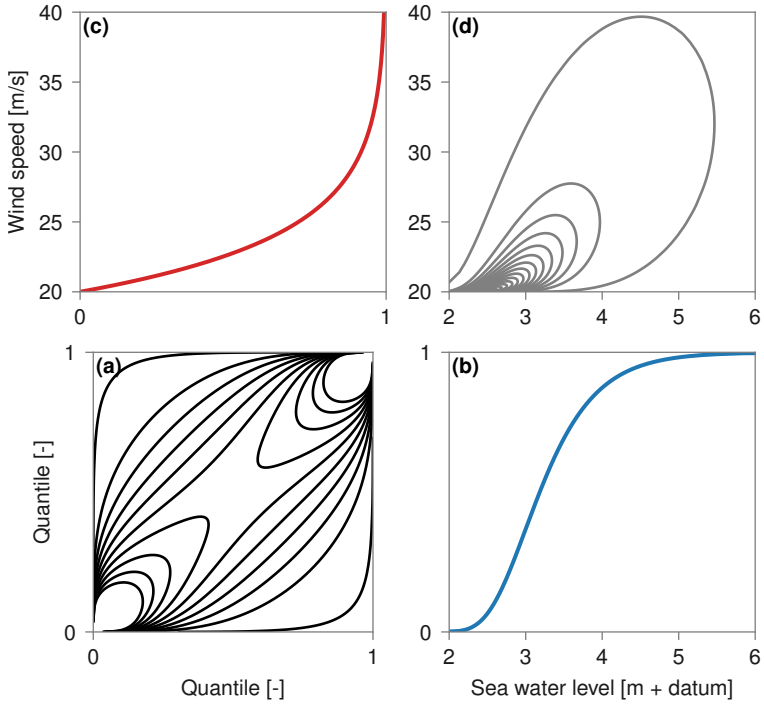


Figure 2.8: Conceptual explanation of a copula (a) and its relation to marginal statistics (b and c), and the joint probability distribution (d).

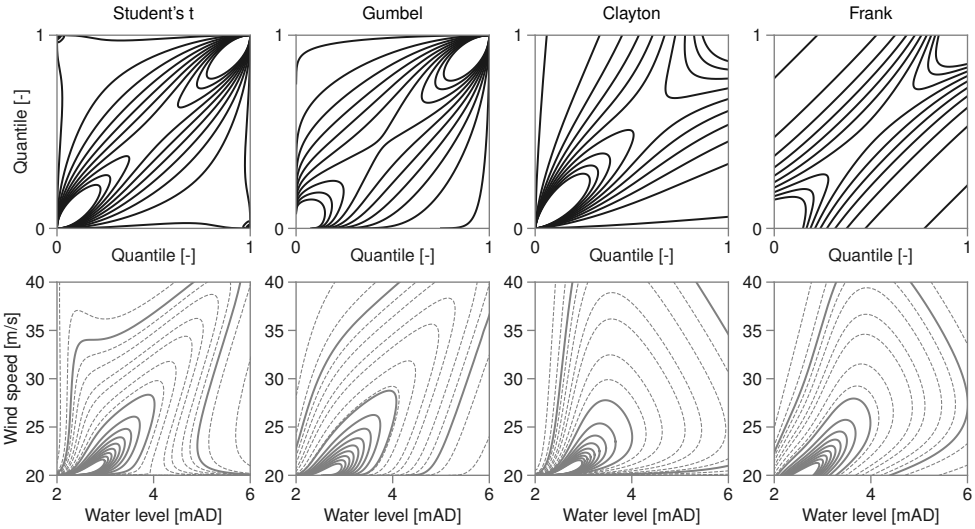


Figure 2.9: Contour lines of different copulas' PDFs on their $[0,1]$ scale (top row) and transformed to physical scale (bottom row).

tail compared to the lower tail, which becomes more apparent when transforming to a physical scale. Figure 2.9 shows a Student's t-copula (with degree of freedom 2), Gumbel, Clayton, and Frank copula, with model coefficients chosen such that the resulting scatter plots have the same Pearson correlation coefficient as the Gaussian copula from Fig. 2.8. This illustrates how the different models have a different tail dependence. All except Student's t-copula are Archimedean copulas.

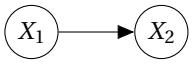
Correlation coefficients can be calculated from sufficiently large datasets, but with increasing number of variables the dataset size needs to increase as well in order to "cover" all regions of the unit-hypercube (i.e., the multivariate version of the unit squares shown in Fig. 2.9). In absence of such data, expert judgment can be used to estimate the correlation matrix. This is not a trivial task, firstly because of the numbers of coefficients to be estimated and secondly because not all combinations of coefficients in the matrix results in a valid (i.e., positive semi-definite) correlation matrix. The simplest example for this last problem is a correlation matrix with three variables X_1, X_2, and X_3, in which both pairs (X_1, X_2) and (X_2, X_3) are fully positively correlated (i.e., a coefficient of 1.0). In this case the pair (X_1, X_3) must then be fully dependent as well because they are related through X_2, with which they are both fully dependent. Any other value than 1.0 between X_1 and X_3 will result in an invalid correlation matrix. In case the correlations are strong, but not perfect, such conditions become less strict and less clear, but they still need to be satisfied to create a valid correlation matrix. Non-parametric Bayesian Networks provide a solution for this.

Non-parametric Bayesian Networks

A Bayesian Network (BN) is a directed acyclic graph (DAG) that represents the dependence between random variables through nodes and arcs(Darwiche, 2009; Pearl, 2000). See Fig. 7.2 for some examples of DAGs. In a BN, every node represents a random variable. These are generally discrete which makes the conditional probability functions, represented by the edges, conditional probability tables. As an example, consider a hypothetical Bayesian network that describes the dependence between X_1 testing positive for COVID and X_2 having it. This network is shown in Fig. 2.10. With both random variables having two states, the BN renders four conditional probabilities to be quantified:

1. having COVID conditional on testing positive, $P(X_2 = \mathrm{T}|X_1 = \mathrm{T})$,
2. having COVID conditional on testing negative, $P(X_2 = \mathrm{T}|X_1 = \mathrm{F})$,
3. not having COVID conditional on testing positive, $P(X_2 = \mathrm{F}|X_1 = \mathrm{T})$, and
4. not having COVID conditional on testing negative, $P(X_2 = \mathrm{F}|X_1 = \mathrm{F})$.

Bayesian network



Condition probability table

	X_2	
X_1	T	F
T	$P(X_2 = \mathrm{T}\|X_1 = \mathrm{T})$	$P(X_2 = \mathrm{F}\|X_1 = \mathrm{T})$
F	$P(X_2 = \mathrm{T}\|X_1 = \mathrm{F})$	$P(X_2 = \mathrm{F}\|X_1 = \mathrm{F})$

Figure 2.10: Bayesian network with two nodes (X1 and X2) connected by a single arc with its conditional probability table.

These probabilities are conditional because having COVID changes the probability of testing positive and, conversely, testing positive changes the probability that someone has COVID. The conditional probabilities shown in the table can be calculated using Bayes' theorem (hence the name of the network), which is explained for this example in Section 2.3.1.

In many models, random variables have more than two states or are continuous, requiring quantification of larger conditional probability tables. It can be challenging to quantify such networks, particularly when the network consists of more than two nodes and multiple arcs exist between them. The number of conditional probabilities to be assessed depends on the number of states of each node and the number of arcs incoming to a particular node (Druzdel & Van Der Gaag, 2000; Renooij, 2001). This increases rapidly with the number of states of the variables in the network (continuous variables can be discretized into several states). The more states considered in a discretization of a continuous variable, the better its representation, but also the larger the number of conditional probabilities to be quantified.

The non-parametric Bayesian network (NPBN) solves this by describing the (conditional) relation using a Gaussian copula. Each arc in a NPBN represents a (conditional) rank correlation, rather than a (conditional) probability table. Similarly to the usual BN, the

structure of the graph defines which child node is dependent on which parent node, and through that, the conditional (in)dependence between nodes. By specifying the (conditional) rank correlations for all edges, the resulting correlation matrix is valid. While NPBNs are based on Gaussian copulas, the marginal distributions of the random variables do not need to be normally distributed. As explained in Section 2.2.3, the multivariate normal distribution can be transformed to quantiles in the $[0,1]$ range, and to any random variables using its percentile function. This absence of a need to parameterize the marginals (since any invertible marginal distribution may be used) is what differentiates it from other types of BN and is why it is called a *non-parametric* Bayesian network. For a more formal and detailed explanation of NPBNs, as well as a description of some applications, refer to (A. Hanea et al., 2015).

2.3. Bayesian statistics

2.3.1. Bayes' theorem

Bayes' theorem, introduced by Thomas Bayes, is an equation used for calculating conditional probabilities. It provides the probability of an event or hypothesis given prior knowledge of the conditions around the event. Bayes' theorem is stated as

$$P(H|D) = \frac{P(H) \cdot P(D|H)}{P(D)}. \tag{2.13}$$

H denotes the hypothesis or variable of interest. D are the data or observations. $P(H|D)$ is the posterior probability, which expresses the probability of the hypothesis given that we have observed D. This posterior probability depends on

1. the initial, or *prior*, estimate of the probability of the hypothesis $P(H)$,
2. the probability of the data given the hypothesis $P(D|H)$, also known as the likelihood, and
3. the evidence or marginal probability, which is the sum or integral of all hypotheses multiplied by their likelihoods $\sum_i^n P(H_i)P(D|H_i)$.

Substituting these terms states Bayes' theorem as

$$posterior = \frac{prior \times likelihood}{marginal\ probability}. \tag{2.14}$$

To illustrate the use of Bayes' theorem, we calculate the probability of a person having COVID (H_1, the hypothesis), given that the person tests positive for COVID-19 (D, the data). Assuming:

- 5% of the population have COVID at the moment of testing. Without knowing anything else about the person, this would be our *prior* estimate, so $P(H_1) = 0.05$.

- The probability of the test turning out positive when having COVID is 95%, or $P(D|H_1) = 0.95$.
- The fraction of false positives is 2%, meaning that the probability of testing positive when not having COVID (H_0) is 2%, or $P(D|H_0) = 0.02$

The posterior probability can then be calculated as follows:

$$\begin{aligned} P(H_1|D) &= \frac{P(H_1) \cdot P(D|H_1)}{P(D)} \\ &= \frac{P(H_1) \cdot P(D|H_1)}{P(H_1) \cdot P(D|H_1) + P(H_0) \cdot P(D|H_0)} = \frac{0.05 \cdot 0.95}{0.05 \cdot 0.95 + 0.95 \cdot 0.02} = 0.714 \end{aligned} \tag{2.15}$$

According to Bayes' theorem and the assumptions, the probability of the person having COVID is 71% even though they tested positive. Perhaps the person finds this evidence not convincing and decides to do another test. Bayes' theorem can now be used again to update the posterior probability. After the last test, the person has updated their prior belief for $P(H_1)$ from 0.05 to 0.714. Doing another test, this posterior becomes the new prior. If the test result is positive again (and we assume the test results to be independent), the updated posterior probability is a convincing 99%. However, if the test is negative, the probability of having COVID reduces to 11%.

While this numerical example illustrates the usefulness of Bayes' theorem, the most important element is that the outcome, in this case the probability of having COVID, does not just depend on the data but also on the initial assumptions about it. More generally, a person's expectations do not just depend on what they observe but also on their beliefs about it. Consequently, a (Bayesian) probability expresses a degree of believe, rather than a physical property. This comprehensive view of statistics is what is known as Bayesian statistics. While its counterpart, frequentist statistics, revolves around drawing conclusions based on just the data, Bayesian statistics reasons from an initial idea or construct of the situation.

In practice, the two approaches are often intertwined and not clearly distinguishable. This thesis contains a few elements of Bayesian statistics that are presented in the remainder of this chapter. They are all Bayesian because they rely on Thomas Bayes' theorem, but the way they apply it, can be quite different.

2.3.2. Dutch flood risk as a Bayesian approach

In the "Fundamentals of flood protection", Kok et al. (2017) explain the view on probability in the context of Dutch flood risk. They describe the frequentist approach, in which a probability is a relative frequency (i.e., a number of observations in a number of years), as problematic due to the lack of knowledge. The flood probability is considered a physical property of which value we are uncertain because the lengths of observations are too short, and number of dike failures are too few. Therefore, using the frequentist approach, it becomes impossible to conclude with certainty whether the probability of flooding meets a particular standard (Kok et al., 2017).

Contrarily, in the Bayesian interpretation, Kok et al. describe "the probability of flooding [to be] a measure of the likelihood that a flood will occur, given the knowledge at our disposal". While the terminology differs a bit from Section 2.3.1, this can be interpreted as the posterior probability, $P(H|D)$, consisting of the hypothesis (a flood will occur) and the data (knowledge at our disposal). Consequently, the probability of flooding is not a physical property, but a degree of uncertainty. In terms of Dutch flood risk assessment, the most important difference between the frequentist and Bayesian approach is the treatment of uncertainty. A Bayesian approach considers a parameter's value to be a degree of believe, which is an uncertainty or probability distribution. The frequentist approach, on the other hand, considers them fixed but unknown values. A frequentist confidence interval can be defined, which (loosely said) indicates the range that contains the true value with (e.g.,) 95% certainty, but it is not a probability distribution.

Figure 2.11 illustrates the two different approaches. Both figures show an exceedance frequency curve of a specific water level in grey. The flood defense is assumed to fail when the water level exceeds 4 meters above datum (mAD). Moreover, the flood defense has a required safety standard of 1/1000 per year. Given this information, does the flood defense meet the standard?

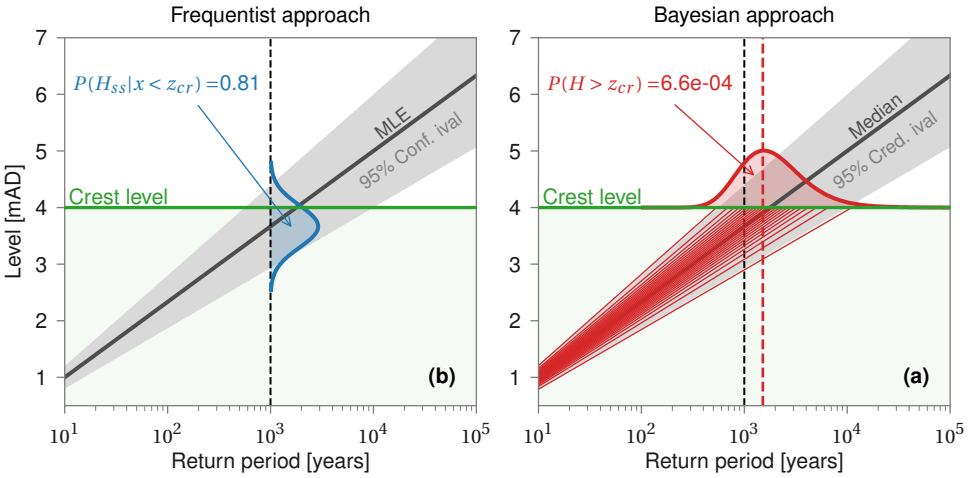


Figure 2.11: Illustration of a frequentist and Bayesian approach for assessing whether a dike fails (i.e., water level exceeds crest level) more often or less often than the safety standard.

The left panel shows the frequentist approach. A probability distribution is fitted to the observed water levels, which has led to a maximum likelihood estimate (MLE) and a confidence interval. The 1000-year water level from the MLE is less than 4 meters. The uncertainty is expressed through the confidence interval, which expresses the probability (confidence) that if the experiment (i.e., observing water levels and fitting a distribution) was repeated 100 times, the confidence interval from that experiment contains the true value of the model parameters 95 times. This is indicated with the blue distribution (assumed to be normal distribution). From this distribution, the probability that the 1000-year ARI water level H_{ss} from the true model parameter x is lower than the crest

level z_{cr}, is 81%. Whether the flood defense matches the safety standard then depends on the chosen confidence level. If a (typical) 95% confidence level is required, the flood defense would be considered not proven to be safe.

The right panel shows the Bayesian approach. Fitting the water level probability distribution with a Bayesian approach gives a probability distribution of model parameters, indicated by the grey *credible* interval. This interpretation of uncertainty gives the flexibility to assign probabilities to different model parameters. This is done in the example by defining 25 red lines that each represent an equal part (4%) of the probability distribution. By calculating the probability with which the crest level is exceeded, conditional to that model parameter, and averaging the results, the total failure probability is calculated: $P(H > z_{cr}) = \int_x f(x)P(H > z_{cr}|x)dx$. The result is $6.56 \cdot 10^{-4}$ (in a year), less than the maximum allowed failure probability according to the safety standard.

The frequentist approach puts the model central. For example, the confidence interval indicates the range that contains the *true model parameter* 95 times, if the experiment is repeating 100 times. The Bayesian approach puts the observations central. There is no true model parameter, only true observations. The credible interval consequently indicated the distribution of model parameters given those observations. The latter allows to integrate the uncertainty into the failure probability, where the frequentist approach can only express a degree of confidence that the true model parameter is captured. Despite these conceptual differences between the two approaches, confidence intervals and credible intervals do often span the same range.

An important aspect of Bayesian statistics are the priors. These do not have to be explicit a-priori distributions but can also be expressed by choosing a modeling approach that is considered better or more suitable. Such priors will differ per person, which complicates a flood probability assessment as the result should ideally not depend on an individual's specific ideas. Kok et al. mention that "In practice, such differences can be overcome by exchanging data, second opinions and the establishment of best practice."

2.3.3. Bayesian sampling techniques

While Bayes' theorem is easy to apply for examples such as described in Section 2.3.1, it quickly becomes more complicated when considering more complicated models or updating probability distributions. When an analytical solution is infeasible (or one lacks the mathematical skills to work it out), sampling techniques such as Markov Chain Monte Carlo (MCMC) provide a solution.

MCMC is a class of algorithms used to sample from probability distributions, useful when an analytical solution is challenging or infeasible to obtain. MCMC is a Monte Carlo technique, meaning it involves random sampling. It does that using a Markov process, which is a stochastic process in which the probability of the future states (i.e., the next sample realization) only depends on the current state. Moreover, the probabilities of transitioning to a the next states (model parameter combination) are proportional to the posterior distribution. This results in a chain of equally probable parameter combinations. After correcting the chains for spin-up and auto-correlation, they contain the

posterior probability distribution of the model-parameters, as well as dependence between them.

To illustrate how MCMC works we use the Metropolis-algorithm (Metropolis et al., 1953). MCMC jumps from one state to the next by comparing Bayes' theorem for the current and next combination of model parameters. $P(H)$ would be the prior probability of the model parameter(s) H, and $P(D|H)$ would be the likelihood of the observations given a combination of model parameters. $P(D)$ is difficult to calculate as it involves integrating the likelihood for all possible model parameters, but it is also independent of the proposed H. Therefore, it can be crossed out as a normalization constant, leaving:

$$P(H|D) \propto P(H) \cdot P(D|H). \tag{2.16}$$

The Metropolis algorithm compares the posterior probability for two parameter combinations:

$$\frac{P(H_{new}|D)}{P(H_{current}|D)} = \frac{P(H_{new}) \cdot P(D|H_{new})}{P(H_{current}) \cdot P(D|H_{current})} \tag{2.17}$$

If this fraction is larger than 1.0 (i.e., the proposed combination H_{new} has a higher posterior probability than $H_{current}$), it is adopted as the current combination. This means the algorithm will converge to the most likely parameter combination, similar to a maximum likelihood estimate. However, to explore the full distribution, it also needs to be able to transition to less likely combinations. Therefore, if the fraction is smaller than 1.0, a random number is drawn between 0 and 1. The less likely combination is then only adopted if the fraction (Eq. (2.17)) is greater than this number. This allows the algorithm to explore the full parameter space, while transitioning to new combinations such that the resulting chain of visited parameters is proportional to the posterior probability.

While MCMC allows the use of priors ($P(H_{new})$ and $P(H_{current})$ in Eq. (2.17)), they are not necessarily required. If the preference is to not express any information through the prior, a non-informative prior is chosen. This is a distribution that results in a uniform posterior distribution when sampled from. Regardless of the prior, the results of the MCMC analysis will be a Bayesian result, meaning that the distribution of parameter combinations are interpreted as an uncertainty or degree of belief. Similarly, a prior probability of a model parameter can be considered in a maximum likelihood estimate as well. While it does not make the interpretation of the results Bayesian, it does incorporate the prior beliefs in the outcome, illustrating the cross-over between Bayesian and frequentist probability.

Figure 2.12 illustrates the chain that is generated by MCMC from one of the GEV distributions fitted in Chapter 6). The grey line shows 1000 steps from a single chain's trace. To avoid auto-correlation, only once every 25 steps a combination is recorded. These are shown by the dots, projected as circles on the three axes planes. The joint distribution is found by combining the trace from several chains (e.g., 50) with different starting points across the distribution, and discarding the first steps as spin-up. This is shown with the (kernel density smoothed) contours projected on the planes, which are based on roughly 10,000 combinations. The dependence, in this case strongest between the location and scale parameter, follows from the "empirical" trace.

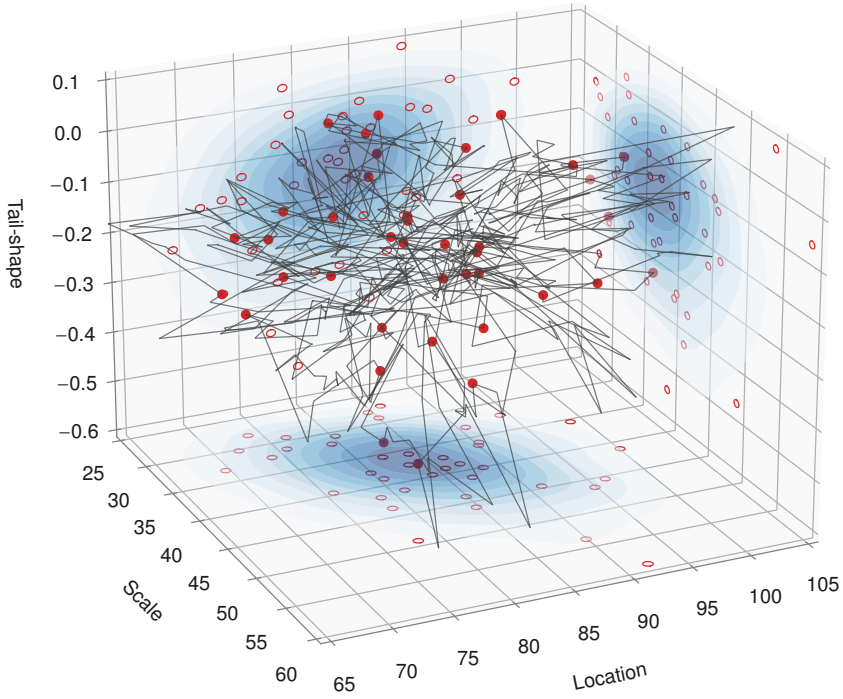


Figure 2.12: Example of an MCMC trace from fitting a generalized extreme value distribution to observed annual maxima.

2.3.4. Expert judgment as a Bayesian prior

Structured expert judgment is a suitable technique to provide priors for a Bayesian analysis. In Bayes' theorem (Eq. (2.13)), the prior expresses the initial belief about parameters before any data are observed. Structured expert judgment elicits experts' beliefs about a topic of interest, typically also without the data being observed. In the context of Bayesian statistics, expert judgments can thus be treated as prior probabilities. In scenarios where data are limited, (structured) expert judgments can be used to derive priors in a scientific manner, such that an informative posterior distribution can still be obtained.

Priors can be used to inform model parameter or model outcomes. The difference between these two approaches was assessed in Chapter 5 (not in a Bayesian context). Chapter 6 uses a prior on the model outcome to inform extreme discharges. When a prior probability is considered for a model outcome, the process presented in Section 2.3.3 involves an extra step in which the model outcome is calculated for the proposed parameter combination H. This is then compared to the prior estimate for that outcome, yielding a prior probability (i.e., not $P(H)$ but $P(f(H))$). This is explained in Section 6.3.3. The remainder of the sampling process is the same.

The use of expert judgment-derived priors is well represented in research, for example

by (Renard et al., 2006) or (Viglione et al., 2013). Software such as RMC-Bestfit provide a graphical user interface and absolve the need for programming, at least for the purpose of fitting extreme value distributions to observations. For more elaborated models, one can use software modules such as PyMC3 (Abril-Pla et al., 2023), emcee (Foreman-Mackey et al., 2013) (both Python), or STAN (Carpenter et al., 2015), which has a bespoke syntax for probabilistic programming.

3

Software for expert elicitation

Two pieces of software were developed in this research to assist researchers and problem owners in doing expert elicitations. The first, ANDURYL, is an aid for eliciting univariate uncertainties using structured expert judgment. The second, Matlatzinca, helps with eliciting statistical dependence using Non-Parametric Bayesian Networks. Both are open source and freely available software that can be found on GitHub. This chapter gives an overview of both pieces of software. For more details, please refer to the articles in which they were published.

The contents of Section 3.1 have been published in: Rongen, G., 't Hart, C.M.P., Leontaris, G., & Morales-Nápoles, O. (2020). Update (1.2) to ANDURIL and ANDURYL: Performance improvements and a graphical user interface. *SoftwareX, 12*, 100497.

The contents of Section 3.2 have been published in: Rongen, G., & Morales-Nápoles, O. (2024). Matlatzinca: A PyBANSHEE-based graphical user interface for elicitation of non-parametric Bayesian networks from experts. *SoftwareX, 26*, 101693.

3.1. ANDURYL

Software implementing Cooke's classical model (Cooke, 1991) for structured expert judgment has been available through a program called Excalibur. It has a graphical use interface and is freely available. However, it is not open source, making it difficult to further develop the (relatively old) piece of software. This limits the usefulness for research applications because new methods cannot be tested through the program and recalculating historical case-studies is a tedious manual task. To resolve this, Leontaris and Morales-Nápoles (2018) created a MATLAB program called ANDURIL[1]. This program provided the functionality to process elicitation results from the Classical Model but did not have a user interface. Following this, 't Hart et al. (2019) created a Python version of ANDURIL, named ANDURYL (note the Pythonic "Y"). The advantage of Python over MATLAB is that it is freely available, removing another obstruction for using software for expert elicitation. However, the lack of a graphical use interface still required users to have programming skills. This was resolved be creating a graphical user interface (GUI) for the Python version (Rongen et al., 2020), which was done in the context of the research presented in this thesis. This interface, programmed with the Python module PyQt5, and can be compiled with PyInstaller (for Windows) resulting in a stand-alone executable. This makes ANDURYL accessible to non-Python users.

An overview of the ANDURYL GUI is shown in Fig. 3.1, Fig. 3.2, and Fig. 3.3. These figures show the different elements of the Classical Model:

- In the top left panel, an overview of the experts is shown. The bottom row of the panel shows a decision maker, a weighted combination of experts.
- The lower left panel shows the items for which estimates are made. Some have a realization. These are the seed questions that are used to assess the expert's statistical accuracy. The other items without realizations (lower in the list) are the variables of interest.
- In the Classical Model, uncertainties are estimated through a number of quantiles or percentiles. The upper right panel shows some of the estimates of the first expert for the different quantiles. The colors indicate how these estimates compare to, if these are known, the realizations or otherwise the total estimated range for the question.

[1] The name ANDURIL refers to "A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty" but you might recognize it as the name of a legendary sword, just like Excalibur. Leontaris and Morales-Nápoles (2018) explain the origins of the name in a comical manner that also includes the motivation, which is why it is quoted here: "In order to avoid confusion of the minority of people, who are not familiar with the universe of Lord of the Rings by J.R.R. Tolkien, the authors would like to clarify the inspiration for the name of the developed Matlab toolbox. Andúril was the name of the sword of Aragorn, the son of Arathorn, which was reforged from the shards of Narsil (the sword that was used by Isildur to cut the One Ring from Sauron's hand). Excalibur is also the name of the legendary sword of King Arthur. Similarly to the sword, the source code of EXCALIBUR software remained accessible only to a few worthy ones. Therefore, the researchers and practitioners could only admire and use the software without being able to further investigate and explore developments of the method. To change this, the existing software had to be 'broken to pieces' and then 'reforged'. Naturally, the name of the resulting new open-source Matlab toolbox is ANDURIL. Hopefully, this will help in bringing peace to troubled researchers and practitioners of Cooke's classical model."

- The lower right panel shows calculated decision makers. Different methods of calculating a decision makers will give different results.
- Distributions (CDF, survival function, PDF, or range) can be plotted per item, for several experts. This is shown in Fig. 3.2. Alternatively, results can be grouped by expert, showing the range of the estimates for all items.
- Robustness results can be visualized for a number of excluded experts or items, as shown in Fig. 3.3. The boxplot shows the statistical accuracies or information scores that are calculated for excluding any combination of one or more experts.

Additional to the user interface, a number of improvements were made on the back end, compared to the version from ('t Hart et al., 2019).

- In (Colson & Cooke, 2017a), 33 post-2006 studies using the Classical Method are presented using CC. These data were used to compare output from ANDURYL to Excalibur. Due to a new implementation for combining expert CDFs into a decision maker, differences between ANDURYL and Excalibur were reduced compared to the last code version ('t Hart et al., 2019). For two of the 33 studies, "Hemophilia" and "Ice sheets", there are still differences in calculated statistical accuracy and information score. It is unknown what causes these differences, as they cannot be traced through the Excalibur code. For four other studies there are small differences, seemingly from rounding errors. For the remaining 26 studies, the results are equal.
- To assess the sensitivity of the study to the experts and questions that are included, a robustness analysis can be done. With 10 experts there are only 10 options for excluding a single expert. This number quickly grows when multiple experts (or items) can be excluded, with 55 options for excluding two experts, and 175 for excluding three. The improved computational performance makes it less demanding to do robustness analysis for a large number of experts.
- The option was added to vary the overshoot and bounds for each item. The elicited percentiles can be varied as well between target items. For the seed items, the elicited percentiles need to be consistent for statistical accuracy calculations.
- Exporting and importing options. The program has options for loading and saving the project in EXCALIBUR format (for compatibility) or as in a more common JSON format. Furthermore, tabular results or distributions as shown in Fig. 3.1 and Fig. 3.2 can be exported to xlsx or csv, or copied to clipboard.
- The ANDURYL code is separated between calculation and user interface functionalities so that the Python-module can also be used from a script or Jupyter notebook. For research purposes this is a useful functionality. An example of such a notebook is shared on notebook to help a user to get started.

Finally, for the research described in Chapter 4, options were added to use different measures of statistical accuracy, and the Metalog distribution rather than the piecewise uniform distribution. These options are available in a separate branch on GitHub.

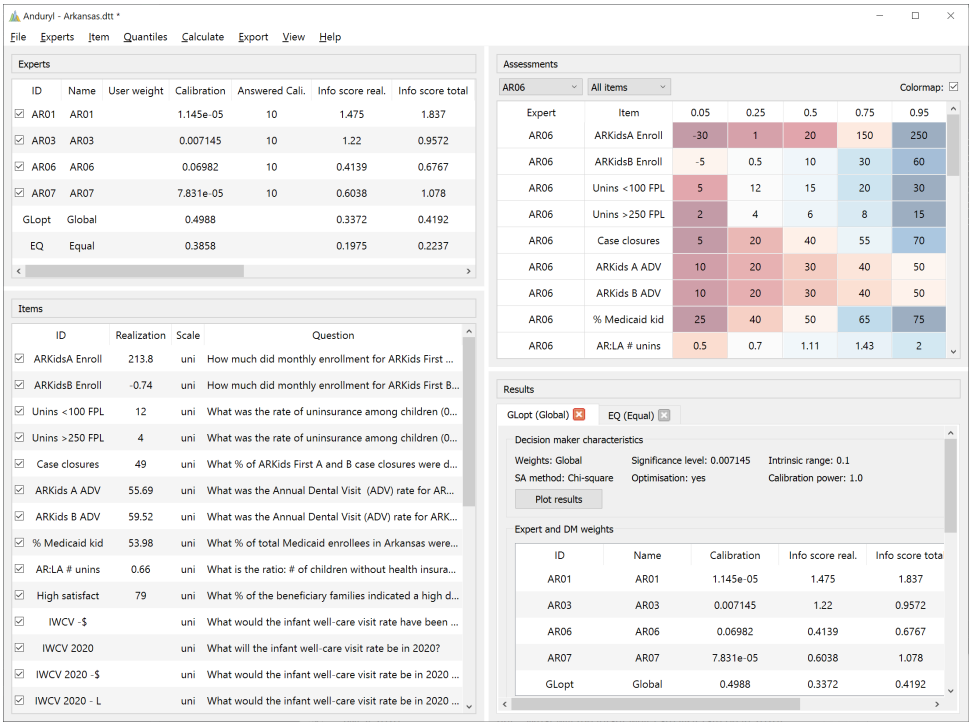


Figure 3.1: Overview of the ANDURYL GUI, with on the background the main window and on the foreground the CDF of each expert and the DM for a specific question.

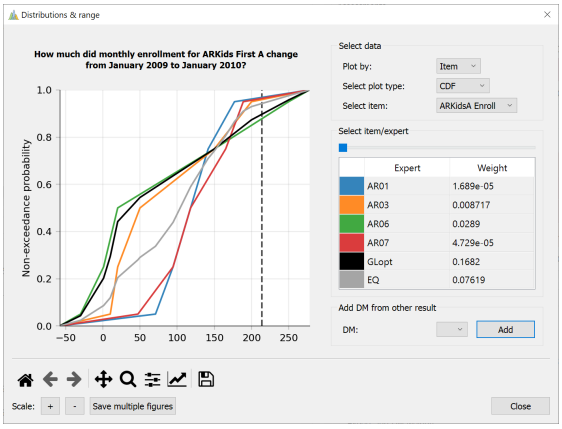


Figure 3.2: Viewing expert and decision maker estimates for separate items.

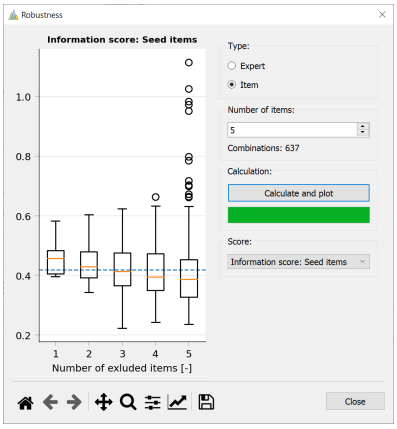


Figure 3.3: Viewing robustness of elicitation to excluding experts or items.

3.2. Matlatzinca

Matlatzinca is a graphical user interface (GUI) that was developed to aid the dependence elicitation described in Chapter 7. The mathematical complexity of multivariate dependence models makes direct elicitation difficult. For example, as explained in Section 2.2.3, not every correlation matrix is a valid one. Using a non-parametric Bayesian Network largely resolves this issue, but it still puts some constraints on the correlations between different variables. Ideally, experts are not burdened with these mathematical limitations during an elicitation and can focus on conveying their knowledge into the dependence model. Matlatzinca[2] was designed to do just that.

Matlatzinca is the first version of a stand-alone, open source, Graphical User Interface (GUI) for PyBANSHEE Koot et al., 2023; Paprotny et al., 2020. PyBANSHEE is a Python module that can be used to quantify, process, and sample from non-parametric Bayesian Networks (NPBNs) A. Hanea et al., 2015. Matlatzinca uses functionalities from PyBANSHEE to calculate conditional correlations. Additionally. Two computational methods that are not present in PyBANSHEE but were added to facilitate expert elicitation, are:

- The option to enter a non-conditional correlation and get the resulting rank correlation (instead of vice versa).
- Determine the limits for this correlation coefficient, to get a valid $(-1,1)$ conditional rank correlation.

Figure 3.4 presents a screenshot of Matlatzinca. The screenshot presents an example of a Bayesian Network (BN) for hydrological modeling quantified by one of the seven experts that participated in a recent application for estimating extreme river discharges in the Meuse river Chapter 7. The GUI consists of three panels:

- The drawing panel Fig. 3.5 left). This is where the Directed Acyclic Graph (DAG) that represents the dependence structure of the BN is drawn. Notice that arcs provide information with regard to the ordering of parents in the DAG, according to the protocol discussed in A. Hanea et al., 2015. That is, which arcs represent conditional rank correlations and the respective conditioning sets.
- The input panel (Fig. 3.5 bottom right). This panel presents the `Nodes` (representing random variables) used in the drawing panel. It also shows the `Edges` that the user must quantify. Two options are available for quantifying arcs, 1) specifying Spearman's conditional rank correlations (`Conditional rank corr.`) or 2) specifying non-conditional rank correlations (`non-conditional rank corr.`). Notice that the interval in which non-conditional rank correlations may vary is also provided by the software (`Range non-cond. rank corr.`). This interval

[2] Matlatzinca may be translated from Nahuatl (the language of the Aztecs) to English as "The people that make nets". This is the name that the Aztecs gave to the inhabitants of the Valley of Toluca in central Mexico who were well known fishermen at the time. Our first release of a GUI for PyBANSHEE is meant for people that wish to quantify (make) Non-parametric Bayesian Networks (Nets) with expert judgments. Hence, we name our GUI Matlatzinca.

depends on the structure of the DAG and the value of the correlations input by users in the ancestors' arcs. Notice that the first parent of a node corresponds to a non-conditional rank correlation, which may take values in $(-1,1)$. *Conditional* rank correlations may also take values in this interval. However, the interval for *non-conditional* rank correlations is restricted by the conditional independence statements embedded in the DAG and the restrictions of the correlation matrix itself. The column `Range non-cond. rank corr.` is updated as users introduce values for either conditional or non-conditional rank correlations.

- The correlation matrix panel. The correlation matrix of the model is shown in the upper right panel of Fig. 3.5. The magnitude and direction of the individual correlation coefficients are represented by circles of different diameters and colors (a colormap may be chosen by the user).

- A separate window that allows the user to relate conditional probabilities and rank correlations can be accessed through `Plot conditional probabilities`. Figure 3.5) shows this, with the conditional probability $P(F_{X_1}(x_{1,q}) > q | F_{X_2}(x_{2,q}) > q)$ for three user-defined percentiles (q) in the distributions of two random variables X_1 and X_2. The relation between rank correlation coefficient and the random variables under the Gaussian copula assumption is presented by (Morales et al., 2008). This window is intended to help users relate rank correlations and conditional probabilities in case they prefer one or the other for quantifying their models.

The different panes are interactively connected. Clicking a node or edge in the left pane highlights the corresponding rows in the tables on the lower right. A project can be saved and loaded from the `File`-menu. Display properties can be adjusted in the `View`-menu, and the `Export`-menu contains options of exporting the nodes, edges, or correlation matrix to CSV or the clipboard. Finally, the documentation, including a Quick-start description are accessible via the `Help`-menu.

Matlatzinca is the first version of a GUI built around PyBANSHEE. Unlike software for elicitation of univariate uncertainty such as ANDURYL, Matlatzinca does not allow for evaluating expert performance and combining multivariate uncertainty into decision makers.

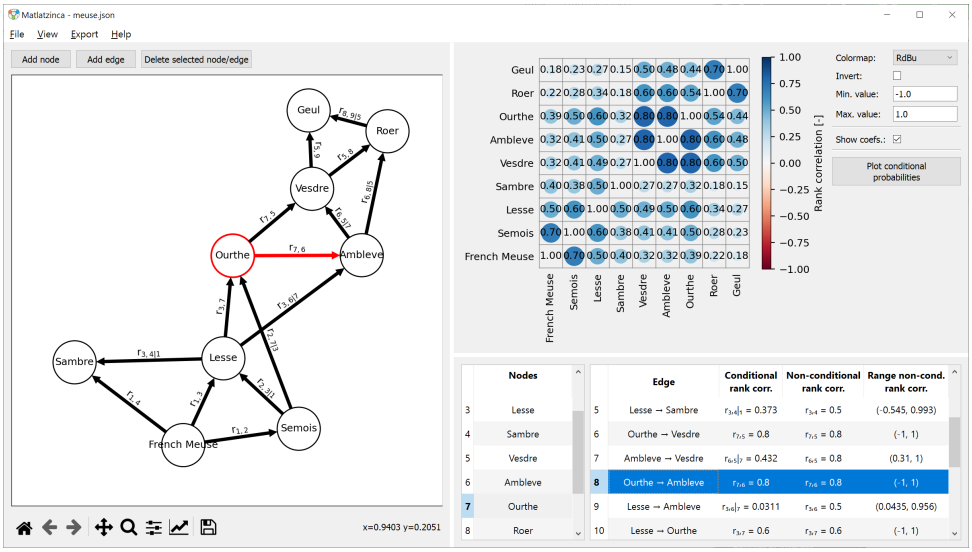


Figure 3.4: PyBANSHEE GUI for elicitation of Non-Parametric Bayesian Networks from experts: Matlatzinca. On the left, the drawing panel. On the top-right, the correlation matrix panel. On the bottom-right, the input panel.

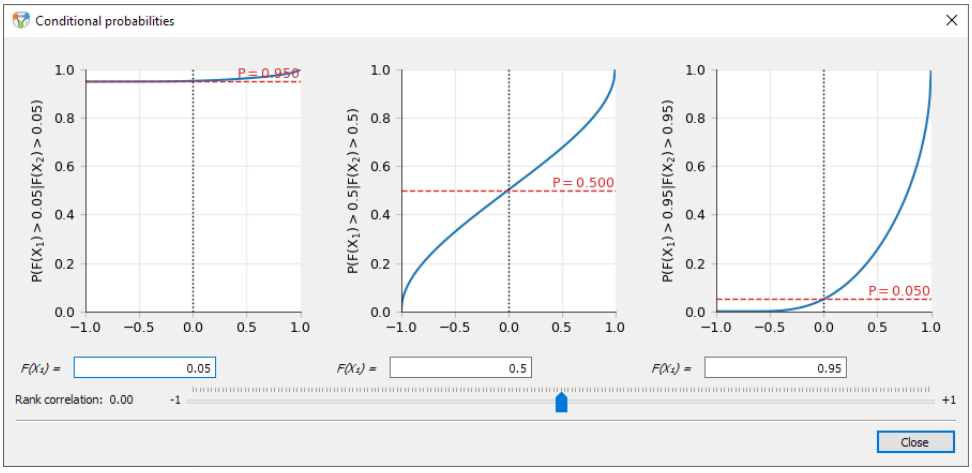


Figure 3.5: Conditional probabilities of interest as a function of rank correlation available in the matrix panel of Matlatzinca

4

Continuous distributions and measures of statistical accuracy for structured expert judgment

This study evaluates five scoring rules, or measures of statistical accuracy, for assessing uncertainty estimates from expert judgment studies and model forecasts. These rules — the Continuously Ranked Probability Score (CRPS), Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), Anderson Darling (AD), and chi-square test — were applied to 6864 expert uncertainty estimates from 49 Classical Model (CM) studies. We compared their sensitivity to various biases and their ability to serve as performance-based weight for expert estimates. Additionally, the piecewise uniform and Metalog distribution were evaluated for their representation of expert estimates because four of the five rules require interpolating the experts' estimates. Simulating biased estimates reveals varying sensitivity of the considered test statistics to biases. Expert weights derived using one measure of statistical accuracy were evaluated with other measures to assess their performance. The main conclusions are (1) CRPS overlooks important biases, while CM and AD behave similarly, as do KS and CvM. (2) All measures except CRPS agree that performance weighting is superior to equal weighting with respect to statistical accuracy. (3) Neither distributions can effectively predict the position of a removed quantile estimate. These insights show the possibilities and limitations of different scoring rules for combining uncertainty estimates from expert or models.

The manuscript related to this chapter has been submitted for publication at the time of writing. Authors: Rongen, G., Nane, T., Morales-Nápoles, O., & Cooke, R.

4.1. Introduction

Uncertainties are both widespread and influential in many fields, from climate modeling and economic forecasting to engineering design and legal decisions. The ability to accurately quantify uncertainties is important for informed decision-making, and it will often increase the value and usefulness to the outcomes. However, limited availability of data, complexity of the problem, financial or even ethical constraints can limit possibilities to accurately quantify these uncertainties. Expert judgment is a method to quantify uncertainty for variables whose uncertainty is difficult to quantify through other means. Comparing it to a statistical or physics-based models, expert judgment provides data in settings where a model would require assumptions or an extrapolation. It can take informal forms, such as asking an experienced person for their expectations, which might be fine for non-critical issues but high-stakes situations demand a more structured approach, one that is replicable, subject to review, and that could be assessed for potential biases, ensuring reliability and integrity.

The Classical Model (CM) is such an approach, which formalizes the process of expert judgment elicitation in such a way that the resulting uncertainty estimates can be treated as scientific data. It combines expert estimates using weights that are based on comparing uncertainty estimates to known outcomes of a number of check (or calibration) questions. Colonna et al. (2022) recently applied the method to combine COVID-forecasting models. They interpreted the different models as experts and their forecasts as estimates. The Classical Model was used to evaluate and combine these forecasting models, by comparing them to the actual course of events. This shows the the value of CM outside the typical field of expert judgment. First presented in (Cooke, 1991), the method has been widely applied and data from these applications have been made available to researchers, first in (Cooke & Goossens, 2008) and most recently in (Cooke et al., 2021). The later reference also gives a light exposition of the CM and introduces the expert judgment data used in this study. A special issue of Reliability Engineering and System Safety hosting the first publication of expert data (Cooke & Goossens, 2008) also contained contributions of many statisticians, risk analysts and practitioners who raised issues regarding CM[1]. Some of these issues, such as in-sample validation and overconfidence were amply addressed in the discussion papers of that special issue. Other concerns, most notably out-of-sample validation, persistence of performance and point value forecast performance spawned a stream of research. Much of this is summarized in (Cooke et al., 2021) and in the references therein.

An important aspect of a structured expert judgment exercise is the understanding of the sensitivity of its results to the number of experts and questions. The effort required to elicit information from experts means that we are never fully certain of each expert's statistical accuracy, underscoring the need for measures of statistical accuracy that utilize the information provided by the experts in the best way. In other words, we need to

[1] The contributions of Bram Wisse, Tim Bedford, John Quigley, Sandra Hoffmann, Paul Fischbeck, Alan Krupnick, Michael McWilliams, O. Morales, D. Kurowicka, A. Roelen, Shi-Woei Lin, Vicki M. Bier, Thomas A. Mazzuchi, William G. Linzey, Armin Brunin, Jouni T. Tuomisto, Andrew Wilson, John S. Evans, Marko Tainio, Roger Cooke, ElSaadany, Xinzheng Huang, Robert Clemen, Anthony O'Hagan and Simon French are gratefully acknowledged.

determine expert weights accurately based on a limited dataset, such that they reflect an expert's relative weight within a panel.

This study aims to explore two main questions: 1) How do different goodness of fit tests, each with known asymptotic or exact distributions, compare in evaluating expert estimates? And, 2) how do two approaches of interpolating a continuous CDF compare in representing expert estimates? Recent work uses the Continuous Ranked Probability Score (CRPS). Nane and Cooke (2024) present a CRPS-based score that assigns a scalar value to each assessment *cum* realization. Under suitable transformation, these scores for individual variables can be summed such that the exact distribution of the sum is available in closed form. This yields a measure of SA which appeals to an interpolated CDF but not to an asymptotic distribution. In total, we compare five different test statistics,

1. the standard χ^2 test in the CM,
2. the CRPS-based statistic,
3. the Kolmogorov-Smirnov (KS) statistic,
4. the Anderson-Darling (AD) statistic, and
5. the Cramer-von Mises (CvM) statistic,

All but χ^2 compare quantiles to continuous CDFs and therefore require a distribution for transforming realizations to quantiles, using the expert estimates. For this we use two classes of distributions. The first is the piecewise uniform (PWU) distribution corresponding to the minimum information assumption in the Classical Model. The second is the Metalogistic, or Metalog, distribution (Keelin, 2016). This recently introduced distribution offers great shape-flexibility, which helps with fitting a probability distribution to the large variety of expert quantile-estimates. Low parameter probability distributions often yield poor fits in these cases.

The five measures of statistical accuracy and two classes of distribution are compared in a variety of analyses, based on two different data sets. All analyses were done using Anduryl, an open-source Python-module and graphical user interface (Rongen et al., 2020). Metalog calculations were conducted using (Adamczewski, 2023). We used 49 expert judgment studies from the past decades, described in (Cooke et al., 2021), comprising 530 experts and 580 calibration variables. Recently published structured expert judgment studies, such as (Rongen et al., 2022a) and (Ren et al., 2024), were not considered because they have not yet been described and compared in an overview study. Additional to the published studies, we simulate expert estimates from distributions with a specific bias for a more clinical comparison. The analyses show

1. the statistical accuracy results from each score,
2. the ability of each measure of statistical accuracy to detect different biases, and
3. how the weights from each measure of statistical accuracy perform when used to create a DM that is evaluated with another measure.

Results are presented for the PWU distribution and the Metalog distribution. Finally, we consider the case-studies with 5 percentile estimates, removing 2 of these 5 percentiles, and see how well both distributions are able to estimate the position of the missing percentile.

4.2. Methods

4.2.1. Measures of statistical accuracy

Several statistical tests judging whether a sample corresponds to a probability distribution are of interest. Three such statistics are applied in this study: the Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), and Anderson-Darling (AD). Additionally, we include the CM chi-square and the recently introduced transformed Continuous Ranked Probability Score.

Classical model – chi-square

If k quantiles are assessed, with n the number of calibration variables assessed by an expert and n_i the number of realizations falling in the i-th quantile interval, then $s = (s_1, \ldots, s_{k+1})$, where $s_i = \frac{n_i}{n}$ is the sample distribution for the expert. The vector $p = (p_1, \ldots, p_{k+1})$ is the expected relative frequency of interquantile realizations, thus if the 5%, 50%, 95% quantiles are elicited, then $p = (0.05, 0.45, 0.45, 0.05)$. Under the hypothesis that the realizations are independently drawn from p, the quantity $2nI(s|p)$ is asymptotically chi-square (χ^2) distributed with k degrees of freedom, where $I(s|p)$ is the Shannon relative information of s with respect to p (Cooke, 1991). While the resulting statistical accuracy is based on the χ^2-distribution, the scoring rule used in the Classical Model is different from the commonly known χ^2-test.

CRPS

The Continuously Ranked Probability score (CRPS) is a measure for comparing forecast or estimates to realizations (T. A. Brown, 1974). Nane and Cooke (2024) present a scale invariant version of this measure. A closed form for the convolutions of scores (i.e., for multiple elicited items) is derived, which enables evaluating experts' statistical accuracy. For a given expert's distribution F for a random variable X and realization y, the Continuous Ranked Probability Score (CRPS) is defined by

$$CRPS(F, y) = \int_{-\infty}^{\infty} \left[F(x) - 1_{\{x \geq y\}}\right]^2 \mathrm{d}x.$$

We want to test the hypothesis that $F(X) \sim U[0,1]$. For this, we consider $CRPS(F_U, v)$, where F_U is the standard uniform distribution and $v = F(y)$, with y the realization. A certain transformation of the CRPS score leads to known distribution (of a squared uniform random variable). Moreover, the transformed CRPS score becomes scale invariant, that is, the score does not depend on the scale on which the quantity of interest is measured. Furthermore, if we assume n independent variables, then the distribution of the

convoluted transformed CRPS score follows an exact rather than an asymptotic distribution (Nane & Cooke, 2024). The details of computing the transformed CRPS score can be found in (Nane & Cooke, 2024). Throughout the manuscript, we will refer to this transformed CRPS instead of the original CRPS score (T. A. Brown, 1974).

Kolmogorov-Smirnov

The Kolmogorov-Smirnov (KS) test compares two samples (two-sided test) or a sample with a distribution (one-sided test) by using the supremum distance (equation (4.1)) between (empirical) cumulative distribution functions (Kolmogorov, 1933; Smirnoff, 1939)

$$D_n = \sup_x |F_n(x) - F(x)| \tag{4.1}$$

In the context of the Classical Model (CM), a perfectly statistically accurate expert is one for whom the quantiles of the realizations for the calibration questions are uniformly distributed. An expert's statistical accuracy is thus tested by comparing these quantiles to a uniform distribution using the one-sided KS-test. The arrow in Fig. 4.1 illustrates the KS test-statistics. In the KS test, the largest difference tends to be found near the median. Consequently, the test statistic is relatively insensitive to deviations in the tail, which, when applied to expert judgments, typically correspond with overconfidence.

For hypothesis testing, the KS distance is used to investigate the probability that the sample comes from the tested distribution. For this, an exact distribution is approximated using the method proposed by Simard and L'Ecuyer (2011). In classical statistics, a probability lower than 0.05 (i.e., the significance level) leads to rejecting the hypothesis that the data is independently sampled from the distribution of interest.

Cramer-von Mises and Anderson-Darling

The Cramer-von Mises (CvM) statistic is the area between the empirical CDF and target CDF (Cramér, 1928; Von Mises, 1928), illustrated by the hatched area in Fig. 4.1. In contrast to the KS-test, CvM considers the full distribution rather than the distance at a single point. The test-statistics is however still relatively insensitive to deviations in the tail.

The Anderson Darling (AD) statistic, based on CvM, compensates this by adding more weight to the tails of the distribution (Anderson & Darling, 1952). The equation for both statistics is:

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, w(x) \, dF(x), \tag{4.2}$$

where n is the sample size, $F(x)$ the hypothesized distribution (uniform, in this study), and $F_n(x)$ is the empirical cumulative distribution function (the expert's percentile points under the assumed probability distribution). The weight $w(x)$ differs for CvM and AD. In CvM, all realizations x have weight 1.0. For AD, more weight is assigned to both tails of the distribution:

$$w(x) = [F(x)(1 - F(x))]^{-1}. \tag{4.3}$$

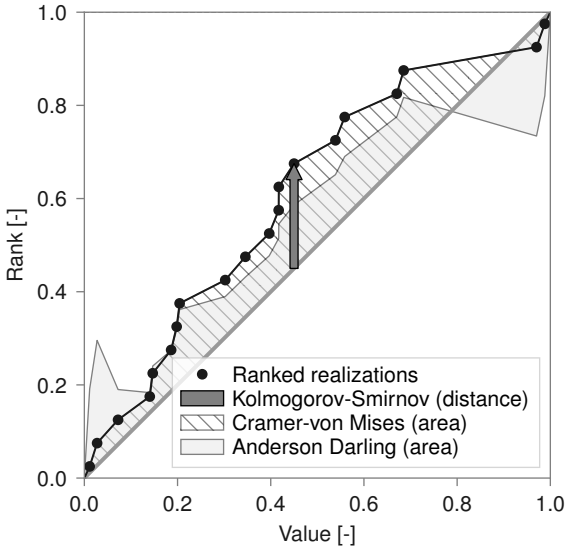


Figure 4.1: Illustration of KS, CvM, and AD test statistics for a sample from a uniform distribution. The sample is plotted by their ranks (the connected dots). The arrow indicates the Kolmogorov-Smirnov (KS) statistic, the hatched area Cramer-von Mises (CvM), and the filled area (a weighted version of the hatched area) the Anderson-Darling (AD) statistic.

By assigning a large weight to the deviation of quantile points in the tail, AD compensates CvM's insensitivity to overconfidence. This is shown by the filled area in Fig. 4.1, which, compared to the hatched area, has a larger distance to the diagonal at the edges. For CvM, distributions from (CSöRgő & Faraway, 1996) are used to convert the test statistic to a p-value. An approximation of the distribution for the AD test statistic, for a uniform distribution, is given by (Marsaglia & Marsaglia, 2004) and (Grace & Wood, 2012). Marsaglia and Marsaglia (2004) cover the full range $a \in (0,\infty)$ (with a being the AD-statistic), while the approximation of Grace and Wood (2012) is specified only for $a \in [3,\infty)$. The latter is more accurate for high values of a, which is why we apply Marsaglia and Marsaglia (2004) for $a \in (0,3)$, Grace and Wood (2012) for $a \in (4,\infty)$, and linearly interpolate between the two for $a \in [3,4]$ to ensure a smooth transition.

Shapiro-Wilk is another test-statistic that is often used for testing normality (Shapiro & Wilk, 1965). The problem with this statistic for our application is that it tests whether a sample is normally distributed with any mean and variance. It does not test whether a sample is standard normal distributed (i.e., $N(0,1)$), so neither can it be used to test whether a sample is uniformly distributed between 0 and 1. Therefore, it was not used in this study.

4.2.2. Metalog distribution

The Metalogistic, or Metalog, distribution is a continuous univariate probability distribution with high shape flexibility introduced by Keelin (2016). It accommodates bounded,

semi-bounded, and unbounded distributions. This makes it an appealing choice for fitting empirical data (e.g., as a continuous replacement for a histogram) but also for modeling expert estimates. The Metalog is a generalized form of the logistic distribution, achieved by substituting the mean and standard deviation in the logistic distribution's quantile function using series expansion. In this study, the three and five term functions $M_3(y)$ and $M_5(y)$ are used:

$$M_3(y) = a_1 + a_2 \ln \frac{y}{1-y} + a_3 (y-0.5) \ln \frac{y}{1-y}$$
$$M_5(y) = M_3(y) + a_4 (y-0.5) + a_5 (y-0.5)^2$$

Here, y denotes the cumulative probability and a_i the constants.

Because the distribution is defined using its quantile function, a unique n-sized vector a can be fitted to any set of n percentiles. This is a useful property for resembling experts' quantile estimates without changing their estimates. However, the quantile function $M_n(y)$ needs to be strictly increasing for all $y \in (0,1)$. This is not necessarily the case for all sets of n percentiles, resulting in invalid or infeasible distributions with negative probability density.

Figure 4.2 shows eight examples of Metalog distributions (smooth grey curves) and piecewise uniform (PWU) distributions (stepped black curves) fitted to either three-percentile estimates (a, b, c) or five-percentile estimates (d, e, f, g, h). For three-percentile estimates, an infeasible a-vector can be solved by imposing a lower or upper bound. This introduces a fourth parameter, making the solution overdetermined. We address this by selecting the bound such that it minimizes the maximum probability density, resulting in the least informative distribution.

For five-percentile estimates, many expert estimates (combinations of five quantiles) lead to infeasible distributions. To be able to process the results for case studies with five quantiles as well, the infeasible estimates are split in two by the median, resulting in two three-quantile estimates (e.g., 0.05, 0.25, 0.50, and 0.50, 0.75, 0.95). This gives a step in density at the median, as shown by the solid line in Fig. 4.2h. Optionally, this can be resolved by imposing a bound on the distribution with the lowest density at the median such as shown by the dashed line. However, this is primarily an aesthetic solution, which is why we chose not to do this. Note that the Metalog distribution can fit the quantile estimates with a feasible (non-negative) distribution, but this would require adding more terms to the a-vector than there are quantiles. Further details on the fitting procedure can be found in Section 4.B.

All tests except χ^2 use quantile points of realizations for evaluating statistical accuracy, rather than the quantile intervals in which the realizations fall. A fitted Metalog distribution provides these quantile points based on the expert estimates. This study examines whether the realization quantiles from the Metalog are a better representation of expert estimates than the realization quantiles generated by a PWU distribution. With respect to the CM, the Metalog also changes the calculation of informativeness since the default approach is based on the piecewise uniform assumption. For further explanation on how this calculation is performed for the Metalog, please refer to Section 4.B.

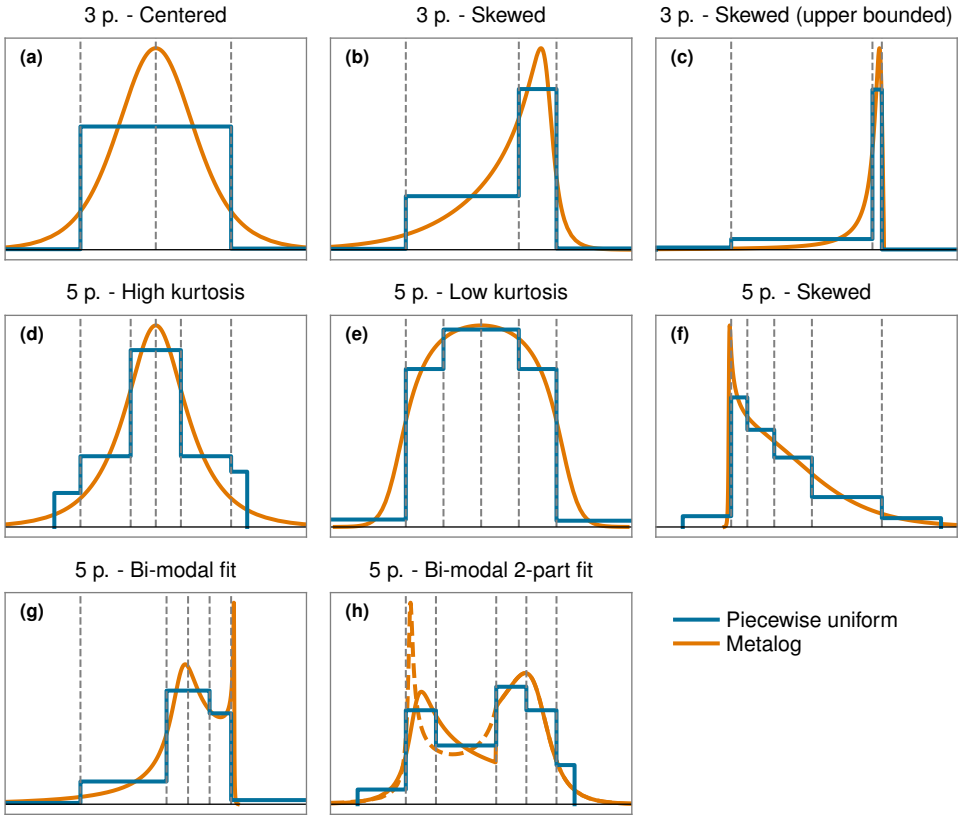


Figure 4.2: Eight examples of Metalog distributions and piecewise uniform distributions fitted to three-percentile (a, b, c) and five-percentile (d, e, f, g, h) expert estimates. The estimates are indicated with the vertical dashed lines.

4.2.3. Comparing measures of statistical accuracy

The five test statistics, as detailed in Section 4.2.1 evaluate statistical accuracy in different ways, leading to different test scores. The sensitivity to biases is explored using the method presented in Section 4.2.3. Section 4.2.3 explains how the statistical accuracy from the different tests are compared given their difference in values. Finally, the method for comparing the quantile estimates from PWU and Metalog is outlined in Section 4.2.3.

Scores' sensitivity to detect biases

To assess the ability of the measures of statistical accuracy to detect biases in experts' individual assessments, we introduce criteria for location bias and underconfidence or overconfidence. Location bias is defined as the absolute difference between the fraction

of realizations below the median estimate and 0.5, or

$$\left| \frac{\sum_{i=1}^{n} (x_i < F_{e,i}(0.5))}{n} - 0.5 \right|, \tag{4.4}$$

with n being the number of items, x_i the realization for item i and $F_{e,i}(0.5)$ expert e's median estimate for item i.

Overconfidence and underconfidence are quantified by the number of realizations below and above the lowest and highest estimated quantile, divided by their expected fraction. Let LQ and UQ be the lower and upper quantile (typically 0.05 and 0.95, 0.10 and 0.90 in two of the 49 cases). The ratio of tail realizations is calculated with

$$\frac{\sum_{i=1}^{n} (x_i < F_{e,i}(LQ)) + \sum_{i=1}^{n} (x_i > F_{e,i}(UQ))}{n(LQ + (1 - UQ))}. \tag{4.5}$$

A value greater than 1.0 indicates overconfidence, a value less than 1.0 indicates underconfidence.

Comparing measures of statistical accuracy via decision makers

As discussed in Section 4.2.1, the Classical Model (CM) relies on the χ^2 statistical accuracy in combination with the information score to assign weights to each expert. These weights are then used to aggregate experts' distributions into a decision maker (DM) using the global weights algorithm with optimization (GLopt) or without (GL). Whenever comparing global weights in the analyses, the values of the weights are the normalized product of statistical accuracy and informativeness. Within this, the measures of statistical accuracy introduced in Section 4.2.1 serve as alternatives for the statistical accuracy term. Consequently, different measures assign different weights to the experts. When applying optimization, this can lead to DMs composed of the estimates of different sets of experts. In addition to the global weights, we consider the equal weight DM (EQ) which assigns the same weight to every expert. Item weights were not considered.

We are interested in comparing the effects of applying different measures of SA within the CM on the decision maker's SA. For this, we cannot simply compare the SA of the DM calculated using each measure, because some measures give on average higher scores than others. For example, KS and CvM are less sensitive to overconfidence, a prevailing bias in CM studies, and therefore give higher SA scores. This does not mean the experts are actually statistically more accurate. To compare the measures, we consider the weights from each measure, evaluated with each of the measures of SA. Both with and without optimization, and for both the PWU and Metalog distribution. As an example, we list the steps in comparing the KS and CRPS weights according to the χ^2 measure of SA:

1. First, experts' weights, as the normalized product of statistical accuracy and informativeness, are calculated based on KS and CRPS measures of SA.
2. Decision makers distributions are obtained for each set of those weights, which we will refer to as $\mathrm{DM}_{\mathrm{KS}}$ and $\mathrm{DM}_{\mathrm{CRPS}}$.

3. The χ^2 measure of statistical accuracy is then also calculated for the DM_{KS} and DM_{CRPS}.

4. This is repeated for all 49 studies. Ranking these SAs gives a set of 98 ranked SA scores.

5. Using the Mann-Whitney rank sum test (Mann & Whitney, 1947), we test whether the DM_{KS} and DM_{CRPS} ranks are statistically equivalent, or whether one is lower (or higher) than the other according to the χ^2 measure of statistical accuracy.

We do this for all combinations of SA measures, such that each pair of DMs is compared with respect to each of the five measures of SA.

Determining the Metalog's and PWU's ability to predict missing quantiles

The choice of the Metalog distribution to represent the probability density between percentile estimates is rooted in the hypothesis that it better aligns with the distribution perceived by experts. This is due to its smooth curve without abrupt changes in probability density at estimated percentiles. To test this hypothesis, we remove the second and fourth percentile from the case-studies involving five elicited percentiles. The removed percentiles are then estimated using both the PWU and Metalog distributions. By comparing the difference between the estimated percentile point and the removed value (e.g., $F^{-1}(0.25) - x_{0.25}$ for the 25th percentile) we can determine which distribution more accurately predicts the location of the removed percentiles.

4.3. Results

We used expert data from 49 studies that are explained in (Cooke et al., 2021). The data comprise 6864 individual expert assessments. The different measures of statistical accuracy (SA) were calculated for the global weights decision maker (DM) with and without optimization, and the equal weights DM. These results allowed us to compare SA across different measures of SA (Section 4.3.1), assess the sensitivity to different biases (Section 4.3.2), and evaluate the ability of the Metalog and PWU distributions to predict missing quantiles (Section 4.3.4).

4.3.1. Individual experts' measures of statistical accuracy

The PWU and Metalog distribution's quantile estimates of the realization, for the 6864 individual experts, are shown in Fig. 4.3. Overconfidence is signaled by the high number of realizations in the tails. The quantile positions for realizations that fall outside the [0.05, 0.95] interval differ most between Metalog and piecewise uniform (PWU). The standard (PWU) approach requires an assumption on the probability density in the $[0.0, 0.05]$ and $[0.95, 1.0]$ range as the 0.0 and 1.0 quantiles are not elicited. These lower and upper bounds are usually placed at the minimum and maximum of *all* experts' estimates and the realization, extended by (typically) a 10% overshoot of this total range. Since only one expert gives the lowest or highest estimate, the estimates of the other experts end up

being extended by (much) more than 10%. This leads to the position of a realization in the (often very wide) tails being relatively close to the elicited outer quantiles (e.g., 0.05 and 0.95). The Metalog distribution does not require an assumption on the tails, except for estimates with very skewed estimates. Therefore, the realizations are placed based on the fitted distribution and experts are judged on their own overconfidence. This tends to result in quantiles closer to 0.0 and 1.0.

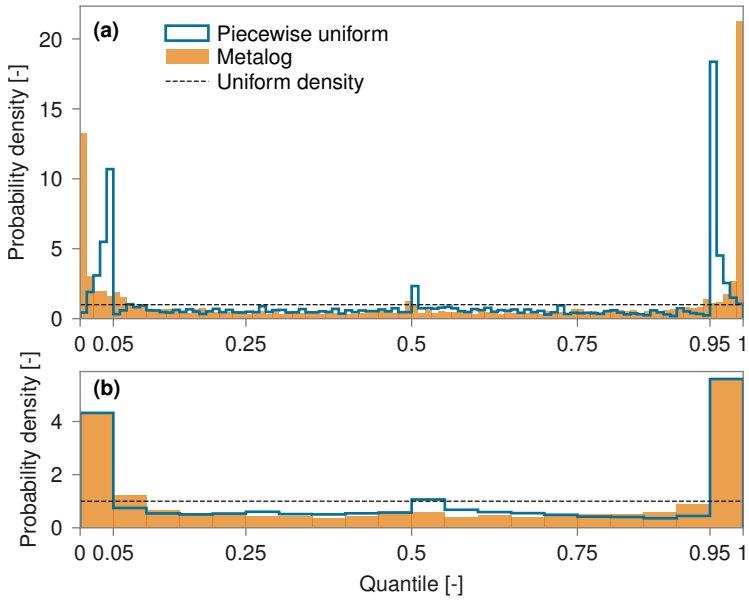


Figure 4.3: Quantiles of realizations for 6864 individual expert quantile assessments with respect to the fitted piecewise uniform (white) and Metalog (grey) distributions. The difference between a) with 100 bins and b) with 20 bins shows the effect of assuming PWU or Metalog on the tail quantiles.

Figure 4.4 shows the statistical accuracy (SA) for the five considered measures, for each of the 530 experts. Histograms of the SA for each individual measure are displayed on the diagonal. The first bin covers the first 5%, i.e., the significance level commonly used in simple hypothesis testing. For an expert with SA less than 5%, the hypothesis that the expert is statistically accurate would be rejected at the 5% level. The dashed lines in the scatter plots also indicate this 5% significance level. The scatter plots in the lower left triangle are obtained under the PWU assumption, those in the upper right triangle under the Metalog assumption. Many scatters are overlapping in the < 0.05 corner, Fig. 4.A.3 in Section 4.A shows more clearly how the measures compare in that range.

The scores χ^2, CRPS, KS, CvM and AD, assign a significance level above $> 5\%$ to 27%, 32%, 58%, 62%, and 46% of the experts when assuming PWU, and 27%, 18%, 49%, 48%, and 17% for Metalog. For all but χ^2, assuming a Metalog distribution leads to lower SAs relative to a PWU distribution. This is because using PWU results in the realizations being at quantiles closer to the 5th and 95th (as illustrated in Fig. 4.3). χ^2 relies on quantile intervals, such that the choice of the inter-quantile distribution does not affect statistical

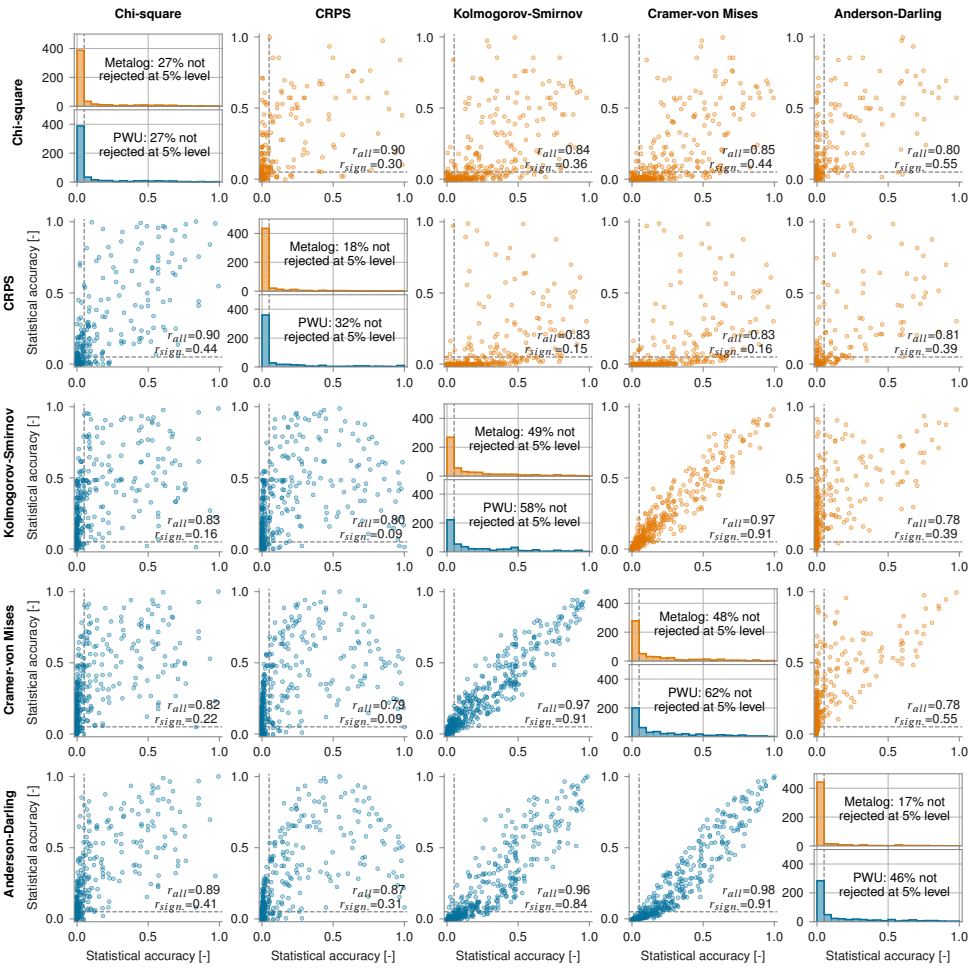


Figure 4.4: Statistical accuracy for the 530 experts based on their quantile assessments in 49 case studies, using the Metalog distribution (upper right panels) and piecewise uniform (lower left panels). The dashed line represents the 5% significance level. The two numbers in the lower right of each panel are the rank correlation between all experts (above), and the rank correlation between all experts with a greater than 0.05 SA in both test (below). Diagonal plots present the histogram of each measure's statistical accuracy for all 530 experts (i.e., the marginal distribution of each SA measure). In each histogram, the percentage of experts with a > 5% significance level is reported.

accuracy.

The rank correlations between the different measures of SA are high. When considering only experts with SA greater than 0.05 for both measures, the correlation is generally low for all combinations except between KS, CvM, and AD. Under the PWU assumption, CvM and KS are relatively similar to each other and to AD. When assuming the Metalog distribution, AD gives significantly lower scores than to KS and CvM. This is due to the extra weight assigned to realizations in the tail (see Fig. 4.1). A high SA for AD does however still ensure high SA with KS and CvM.

Figure 4.A.3 shows the same results plotted on a logarithmic scale, demonstrating again that KS, CvM, and AD are relatively similar measures of statistical accuracy. CRPS and χ^2 also show some resemblance due to their high sensitivity to overconfidence. Moreover, KS and CvM are less likely to yield very low ($< 10^{-5}$) scores, while χ^2 tends to give the lowest scores. In terms of expert weights, the linear representation is more relevant, as it will generally matter less for the DM whether one of the contributing experts gets a 10^{-3} or 10^{-10} score.

4.3.2. Analysis of sensitivity to biases

We analyzed the sensitivity of measures of statistical accuracy to under- and overconfidence and location bias (i.e., overestimating or underestimating). The method for calculating the biases was explained in Section 4.2.3. First, we examined the results for individual expert assessments, as depicted in Fig. 4.5. CRPS is known to be location bias insensitive. However, the CRPS location-bias scatter plot does not show a very different pattern from the other measures, indicating that experts who score high with CRPS are not strongly location-biased. χ^2, CRPS, and AD (under Metalog assumption) are most sensitive to overconfidence, while KS, CvM, and AD (under PWU assumption) are least sensitive to overconfidence. All scores are sensitive to underconfidence, however CRPS actually rewards it, which is further discussed in Section 4.4.1.

Another approach to assess the sensitivity of measures of statistical accuracy to biases is by sampling from known distributions. We simulated four experts with different biases,

1. the perfectly statistically accurate (no bias),
2. overconfident,
3. underconfident,
4. location-biased (overestimating) expert.

The results of the simulation are shown in Fig. 4.6. The four columns correspond to the four experts, with the top row showing the beta-distribution from which realization quantiles are sampled. For each expert, 5 to 50 values are sampled from the distribution. Repeating this process 10.000 times gives the distribution of p-values, indicated with the colored bands.

For the perfectly calibrated expert, all test statistics produce a uniform distribution for the p-value, which aligns with the asymptotic or exact distribution. χ^2 requires more

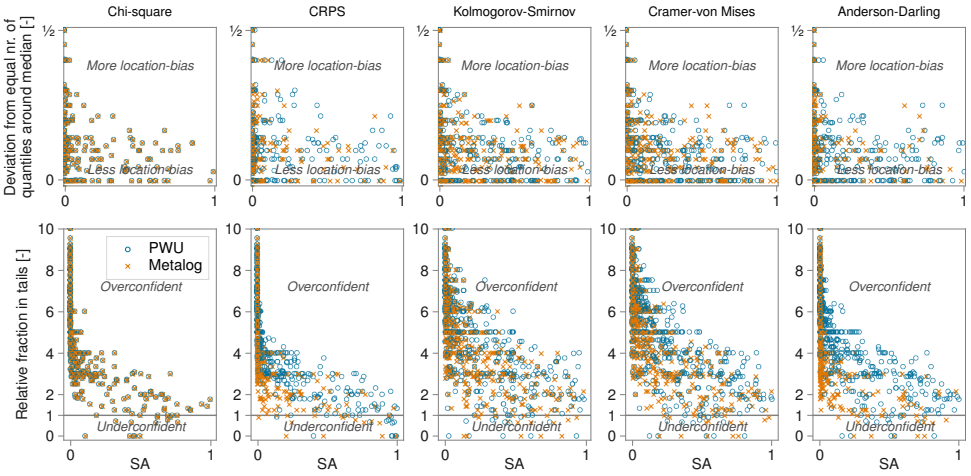


Figure 4.5: Sensitivity of the measures of statistical accuracy to biases. The top row shows sensitivity to under- and overestimating experts (location-bias), calculated using Eq. 4.4. The bottom row shows under- and over-confidence, calculated using Eq. 4.5. The crosses indicate statistical accuracy calculated for the Metalog, the circles for the piecewise uniform distribution.

realizations to reach this uniform result because the χ^2 distribution is an asymptotic rather than an exact distribution of the χ^2 test-statistic. For this reason, a p-value equal to 1 is only possible with 20 or more calibration variables, when eliciting the 5th, 50th and 95th percentile.

CRPS shows the highest sensitivity to overconfidence, followed by χ^2 and AD, and finally KS and CvM. For underconfidence, a similar sensitivity pattern emerges, except that CRPS rewards rather than penalizes underconfidence. An expert with location bias gets the lowest p-values from AD, KS, and CvM, followed by χ^2. CRPS does not pick up location bias, as explained by Nane and Cooke (2024), if the expert is not under- or overconfident.

Overall, KS and CvM respond similarly to biases and AD and χ^2 do as well. For small sample sizes, the continuous measures of statistical accuracy are less capricious, because of the use of exact distributions. Note that using 10 variables was deemed sufficient to select a statistically accurate expert over an overconfident expert. Note that in Fig. 4.6 the mean χ^2 score for 10 experts using the asymptotic distribution is not 0.50 but 0.40 (Cooke, 2014). The continuous measures' use of an exact distribution comes at the expense of assuming a distribution to assign a realization to a quantile. The uncertainty introduced by this assumption is not considered in this analysis beyond eyeballing Fig. 4.4

4.3.3. Comparison of decision maker statistical accuracy

The previous sections presented the individual statistical accuracy measures and sensitivity to biases. This section compares the weights derived using the different measures

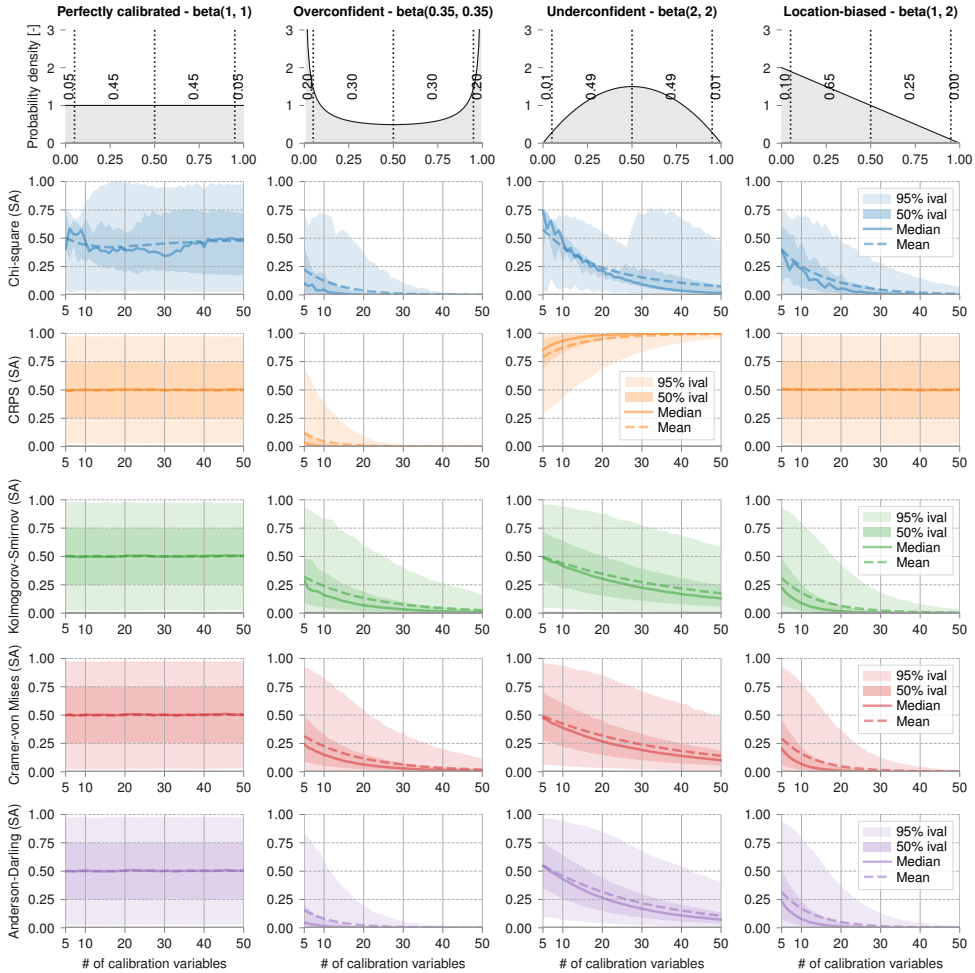


Figure 4.6: Distribution of the DM's statistical accuracy for the different measures of statistical accuracy, resulting from drawing 10.000 samples of 5 up to 50 realization quantiles with different biases and evaluating their statistical accuracy. The four (non-)biases are: perfectly calibrated, overconfident, underconfident, and location-biased (overestimating).

and the resulting statistical accuracy, following the procedure set out in Section 4.2.3. Section 4.A presents illustrations of the intermediate steps that are followed in deriving the results presented in this section.

The measures of statistical accuracy are compared by calculating decision maker weights using each measure of SA (recall that the weights are the normalized product of SA and informativeness), and evaluating the DM's statistical accuracy with, again, each of the measures of SA. This yields 49 values per combination of weight SA and score SA, whose means are shown in Table 4.1. Note that for χ^2 the mean value of 0.40 accords with the value shown in Fig. 4.6 for 10 to 20 calibration variables. For most combinations, the values on the diagonal are highest. For these, the same measure of SA is used for weights and score. This difference is larger for the GLopt DM than for the GL DM. Note that all SA measures except CRPS have a "low opinion of equal weighting" with χ^2 having the lowest. CRPS DM's statistical accuracy is actually slightly higher for equal weighting than for CRPS weighting.

Dist.	DM	SA (w.) / SA (sc.)	χ^2	CRPS	KS	CvM	AD	EQ
PWU	GL	χ^2	0.40	0.36	0.43	0.46	0.42	0.32
		CRPS	0.63	0.67	0.67	0.67	0.67	0.69
		KS	0.52	0.41	0.57	0.55	0.54	0.39
		CvM	0.51	0.40	0.56	0.57	0.55	0.38
		AD	0.51	0.39	0.56	0.57	0.55	0.39
	GLopt	χ^2	0.55	0.38	0.39	0.38	0.43	0.32
		CRPS	0.50	0.62	0.35	0.36	0.42	0.69
		KS	0.49	0.42	0.71	0.67	0.66	0.39
		CvM	0.49	0.43	0.69	0.72	0.68	0.38
		AD	0.49	0.44	0.63	0.66	0.65	0.39
Metalog	GL	χ^2	0.39	0.33	0.46	0.44	0.37	0.32
		CRPS	0.48	0.46	0.54	0.56	0.40	0.59
		KS	0.51	0.43	0.58	0.58	0.46	0.40
		CvM	0.50	0.43	0.59	0.60	0.45	0.39
		AD	0.45	0.36	0.53	0.54	0.36	0.40
	GLopt	χ^2	0.52	0.32	0.42	0.43	0.40	0.32
		CRPS	0.34	0.46	0.27	0.31	0.33	0.59
		KS	0.44	0.44	0.70	0.70	0.49	0.40
		CvM	0.46	0.44	0.67	0.71	0.50	0.39
		AD	0.34	0.36	0.39	0.47	0.41	0.40

Table 4.1: Average DM statistical accuracy for the 49 case studies, calculated with weights from the measures of SA on the columns "SA (w.)" and Equal weights "EQ", and scored using the measures of SA on the rows "SA (sc.)". The results for the GL and GLopt decision maker, as well as the PWU and Metalog distribution are shown.

Table 4.2 displays the p-values of the Mann-Whitney test that compares whether the ranks of SA (for which the means are shown in Table 4.1) are significantly different from

each other. The top five rows compare the ranks under the piecewise uniform assumption and the global weights DM. The only significant number is the 0.025 between CRPS (row) and CvM (column). This suggests that the SA calculated with DM weights from CvM evaluated with χ^2 is significantly higher than the SA calculated with DM weights from CRPS evaluated with χ^2. Or, $P(SA_{DM_{CvM}|\chi^2} > r)$ is greater than $P(SA_{DM_{CRPS}|\chi^2} > r)$ for all r in $(0,1)$.

Dist.	DM	SA	χ^2	CRPS	KS	CvM	AD
PWU	GL	χ^2		0.818	0.217	0.121	0.341
		CRPS	0.184		0.055	**0.025**	0.101
		KS	0.785	0.945		0.345	0.649
		CvM	0.881	0.976	0.657		0.809
		AD	0.661	0.900	0.353	0.193	
	GLopt	χ^2		1.000	0.998	0.999	0.987
		CRPS	**0.001**		0.460	0.582	0.182
		KS	**0.002**	0.542		0.602	0.230
		CvM	**0.001**	0.421	0.400		0.157
		AD	**0.013**	0.820	0.772	0.844	
Metalog	GL	χ^2		0.880	0.093	0.198	0.731
		CRPS	0.121		**0.009**	**0.027**	0.275
		KS	0.908	0.991		0.660	0.962
		CvM	0.804	0.974	0.343		0.919
		AD	0.272	0.727	**0.039**	0.082	
	GLopt	χ^2		0.999	0.959	0.963	0.981
		CRPS	**0.001**		**0.040**	**0.027**	0.117
		KS	**0.042**	0.961		0.409	0.662
		CvM	**0.038**	0.974	0.594		0.702
		AD	**0.020**	0.884	0.340	0.301	

Table 4.2: p-values for the Mann-Whitney rank-sum test. A p-value less than 0.05 (bold) suggests that the ranks of the SAs calculated with the measure of SA in the row is less than the ranks of the SAs calculated with the measure of SA in the column. The SA-values are ranked according to the 49 χ^2 DM's SAs. Both the ranks using PWU and Metalog distribution, as well as using GL and GLopt decision maker, are compared.

Table 4.2 and Table 4.A.1 show that for the global weights DM without optimization the differences are mostly insignificant. The exception are the weights from $CRPS$, which often score significantly lower, especially when evaluated according to the KS, CvM, or AD test.

For the global weights DM *with* optimization, the χ^2 SA calculated with weights from every measure other than χ^2 itself, are significantly lower (see the first column in Table 4.2, rows corresponding to GLopt). Similarly, the CRPS SA calculated with global optimized weights from the other measures is considered significantly lower as well (see Table 4.A.1). And again, KS, CvM, and AD behave similarly as a group; the SAs calculated with weights from χ^2 and CRPS are significantly lower than the SAs calculated with weights from the measures themselves, but the SA from weights in between KS, CvM,

and AD are not significantly lower (or higher). For the Metalog distribution, AD's sensitivity to overconfidence makes it behave more similar to χ^2 and CRPS, and less similar to KS and CvM.

Based on the comparison in this section, and the analyses from Section 4.3.2, the measures of statistical accuracy can be divided into three categories with similar response to specific characteristics of expert assessments:

1. KS, CvM, and AD value quantiles close to their ranked position.
2. χ^2 values a proportional number of quantiles in bins.
3. CRPS values the median estimate close to the realization's quantile.

Figure 4.4 shows that while the correlation between each of these three categories is high, the correlation for the experts that score $SA > 0.05$ in both test is mostly low. Using the global weight algorithm gives enough spread in weight for the differences between KS, CvM, AD on one side, and χ^2 on the other, to be (mostly) insignificant. This is because all four measures give a high statistical accuracy to a close to uniform distribution of quantiles between 0 and 1. However, CRPS responds to a much different characteristic, which makes the difference between it and the other measures significant also under global weights.

When applying optimization, the weight is further concentrated onto a few experts. Referring again to Fig. 4.4, if the weights are assigned to one (or a few) of the experts with a high SA for χ^2, it may not result in a high SA for one of the other measures. This is expressed by the significantly lower p-values in Tables 4.A.1 and 4.2, when applying optimized weights from another measure to the measure under evaluation itself.

This leaves the question why χ^2 behaves differently from KS, CvM and AD (i.e., why the correlation for $SA > 0.05$ in Fig. 4.4 is low). Although all reward a uniform interquantile distribution, this may result from evaluating quantiles on a continuous scale against evaluating them in bins. When realizations are close to the elicited quantile, a small difference in weight might cause a shift to another interval, which can make a large difference in χ^2 SA score. Additionally, the distribution of quantiles within a quantile interval does not matter for χ^2, while KS, CvM and AD reward them being spread out. GL combines all experts so that the CDF has jump points at the quantile values for each expert. For 10 experts there are 30 jump points, the CDF looks rather continuous and the effect of interpolating a continuous CDF is attenuated. GLopt, on the other hand, typically weights only one or two experts and hence has much fewer jump points. This might amplify distortions introduced by interpolating a continuous CDF.

Finally, the behavior of AD heavily depends on assuming the PWU or Metalog distribution. AD penalizes quantiles close to 0.0 or 1.0 much more than quantiles close to 0.05 and 0.95 (refer to the weight in equation (4.3)). This means that assuming the Metalog distribution will make AD (much) more sensitive to overconfidence, and therefore behave more similar to χ^2 and CRPS than to KS and CvM. While this section explores some of the aspects that cause the differences in behavior, the last word on this has not been said.

4.3.4. Assessing accuracy of interpolated quantiles

The choice of five versus three elicited quantiles is made by the analyst for the whole study, not per variable. Opting for three percentiles lowers the elicitation burden, whereas five percentiles more accurately represent the experts' distributions.

An important reason for choosing a Metalog distribution for fitting experts' percentile estimates and estimating realization quantiles, is the hypothesis that it more accurately describes the distribution envisioned by experts. Unlike the PWU distribution, the Metalog distribution lacks discontinuities in probability density at estimated percentiles and, in its the three-percentile version, resembles a bell-shaped curve that is commonly observed in samples. To test this hypothesis, we consider the cases with five elicited percentiles, remove the second and fourth quantile, and estimate their position using both distributions (see Section 4.2.3).

The results are illustrated in Fig. 4.7, which displays the difference between distribution-estimated percentiles and expert-estimated percentiles, for $F(X_{q=0.25}) - 0.25$, with F being the CDF of Metalog or PWU and $X_{q=0.25}$ the expert estimate for the 25^{th} percentile. The dashed line represents a perfect prediction by the distribution. For the 25^{th} percentile, values to the left indicate that the distribution assigned a lower percentile to the experts' estimate. Conversely, values to the right indicate that the distribution assigns a higher percentile than the experts. Panels (a) and (b) show high bin values at 0.025. These values correspond to cases where an expert assigns a value to the 25th percentile precisely between the 5th and 50th percentiles. This situation accounts for approximately 20% of expert estimates. Filtering these estimates results in the black histogram lines. On average, PWU performs better than Metalog. Nevertheless, both distributions show significant deviations when estimating the missing percentiles. It seems that the distributions both lack predictive power for the missing percentiles.

Based on the analysis presented in this section, the Metalog does not offer a better representation of experts' estimates compared to PWU. The smooth and more informative distribution is too precise (i.e., it concentrates probability density more than experts appear to do). Consequently, the best approach to obtain a more accurate representation of experts' probability density functions (PDFs) seems to be eliciting more percentiles.

4.4. Discussion and conclusions

This study set out to test five different measures of statistical accuracy for scoring experts in an expert judgment study. The results are applicable to evaluating and combining uncertainty estimates in a broader context as well, for example in forecasting. The newly evaluated test statistics — the Continuous Ranked Probability Score (CRPS), Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), and Anderson-Darling (AD) — were assessed as an alternative to the relative information based chi-square test (χ^2) used in the Classical Model. Where χ^2 interprets and scores the estimates through discrete quantile intervals, the four alternatives map realizations on a continuous CDF-scale and accordingly calculate the statistical accuracy based on a continuous distribution. This makes the assumed distribution that connects the expert estimated percentiles relevant.

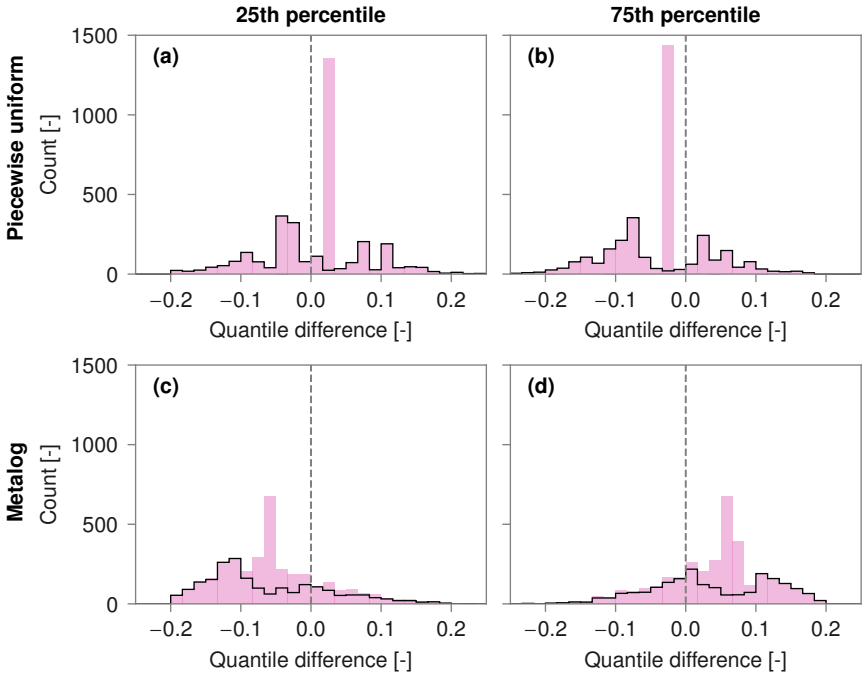


Figure 4.7: Difference between the expert actual estimated 25th and 75th percentile, and the positions according to the PWU and Metalog distribution fitted to the five-percentile cases with the 25th and 75th percentile removed.

In this context, the Metalog distribution was explored as an alternative to the piecewise uniform (PWU) assumption that is typically employed to model the estimated probability density in the Classical Model. The test statistics were assessed through 49 expert judgment studies carried out throughout the last decades, and by sampling estimates from distributions with a specific bias. The study's findings are discussed in two parts: the performance of various test statistics (Section 4.4.1) and the performance of the Metalog distribution (Section 4.4.2).

4.4.1. Performance of different test statistics

Comparative analyses of the five measures of statistical accuracy reveal varying sensitivity to different biases. χ^2 is sensitive to overconfidence, underconfidence and location bias. KS and CvM are sensitive to location bias and underconfidence but less sensitive to overconfidence. AD performs relatively similar to KS and CvM when assuming a PWU distribution. The assumed overconfidence related to the Metalog distribution however makes the AD weights much stricter on overconfidence. Because of this, AD behaves more similar to χ^2 under Metalog.

The new scale-invariant CRPS is sensitive to overconfidence but is insensitive to location bias and rewards underconfidence. A relationship between CRPS and underconfi-

dence is illustrated in Fig. 4.8, which plots statistical accuracy (a) and combined score (b) to the average quantile distance to median (an alternative to the fraction above median displayed in Fig. 4.5).This distance is calculated as $\sum_{i=1}^{N} |F_{i,j}(x_i) - 0.5|/N$, in which

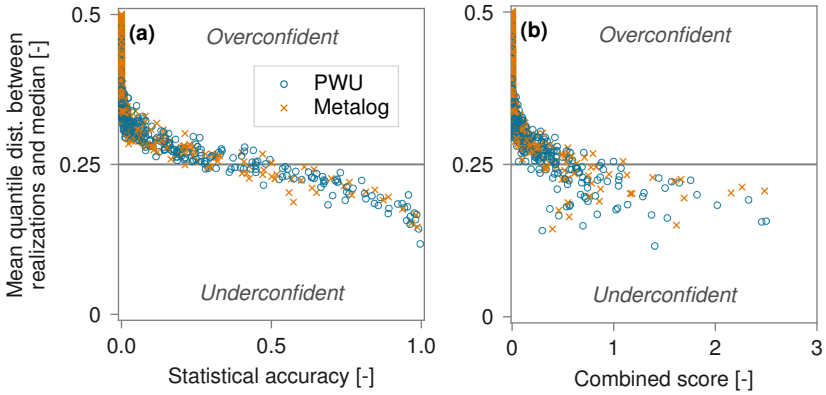


Figure 4.8: Mean absolute difference between realization's quantile and median for CRPS. Plotted against statistical accuracy (a), and plotted against combined score (b).

$F_{i,j}$ is expert j's estimate for item i with realization x_i. An unbiased, uniform, quantile distribution would have an expected distance of 0.25. The figure shows that there is a strong relationship between CRPS and distance to median. Considering the combined score, which includes the informativeness, weakens the relationship. However, the lower informativeness of underconfident experts does not offset the high statistical accuracy, such that underconfident experts still achieve high combined scores. This means that a perfectly accurate expert could achieve a higher statistical accuracy, and likely a higher weight as well, when deliberately making underconfident estimates. It therefore does not encourage experts to state their true, unbiased, beliefs.

Measures of statistical accuracy were compared by calculating decision maker weights with one measure and evaluating them against DMs from the other measures. The results show that for global weights without optimization the differences between SA scores based on weights from different measures are mostly insignificant. The exception is CRPS, which weights lead to significantly lower SA when evaluated with other measures. Applying *optimized* global weights further concentrates weights on experts with a high SA in the specific test-statistic, for which reason they perform worse when applied to most other measures of statistical accuracy.

Under the PWU assumption, SAs calculated with DM-weights from AD are considered not statistically significantly different when evaluated by KS and CvM (see Table 4.A.1). However, when calculated under the Metalog distribution, the SA from AD weights rank significantly lower when evaluated using KS and CvM. The inverse is not the case, weights from KS and CvM do not rank significantly different when evaluated by AD under either PWU or Metalog. This indicates that penalizing more on overconfidence (as AD is under Metalog) leads to significantly lower SA, while penalizing less on overconfidence (i.e., using KS or CvM weights under Metalog) does not significantly reduce SA.

4.4.2. Performance of the Metalog distribution

The Metalog distribution (Keelin, 2016) was explored as an alternative to the piecewise uniform (PWU) distribution that is typically used in CM. CRPS, KS, CvM, and AD require the realizations' quantile positions rather than intervals and the Metalog offers a flexible distribution for placing those. The most significant differences between the Metalog and PWU are the quantile positions for realizations that fall outside the $[0.05, 0.95]$ interval, as shown in Fig. 4.3. Using the Metalog means assuming a higher degree of overconfidence, yielding worse results for measures of statistical accuracy that penalize this. However, this issue is not intrinsic to the Metalog distribution itself, but resides in the (overshoot) range that is assumed for the PWU distribution.

The appealing feature of the Metalog is its smooth, bell-shaped curve, which may be more intuitive to experts (a 'soft' argument, but nonetheless relevant in the field of expert judgment). Typically, a continuous variable should not exhibit a spike in probability density at estimated percentiles. Assuming a bell-shaped curve increases probability density closer to the median while reducing it toward the tails (see Fig. 4.2 a). Paradoxically, when removing the second and fourth percentile from the five-percentile cases and estimating their removed position, the Metalog distribution mostly overestimates the probability density within the $[0.25, 0.75]$ quantile interval. This implies that, in these case studies, experts more often estimated a platykurtic (negative kurtosis, thin-tailed) than a leptokurtic (positive kurtosis, fat-tailed) distribution. The PWU distribution also underperformed in this experiment, highlighting that the best approach for an analyst to obtain an accurate representation of the full distribution is to assess more percentiles.

Fitting a Metalog distribution to all expert estimates in the 49 case-studies proved challenging. While the Metalog distribution offers high shape flexibility, it could not accommodate highly skewed three-percentile estimates without imposing bounds. Additionally, many five-percentile estimates could not be fitted without dividing the distribution into two three-percentile Metalog distribution parts (such as shown in Fig. 4.2 g).

4.4.3. Final remarks

All the five measures of statistical accuracy have different effects on the resulting combined estimate (DM), such as the number of experts that are included with significant weight, the sensitivity to different biases, and the assumptions required for calculating the weights. When assuming a piecewise uniform distribution, KS, CvM, and AD behave similar, but different to χ^2, which again behaves different to CRPS. When assuming the Metalog, the higher degree of assumed overconfidence makes AD behave more similar to χ^2 and CRPS, and less to KS and CvM. All measures except CRPS agree that performance weighting is superior to equal weighting with respect to statistical accuracy.

Neither the PWU nor Metalog distribution did a good job of predicting missing quantile assessments. This, together with the issues in consistently fitting the Metalog distribution to expert estimates and the minimal assumptions in the PWU distribution, favors the PWU distribution as the standard approach in the CM.

Nonetheless, having various options of distributions, as well as different measures of statistical accuracy, provides analysts with flexibility to tailor the approach to their specific

study. These options are available through the open-source Anduryl software, aiding use and further development. For example, a smooth (Metalog) distribution might be preferable over a stepped PDF in a scenario involving gradient-based sampling or optimization. Or perhaps a researcher might decide that a measure of statistical accuracy that is less strict on overconfidence, likely assigning more weights to low scoring experts, is preferred for a specific study. Such decisions are up to the analyst to make and this study aims to provide the knowledge and insights for making it a well-informed decision.

Appendix 4.A: Background on comparison between measures of statistical accuracy

The procedure for cross-comparing the test statistics was explained in Section 4.2.3, and the results were presented in Section 4.3.3. This appendix section gives additional explanation and illustrations on the intermediate steps.

When applying the different measures of statistical accuracy to the 49 studies, some measures give higher statistical accuracy (on average) than others. Comparing the obtained statistical accuracies is therefore biased. To overcome this, empirical distributions were derived for each measure (recall step 4 in the list in Section 4.2.3). These empirical distributions are shown in Fig. 4.A.1. Each panel contains five curves, of which each is constructed by calculating the decision maker in the 49 studies and ranking the resulting SA. For example, the CRPS distribution (squares) was constructed by calculating the CRPS statistical accuracy for all experts and combining that with the informativeness to derive weights. These weights were combined into a decision maker, for which the CRPS SA was calculated. This gives one of the 49 markers in the empirical distribution. Repeating this for all cases using the global weight DM with and without optimization, and the piecewise uniform and Metalog distribution, results in the five distributions in each of the four panels of Fig. 4.A.1.

The empirical distributions of each method's DM-scores are primarily used for comparing statistical accuracy. However, they also show that:

1. Measures of statistical accuracy that are less sensitive to overconfidence (i.e., KS, CvM) tend to return higher SA scores for DM, especially under the Metalog assumption. This is because overconfidence is a prevailing bias in expert judgment studies.

2. Optimization results in higher DM SA. Note that this is not necessarily the primary goal of optimization, which typically results in increased informativeness while not decreasing the SA.

3. Assuming the Metalog distribution gives lower scores than assuming PWU for AD and CvM, due to the measures' sensitivity to overconfidence. χ^2 is sensitive to overconfidence as well, but unaffected by the choice of distribution because it utilizes quantile intervals. At the same time, it profits from the higher informative-

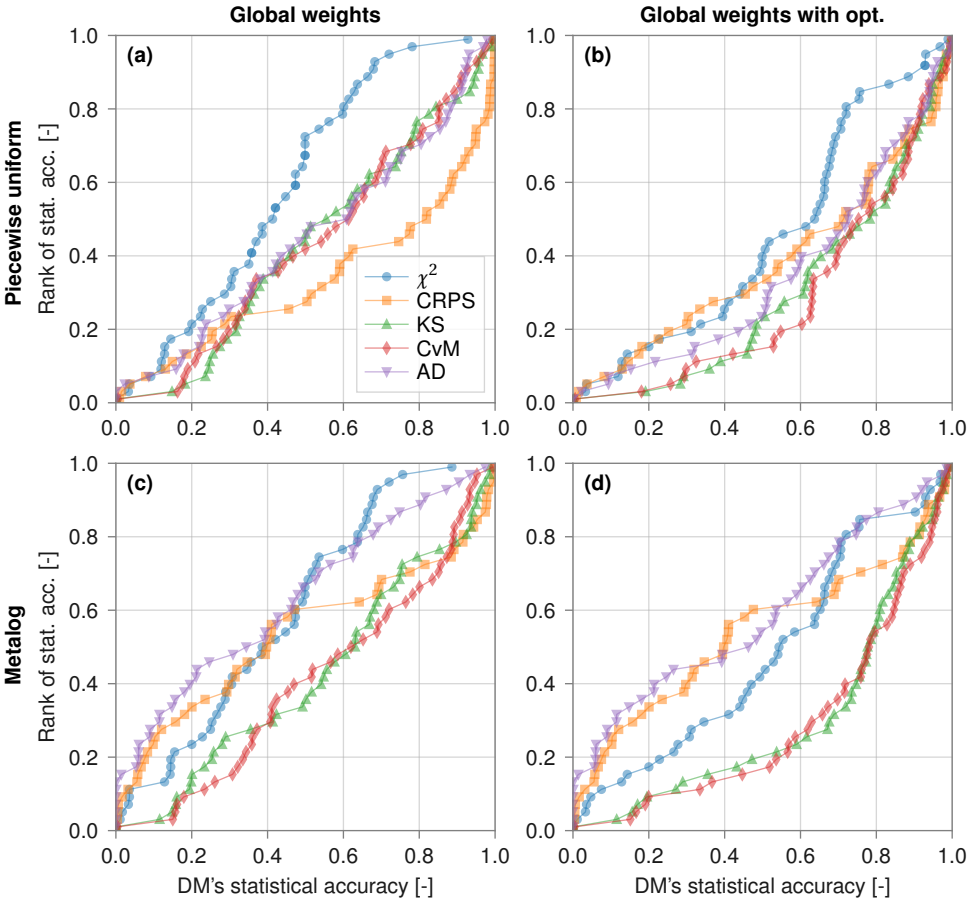


Figure 4.A.1: Empirical distributions of the decision makers' statistical accuracies in the 49 case-studies. a and b for piecewise uniform distribution, c and d for the Metalog; a and c for the global weights DM without optimization, and b and d with optimization.

ness of the Metalog distribution (the same reason KS and CvM score higher with Metalog as well.)

Using the empirical distributions, a χ^2 SA calculated with DM weights from, for example, the CRPS, can be compared to the χ^2 SA calculated with DM weights from χ^2 itself. Doing this for all 49 cases results in a set of ranks per measure of SA. Figure 4.A.2 displays these ranks for SA calculated with weights from each measure of SA and evaluated using each measure of SA. For example:

- Figure 4.A.2 (a) shows the χ^2 SA calculated with global DM_{χ^2} weights and compared to the empirical distribution of DM_{χ^2} SA. This is in fact comparing the χ^2 ranks to the χ^2 ranks itself, resulting in a uniform distribution.

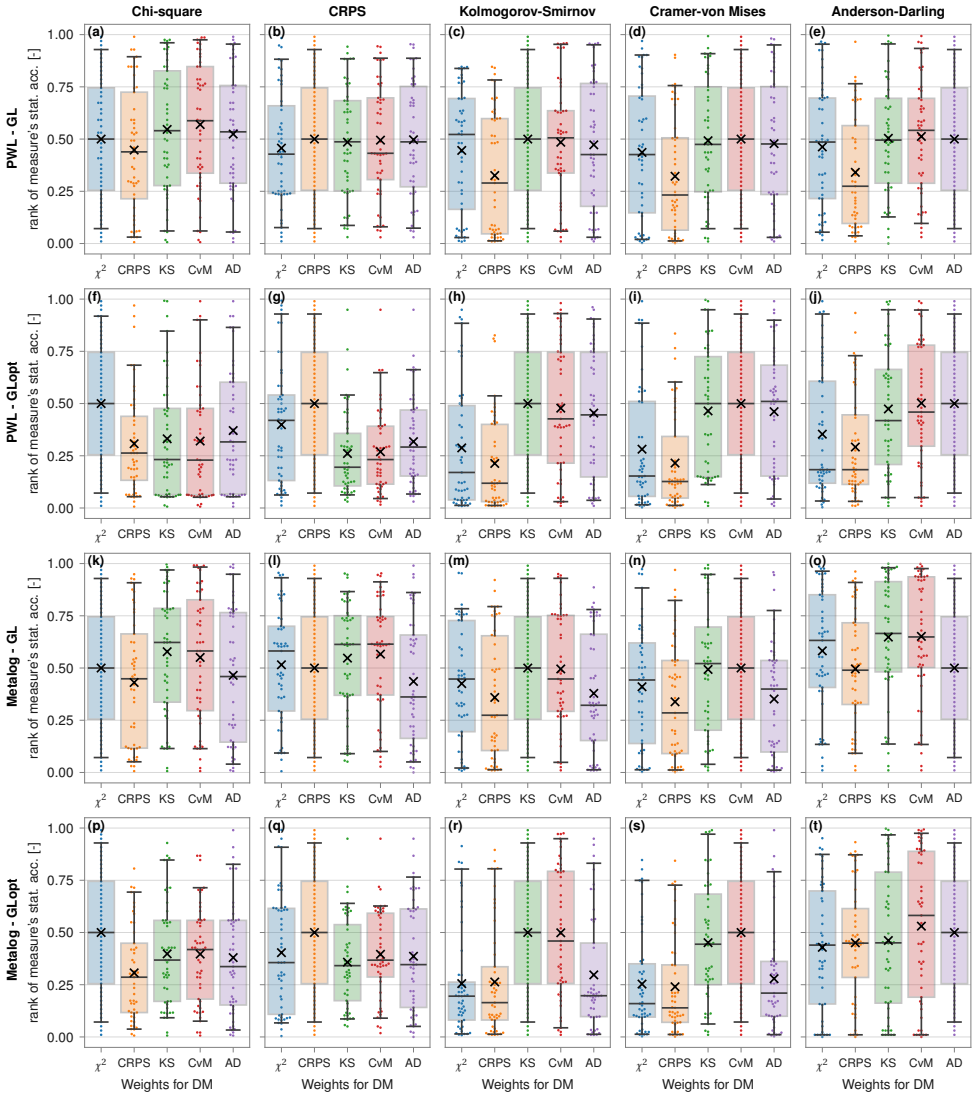


Figure 4.A.2: Comparison of all measures of statistical accuracy. DMs composed using each measure's weight (each box plot) are compared to the empirical distribution of every measure (each column). The comparison is done with and without optimization, and for the PWU and Metalog distribution. Each box indicates the 25th to 75th percentile range, with the horizontal line being the median. The fliers indicate the 5th and 95th percentile, and the cross is positioned at the mean.

SA (sc.)		CRPS				Kolmogorov-Smirnov				Cramer-von Mises				Anderson-Darling			
Dist.	DM	SA (w.)	χ^2	CRPS	KS	CvM	AD	χ^2	CRPS	KS	CvM	AD	χ^2	CRPS	KS	CvM	AD	χ^2	CRPS	KS	CvM	AD
PWU	GL	χ^2		0.23	0.23	0.22	0.23		0.97	0.18	0.31	0.32		0.96	0.16	0.14	0.23		0.98	0.20	0.17	0.26
		CRPS	0.77		0.61	0.54	0.52	**0.03**		**0.00**	**0.00**	**0.01**	**0.04**		**0.00**	**0.00**	**0.00**	**0.02**		**0.00**	**0.00**	**0.00**
		KS	0.77	0.40		0.46	0.44	0.82	1.00		0.60	0.69	0.84	1.00		0.45	0.62	0.80	1.00		0.40	0.52
		CvM	0.78	0.47	0.54		0.49	0.69	1.00	0.41		0.56	0.86	1.00	0.55		0.65	0.83	1.00	0.60		0.60
		AD	0.77	0.48	0.57	0.52		0.68	0.99	0.32	0.45		0.77	1.00	0.39	0.36		0.75	1.00	0.48	0.40	
	GLopt	χ^2		**0.04**	0.99	0.98	0.87		0.86	**0.00**	**0.00**	**0.00**		0.86	**0.00**	**0.00**	**0.00**		0.83	**0.02**	**0.00**	**0.01**
		CRPS	0.96		1.00	1.00	1.00	0.15		**0.00**	**0.00**	**0.00**	0.14		**0.00**	**0.00**	**0.00**	0.17		**0.00**	**0.00**	**0.00**
		KS	**0.01**	**0.00**		0.42	0.08	1.00	1.00		0.65	0.79	1.00	1.00		0.27	0.57	0.98	1.00		0.32	0.33
		CvM	**0.02**	**0.00**	0.58		0.10	1.00	1.00	0.36		0.61	1.00	1.00	0.74		0.75	1.00	1.00	0.68		0.51
		AD	0.13	**0.00**	0.92	0.90		1.00	1.00	0.22	0.39		1.00	1.00	0.44	0.25		0.99	1.00	0.67	0.49	
Metalog	GL	χ^2		0.61	0.25	0.16	0.91		0.86	0.10	0.15	0.82		0.89	0.09	0.07	0.84		0.94	0.11	0.11	0.92
		CRPS	0.39		0.21	0.13	0.86	0.14		**0.01**	**0.01**	0.37	0.11		**0.00**	**0.00**	0.39	0.06		**0.00**	**0.00**	0.47
		KS	0.76	0.79		0.39	0.97	0.90	0.99		0.53	0.98	0.91	1.00		0.45	0.99	0.90	1.00		0.46	0.99
		CvM	0.85	0.87	0.62		0.98	0.85	0.99	0.47		0.98	0.93	1.00	0.55		0.99	0.89	1.00	0.54		1.00
		AD	0.09	0.14	**0.03**	**0.02**		0.18	0.63	**0.02**	**0.02**		0.16	0.62	**0.01**	**0.01**		0.08	0.53	**0.01**	**0.00**	
	GLopt	χ^2		**0.05**	0.72	0.44	0.56		0.55	**0.00**	**0.00**	0.28		0.68	**0.00**	**0.00**	0.28		0.30	0.28	0.06	0.12
		CRPS	0.95		0.99	0.96	0.97	0.45		**0.00**	**0.00**	0.26	0.32		**0.00**	**0.00**	0.17	0.70		0.46	0.10	0.20
		KS	0.28	**0.01**		0.15	0.34	1.00	1.00		0.50	1.00	1.00	1.00		0.20	1.00	0.72	0.54		0.19	0.26
		CvM	0.57	**0.04**	0.85		0.62	1.00	1.00	0.50		1.00	1.00	1.00	0.80		1.00	0.94	0.91	0.81		0.71
		AD	0.44	**0.03**	0.66	0.38		0.72	0.74	**0.00**	**0.00**		0.73	0.83	**0.00**	**0.00**		0.88	0.80	0.74	0.29	

Table 4.A.1: p-values for the Mann-Whitney rank-sum test evaluated using the measures of SA: CRPS, KS, CvM, and AD. A p-value less than 0.05 (bold) suggests that the ranks of the SAs calculated with the measure of SA in the row is less than the ranks of the SAs calculated with the measure of SA in the column. The SA-values are ranked according to the empirical distributions of DM SA displayed in Fig. 4.A.1. Both the ranks using PWU and Metalog distribution, as well as using GL and GLopt decision maker, are compared. SA (w.) indicates the measure of SA used to calculate the weights, SA (sc.) is the measure used to score the weights

- Figure 4.A.2 (a) shows the χ^2 SA calculated with global DM_{CRPS} weights and compared to the empirical distribution of DM_{χ^2} SA. The ranks for DM_{CRPS} are on average slightly lower.
- Figure 4.A.2 (f) shows the χ^2 SA calculated with global DM_{CRPS} *optimized* weights and compared to the empirical distribution of *optimized* DM_{χ^2} SA. Now the CRPS weights give substantially lower statistical accuracy.

Whether substantially lower is also significantly lower is tested using the Mann Whitney test. The resulting p-values of those tests, for the DM weights evaluated using χ^2, were presented in Table 4.2. The first five rows in the table correspond to Fig. 4.A.2 a. It shows the probability (p-value) that each of the samples is lower than the other sample. For example, the statistically significant p-value of 0.025 for $DM_{CRPS} < DM_{CvM}$ is represented by the CvM (red) box plot showing higher values than the CRPS box plot (second from the left of each plot). Figure 4.A.2 f, k, and p correspond to the remaining rows in Table 4.2.

The p-values of the rank comparison evaluated for the other measures of statistical accuracy are shown in Table 4.A.1. These correspond to the remaining panels in Fig. 4.A.2.

Finally, Fig. 4.A.3 shows the same figure as Fig. 4.4 but now on a logarithmic scale. This shows the behavior of the measures to experts that score a very low statistical accuracy.

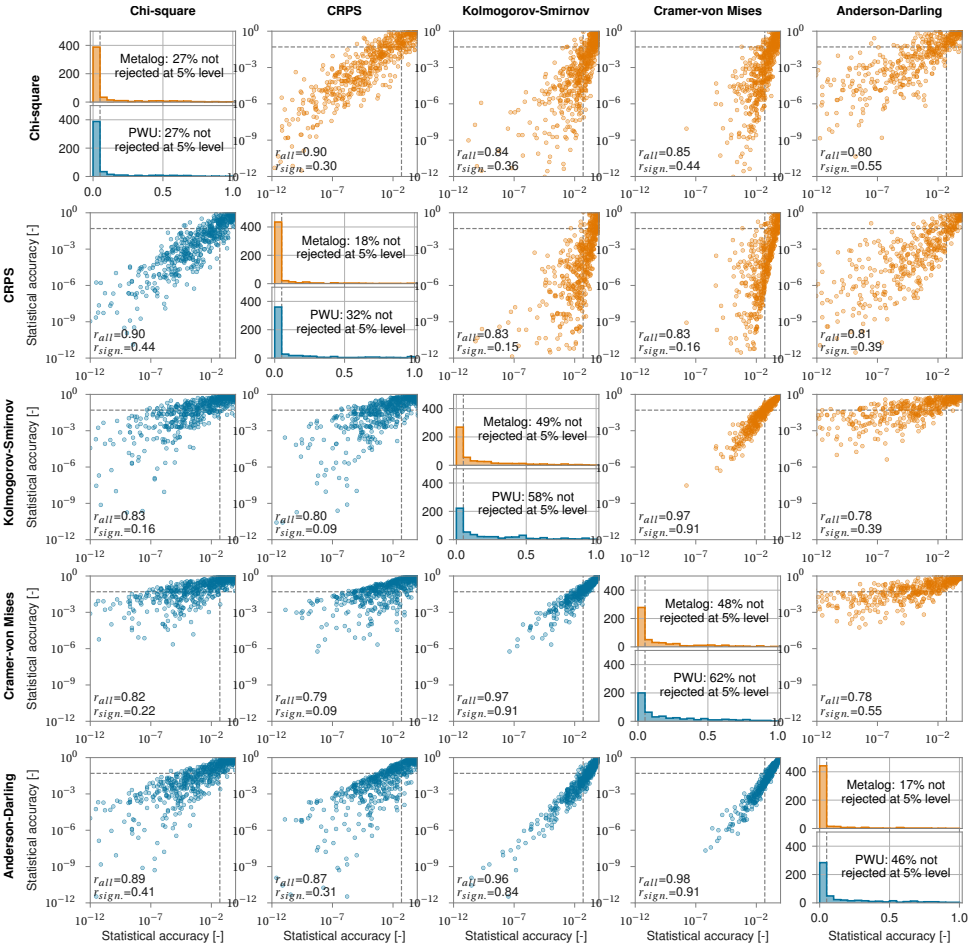


Figure 4.A.3: Statistical accuracies for the 530 experts based on their quantile assessments in 49 case studies, using the Metalog distribution (upper right panel) and piecewise uniform (lower left panels). Similar to Fig. 4.4 but on a logarithmic scale. The axes are limited to 10^{-12}, not showing SA values below this limit. The two numbers in the lower left of each panel are the rank correlation between all experts, and the rank correlation between all experts with a greater than 0.05 SA in both test. Diagonal plots present the histogram of each measure's statistical accuracy for all 530 experts (i.e., the marginal of each scatter plot). In each histogram, the percentage of experts with a > 5% significance level is reported.

Appendix 4.B: Metalog distribution

4.B.1. Information score for the Metalog distribution

In the Classical Model, the information score compares the probability density distribution $f_{e,i}$ for item i elicited from expert e, to a background density g_i. This background density is uniform across the intrinsic range $[L^*, U^*]$. This range is defined for each item by collecting all experts' estimates and the realization. The minimum and maximum of this set form the lower bound L and upper bound U. An overshoot k (typically 0.1) is than added to obtain $[L^*, U^*] = [L - k(U-L)/100, U + k(U-L)/100]$. Typically, the 5th, 50th, and 95th percentiles are elicited. This creates a probability vector with 4 quantile intervals, $p = (0.05, 0.45, 0.45, 0.05)$. When assuming the piecewise uniform distribution, the expert density $f(e,i)$ is composed of a uniform distribution between each subsequent pair of values in the vector X: $(L^*, x_{0.05}, x_{0.50}, x_{0.95}, U^*)$. The information score is calculated by comparing the interquantile range to the bin size p_i:

$$I_{PWU}(e,i) = \log(U^* - L^*) + \sum_{i=1}^{4} \left[p_i \cdot \log\left(\frac{p_i}{x_{i+1} - x_i} \right) \right] \tag{4.6}$$

If an expert would estimate values for the three percentiles that result in a uniform distribution on $[L^*, U^*]$ the information score would be zero. Any deviation from this results in an information score above zero.

When assuming the Metalog distribution, the interquantile probability is not uniformly distributed. The information score is therefore calculated by integrating the expert density $f(e,i)$ over the range $[L^*, U^*]$:

$$I_{ML}(e,i) = \log(U^* - L^*) + \int_{x=L^*}^{U^*} \left[p(x) \cdot \log\left(\frac{p(x)}{dx} \right) \right] dx \tag{4.7}$$

In which $p(x)$ is the uniform background probability density in the range $[L^*, U^*]$. However, the unbounded version of the Metalog ranges from $-\infty$ to ∞ and is thus not limited to $[L^*, U^*]$. Using the infinity range would lead to a probability density of zero for the background probability and an infinite informativeness for the expert. Therefore, only the range $[L^*, U^*]$ is considered for calculating the informativeness. Because of the need for limits on the background probability density, a choice for the overshoot is still required when using the Metalog to interpolate expert percentile estimates.

4.B.2. Method for fitting a distribution to varying percentiles

Not all three-percentile or five-percentile expert estimates result in feasible Metalog distributions (i.e., $f(x) > 0$ for all x, with $f(x)$ being the PDF of x). For symmetric three-percentiles cases, the constraints for a feasible Metalog are given by $a_2 > 0$ and $|a_3|/a_2 < 1.66711$. For an unbounded Metalog distribution with a 5th percentile value of 10, and a 95th percentile value of 90, the median should be in between 20 and 80. If this is not the case, a feasible distribution can be achieved be imposing a lower or upper bound such that the constraints are met. This leads to highly skewed distributions, with one bound and one very thick tail. For the 3-percentile case, the steps in fitting a Metalog distribution are:

1. Check whether the linear least squares fitted a-vector is feasible.
2. If not, determine whether the expert estimate is left or right skewed. Find the lower and upper limit of the of the lower bound (left skewed) or upper bound (left skewed) that meets the constraints (i.e., $a_2 > 0$ and $|a_3|/a_2 < 1.66711$).
3. Iterate towards the bound that results in the distribution with the lowest maximum probability density (i.e., the least informative distribution).

For the five-percentile case it is more difficult to obtain a feasible fit. This is partly due to many expert estimates being more or less uniform in between the 5th and 95th percentile, such as shown in Fig. 4.2 e. For non-symmetric cases this often leads to not finding a feasible solution using a 5-term a-vector. Figure 4.2 h. is an example of this. In some cases, a solution is to impose a bound on one side just like for the skewed three-percentile cases. However, this does not work in all cases. Another solution is to use two three-percentile Metalog distribution, one representing the 5th, 25h and 50th percentiles, and one representing the 50th, 75th and 95th percentiles. This leads to a discontinuity in probability density at the median, which can be removed by imposing bounds on one (or both) of the 3-percentile Metalog parts (e.g., adding an upper bound to the left part in Fig. 4.2 g). We chose not to do this as it is only an aesthetic solution, which affects the tail probabilities and tends to create spikes in the probability density functions. The steps for fitting a five-percentile Metalog are therefore:

1. Check whether the linear least squares fitted a-vector is feasible.
2. If not, find two feasible three-percentile Metalog distributions (following the steps above).
3. Optionally, remove the discontinuity in probability density at the median, by iteratively shifting the lower or upper bound of the right or left distribution (whichever has the lowest probability density at the median) until the gap is removed.

5

Reliability analysis of the Dutch flood defense system

The Wettelijk Beoordelingsinstrumentarium (WBI) is the legal set of instruments for flood risk analysis in the Netherlands. Often, engineers have the impression that some failure probabilities of flood defenses resulting from these instruments are overestimated. In an effort to better estimate the failure probabilities of dikes along the Dutch river Rhine, this study sets out to assess them with experts and compare them to model results. We used the Classical Model (a.k.a., Cooke's method) for combining experts' estimates in a structured way and follow two approaches to estimate a system failure probability. In the first approach, experts estimate discharges that lead to at least one dike failure. This gives plausible results; failure probabilities between 1/30 and 1/17.000 in a year. The second approach is based on adjusting existing model-based assessment results, by estimating the model-bias and incorporating additional dependencies. This mostly leads to large, implausible, failure probabilities: Experts tend to give more conservative answers as they are asked for detailed estimates without clear reference values. This results in large uncertainty and consequently (too) high failure probabilities. Our research shows that when applied in a clear frame of reference, structured expert judgments can be successfully used for estimating the reliability of Dutch flood defenses.

The contents of this chapter have been published in: Rongen, G., Morales-Nápoles, O., & Kok, M. (2022). Expert judgment-based reliability analysis of the dutch flood defense system. *Reliability Engineering & System Safety, 224*, 108535.

5.1. Introduction

A large part of the Netherlands would regularly flood were it not protected by dikes. To prevent this, the Dutch have been building and maintaining dikes for centuries. Nevertheless, rivers and sea have flooded parts of the Netherlands dozens of times over the past centuries. Nowadays, dikes must have a failure probability less than 1/100 to 1/30,000 in a year to be considered safe according to legal standards (Ministry of Infrastructure and Environment, 2016). These small probabilities correspond to breach events during extreme river discharges or storm events that have never been measured in recent history, such that failure probability calculations cannot be verified used empirical data. Thus, modeling is needed to estimate probabilities of such extreme flood events and possible dike failures.

The use of probabilistic methods for design and assessment of flood defenses is common practice in the Netherlands (see (Vrijling, 2001) for a historical account and (Torres-Alves & Morales-Nápoles, 2020) for a recent example of the use of the traditional Dutch standard). The Wettelijk Beoordelingsinstrumentarium (WBI) is the current official set of models and tools for flood risk assessment (Slomp et al., 2016). ILT – Informatiehuis Water (2024) show that the methods generally give plausible reliability assessments for a short stretch of dike. However, the combined failure probabilities are overestimated when these assessments are combined for all dikes in the Netherlands. If such probabilities were correct, the Netherlands would have a dike breach every few years, which we know from experience is not the case. This indicates that parts of the WBI method are likely inaccurate. If this is ignored and the results are treated as correct, major (and probably economically unfeasible) dike improvements throughout the Netherlands would be enforced, because the Dutch safety standards is based on an optimal economic safety target (Dupuits et al., 2017; Jonkman et al., 2011; Kind, 2014). Hence, overestimating failure probabilities leads to ineffective spending of public money.

This study sets out to derive more plausible dike failure probabilities with the aid of experts in the field of flood risk management. We follow two approaches to estimate the system failure probability, both based on estimating fragility functions or fragility curves (Sayers & Meadowcroft, 2005; Van der Meer et al., 2009). These are widely used in flood risk assessments (Kok et al., 2017; Nofal et al., 2020). The first approach directly collects the experts' estimates of dike failures on the system-level (i.e., the probability that at least one dike fails) to derive more plausible estimate of the system failure probability than currently follows from the WBI. The second, more detailed, approach collects detailed estimates of failure mechanisms on a dike level and combines these with model-based results, considering load dependence. From this, we determine how credible the experts' estimates are. Additionally, it might indicate the cause of the implausible results following from the WBI-method. The two approaches discussed in this research (direct and more detailed) relate to research on risk estimates from fault trees as discussed for example in (Fischhoff et al., 1978; Fox & Clemen, 2005).

To account for dependence in hydraulic loads, relevant studies often rely on Monte Carlo-based reliability methods (e.g., Curran et al., 2020; Dupuits et al., 2019; R. B. Jongejan et al., 2020; Klerk et al., 2014). In this study, we assume both full dependence in load

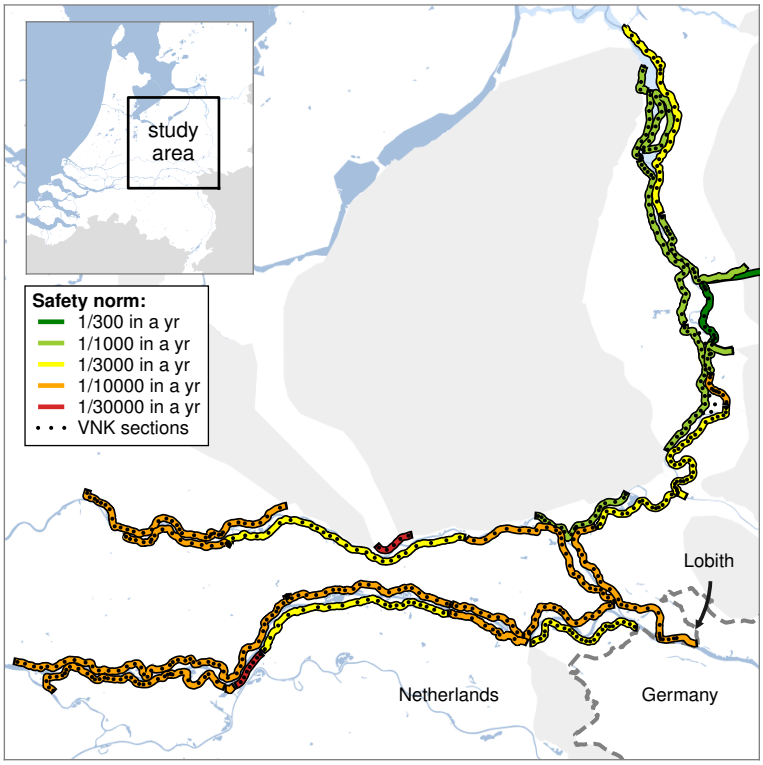


Figure 5.1: Overview of the part of the river Rhine considered in this study. The colored lines indicate the dike sections' norm, while the black dots represent the smaller VNK sections.

(i.e., a correlation coefficient of 1.0) and independence in dike strength (i.e., a correlation coefficient of 0.0). This gives a conceivable dependence structure befitting the water system and available data. To ensure scrutability, empirical control, neutrality, and fairness, when gathering expert estimates, we applied structured expert judgment with the Classical Model, also known as Cooke's method (Cooke & Goossens, 2008; Hokstada et al., 1998). A similar study to estimate reliability of French flood defenses was conducted by (Hathout et al., 2019). This study's focus is however the system scale rather than the reliability of individual dikes.

The area under investigation is the part of the Dutch river Rhine without tidal or lake level influence. This concerns the river branches Waal, Nederrijn, Lek, Pannerdensch Kanaal and IJssel, as shown in Fig. 5.1. The system consists of 28 dike trajectories, for which the color indicates the required safety level or norm. This is the maximum allowable probability of failure of the trajectory in a year. The 525 black dots indicate the dike sections as used in *The Flood Risk in the Netherlands* project (VNK2) (R. Jongejan & Maaskant, 2015; Projectbureau VNK2, 2010).

5.2. Method for safety assessments with expert estimates

This section explains the methods applied in this research, being the dike failure mechanisms (Section 5.2.1), the WBI method for dike safety assessment (Section 5.2.2), how expert judgment is incorporated in it (Section 5.2.3), and finally, the practicalities of the expert elicitation (Section 5.2.4).

In line with this thesis, this research applied the Classical Model for structured expert judgment, a method for eliciting and combining expert judgments based on empirical control, with the aim to reach rational consensus (see Section 2.1 for an elaborate description). The experts' assessments were processed with ANDURYL v1.2, which was presented in Chapter 3.

5.2.1. Failure mechanisms

Dike failure due to the failure mechanisms overtopping and overflow, piping, and macro-instability are considered. These are the most important mechanisms for the river Rhine, based on the failure probabilities calculated in VNK2. 't Hart et al., 2016 gives a description of these failure mechanisms. The WBI method provides a model for calculating failure probability for each of these mechanisms. The piping failure probability is calculated with the adjusted model of Sellmeijer (Knoeff et al., 2009), and the macro-instability failure probability with D-Stability (Van der Meij, 2019). (Geerse, 2011) gives a description of the method for overtopping and overflow failure, while the critical overtopping discharges are derived from (van Hoven, 2019). Figure 5.2 shows a typical dike section for which the three failure mechanisms are considered. This is a simplification; slightly different characteristics were considered for each mechanism.

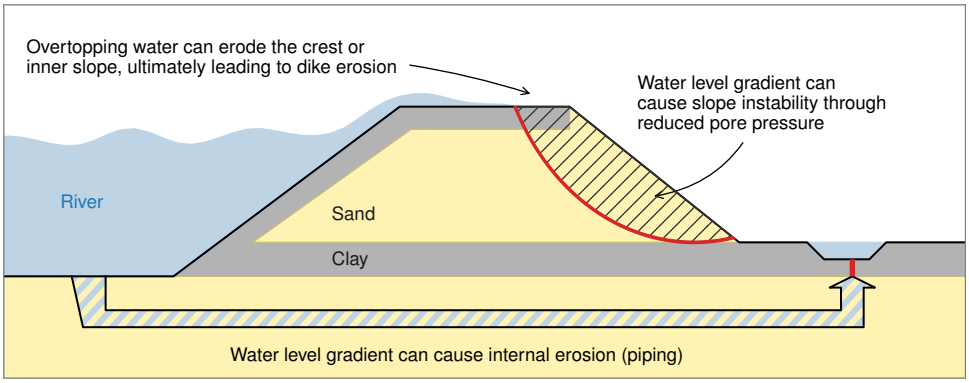


Figure 5.2: Conceptual cross section of a typical dike section and the considered failure mechanisms

5.2.2. Current WBI-model for a dike section's failure probability

A dike in the Netherlands has a safety norm, that is, a maximum allowable failure probability in a one-year period. A dike *section* consists of several stretches that together must

meet this standard. A stretch of dike with comparable properties is a dike *segment*, which can fail through to various failure mechanisms. Ultimately, the combined failure probability of all dike segments and all failure mechanisms, calculated with the WBI, must meet the safety level for the section.

The failure probability of a dike section is a function of the failure probabilities of the separate segments:

$$P_{sc}(Z<0) = \min\left(N \cdot \max_i\left(P_{sg,i}(Z<0)\right); 1-\prod_i^n\left(1-P_{sg,i}\right)\right) \tag{5.1}$$

In which $P_{sc}(Z<0)$ is the section failure probability, Z is the limit state function (less than zero implies failure) and $P_{sg,i}(Z<0)$ is the failure probability of segment i. For conciseness, we write $P(Z<0)$ as P_f . Section 2.2.2 explained how Eq. (5.1) combines segment failure probabilities independently, unless a lower failure probability is achieved by considering the number of times N a representative distance fits into the section length.

The failure probability of a segment i is calculated from the independent combination of the different failure mechanisms:

$$P_{f,sg} = 1-\prod_{j=1}^{3}\left(1-P_{f,sg,j}\right) = 1-(1-P_{f,sg,pip})\cdot(1-P_{f,sg,ms})\cdot(1-P_{f,sg,oo}) \tag{5.2}$$

In which j denotes the failure mechanisms, piping (pip), macro-instability (ms), and overtopping and overflow (oo). The failure probability is calculated per event, which for the rivers is a single flood wave. The standard WBI procedure is to consider six months per year during which every month can have one flood wave (Geerse, 2011). This means the maximum exceedance frequency of the limit state is six times per year. Determining the failure probability for a single segment (i.e., one of the parts in Eq. (5.2)) is challenging. Especially for the geotechnical mechanisms piping and macro-instability, for which the missing knowledge of the subsoil and saturation causes large uncertainty.

The high water levels in the study area are caused by the same high discharge at Lobith (Chbab, 2017). This simplifies the dependence; given the peak discharge at Lobith, the peak water levels in the complete the river system are known with limited uncertainty. We simplify this by assuming that the local water levels are directly related to the discharge at Lobith.

A comprehensible way of expressing the relation between load and (conditional) failure probability is with a fragility curve, which expresses the failure probability given the load (Van der Meer et al., 2009). A load can be, for example, water level, wind speed, discharge, or any combination of these (in which case we would have a fragility (hyper)plane (Nofal et al., 2020). In case of a peak discharge k, the failure probability can be expressed as:

$$P(Z<0) = \int_{k=0}^{k=\infty} f(k)P(Z<0|k)\,dk \tag{5.3}$$

In which $f(k)$ is the probability density of the peak discharge in a year. By integrating the conditional exceedance probabilities (i.e., fragility curves) with the probability

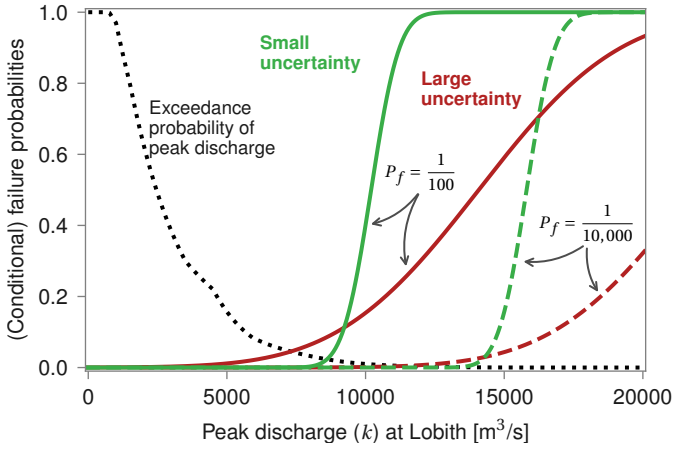


Figure 5.3: Examples of fragility curves (solid and dashed lines) that represent different failure probabilities and degrees of uncertainty.

density of the water level, the total failure probability in a year is calculated. Examples of fragility curves are given in Fig. 5.3. The mildly sloped lines are fragility curves with large uncertainty about the discharge that leads to failure, where the steep curves represent less uncertain fragility curves. Because the exceedance probability of more extreme events often decreases roughly exponentially, the lower (left) tail of the fragility curve is most determinative for the total failure probability. The fragility curve is used throughout this study because it can be related to experts' uncertainty estimates. The answer to the question: "At what discharge does dike X fail by mechanism Y with a 50% probability?", is similar to, "at what k is $P_X(Z_Y < 0|k) = 0.5$?". This represents a point on the curve in Fig. 5.3. Additionally, by expressing a failure probability in the form of Eq. (5.3), the dependence of water levels to the discharge at Lobith can be considered as well, by combining the failure probabilities conditional to k. This is illustrated in Fig. 5.4, where the combination of all blue dots at 12,000 m^3/s (conditional failure probabilities) are independently combined to the red dot, which is the combined failure probability conditional to the peak discharge k. The red line itself is also a fragility curve. Using the dependence between water levels along the river (through the peak discharge at Lobith) is not standard procedure in the WBI-method. It was added it in this study because it arguably gives a more accurate description of dike failure (see Section 2.2.2 for a discussion on this subject).

5.2.3. Dike assessment with experts

The two approaches followed for estimating failure probabilities with experts are both based on quantifying the term $P(Z < 0|k)$ in Eq. (5.3):

1. In the first approach, experts are asked to estimate the *system* conditional failure probability, $P_{sys}(Z < 0|k)$, i.e., at what discharges they believe a dike would fail. Integrating these estimates with the discharges as shown in Eq. (5.3) gives the total

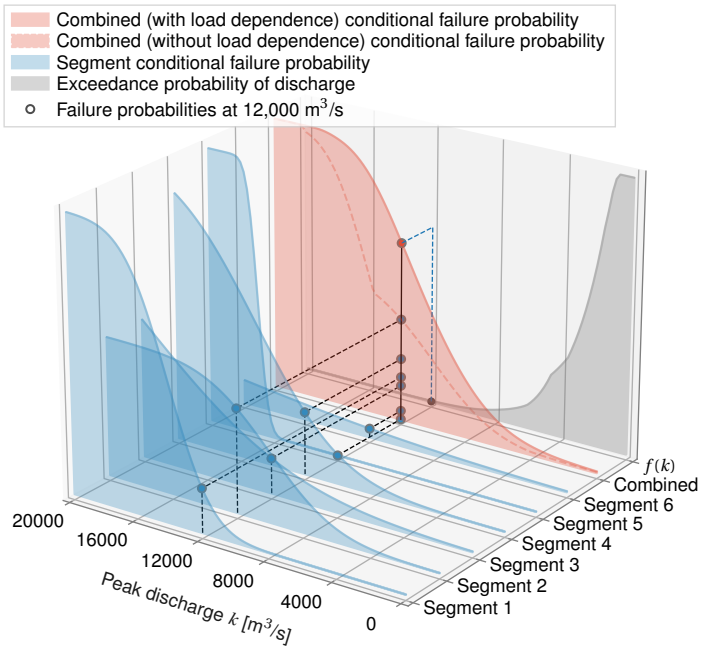


Figure 5.4: Illustration of how failure probabilities are combined, conditional to the load. The red solid curve considers *strength* independent, the dashed red curve considers *strength* dependent. The total failure probability results from integrating the conditional failure probabilities with the discharge statistics (indicated with the black curve).

failure probability. The peak discharges at Lobith were adopted from the WBI; they are not an elicited quantity in this study.

2. In the second approach, experts are questioned about several aspects of the safety assessment on a dike scale. These are the bias in the model-based failure probability calculations per failure mechanisms and the uncertainty in the relation between load and failure probability. These estimates are used to modify existing model outcomes and combine this to the term $P_{sys}(Z<0|k)$, from which the total failure probability is calculated.

The first option is relatively easy for the experts to relate to experience but gives little information about failure probabilities on the dike level. This information is more readily available from the second option, but it is hard for the experts to relate this to a total failure probability of the system. Therefore, we use the first approach to derive a more credible probability estimate, and the second approach to assess the values of these estimates and determine what causes potential differences.

For the second approach an estimate of the failure probability for each dike segment (represented by one of the black dots in Fig. 5.1) is needed. Recall that the study area contains 525 of such segments. Eliciting a failure probability from each expert for each

of these segments would be unrealistic within this study. Therefore, the individual failure probabilities were derived from two reference sources:

- For the current dike safety, these are the failure probabilities from VNK2 (Projectbureau VNK2, 2010). This project was finished in 2015 and considered the near future system layout (including the river engineering works 'Room for the River'), making it a useful source for the assessment of dike safety at present.
- The dike safety levels required by law (norms or standards). When a dike is designed by the WBI-method to meet this standard, we know that the failure probability of a trajectory should be lower than the safety standard (when assessed with the same tools with which it was designed.)

The total failure probability can then be calculated by updating them based on expert judgments and combining them into the system failure probability. This is expressed in the following equation:

$$P_{sys}(Z<0) = \int_{k=0}^{\infty} f(k)\Big(1 - \prod_{i=1}^{N_{sg}} \prod_{j=1}^{3} \{P_{ij}(Z<0|k-\Delta k_{ij})\}\Big) dk \tag{5.4}$$

in which j denoted one of the three failure mechanisms. A potential load-reducing effect of the first breach on the next breaches (a relevant topic for dike reliability (Dupuits et al., 2019)) is not taken into account, because the system is considered to have failed after the first breach. The experts' uncertainty in the discharge at which the dike fails determines how much $P(Z<0|k)$ varies for different values of the peak discharge k. The estimated bias in the failure mechanism shows whether failure takes place at consistently higher or lower discharges. It is determined by comparing the failure probability from the experts estimate with the model results. The quantity Δk_{ij} is calibrated such that the difference in failure probability is matched after integrating. Referring back to Fig. 5.3, the uncertainty expressed by the rate of change in P_{ij} for variations in K, indicates whether the fragility curve behaves more like the steep (small uncertainty) curve or gradually rising (large uncertainty) curve in Fig. 5.3. The bias (Δk_{ij}) indicates whether it is one of the solid curves or dashed curves in the same figure.

In this study, dike properties are assumed to be constant during an event. Therefore, only the highest water level in the event is considered when calculating the failure probability for piping and stability. For overtopping wind-generated waves play a key role, which vary on shorter time scales than water levels. Consequently, they are considered on a smaller time scale of 12 hours, as explained in (Geerse, 2011). The failure probability given a peak discharge k is then calculated with:

$$P(Z_{ov}<0|k) = P_{30\mathrm{d}}(H_{ov}>h_{ov}|k) = 1 - \prod_{t=1}^{N} \left[1 - P_{12\mathrm{h},t}\{H_{ov}>h_{ov}|q(t|k)\}\right] \tag{5.5}$$

Where $P_{30\mathrm{d}}(H_{ov}>h_{ov}|k)$ is the exceedance probability of the critical overtopping discharge h_{ov} in the 30-day duration related to peak discharge k. $P_{12\mathrm{h},t}(H_{ov}>h_{ov}|q(t|k))$ is the exceedance probability of this quantity in a 12-hour window t which depends on

the average river discharge during these 12 hours $q(t|k)$. The 12-hour average river discharge $q(t|k)$ is a function of a standardized hydrograph shape and peak discharge k. The conditional failure probability in Eq. (5.5) can be substituted in Eq. (5.4) as one of the failure mechanisms j. Compensating for a bias is not necessary, as the expert's assessment for the critical overtopping distribution H_{ov} can be estimated directly. For a detailed explanation we refer to (Geerse, 2011). Note that wind conditions are assumed spatially uniform in this study.

5.2.4. Questionnaire, experts, and practical set-up

The conducted questionnaire contains 22 questions, which are listed in Section 5.A. The questionnaire was accompanied with a detailed problem description per question. The first 10 items are seed questions, used to determine the experts' statistical accuracy. Items 11 and 12 are used to determine $P_{12\text{h}}(h_{ov} > H_{ov})$ from Eq. (5.5). After which item 13 to 16 address the bias for the piping and macro-instability models (Δk_{ij} in Eq. (5.4)). Then, items 17 to 20 concern the uncertainty in the load at which failure occurs (the rate of increase in $P(Z < 0|k)$ for changing k). Finally, items 21 and 22 were used to assess the failure probability on system level.

13 experts participated in the elicitation. Table 5.A.1 in lists their names, professional interests, and expertise. Most experts are flood risk advisor, affiliated to national and regional governments' water authorities, universities, research institutes, consultancy firms and independent consultants. Their expertise ranges from geotechnical engineering (related to piping and stability failure) to revetments (overtopping) and hydraulic loads (river discharges). The questionnaire was refined based on two dry runs with two different experts that have a similar background as the 13 that took part in the elicitation. The expert session was held with video conferencing due to COVID-19 pandemic restrictions. Two half-day sessions were organized with two occasions for each session. The first session contained a presentation of the study topic and the Classical Model, followed by answering the questionnaire individually. Experts were able to ask questions during the elicitation, which were then discussed plenary. The second session was used to present and discuss the results. To get their best assessment on the target items, experts had the opportunity to change their assessments in case they misinterpreted the question. Expert D used this possibility to change two seed items. Because the expert had seen the answers (realizations) at this point, the results were changed to 'not answered,' even though the questions were clearly wrongly interpreted. This led to slightly higher statistical accuracies for the other experts as well, as the minimum number of answered seed questions by all is used in calculating the statistical accuracy (see Eq. (2.1), N is 2 less). Expert F chose to withdraw from the expert session before filling out the target questions. This expert's results were removed, leaving twelve experts' assessments. An overview of the experts' and DMs' estimates for all items, is shown in Fig. 5.A.1 and Fig. 5.A.2.

5.3. Expert estimates and resulting failure probabilities

Section 5.3.1 presents the Classical Model results. Section 5.3.2 shows the experts' estimates for the system level failure, while Section 5.3.3 present the dike level estimates. The failure probabilities resulting from both are comparing in Section 5.3.4.

5.3.1. Results for the Classical Model

	Statistical Accuracy	Information score		Weight
		All	Seed	
Exp A	0.000304	0.923	0.696	0.000212
Exp B	0.057	1.000	0.980	0.0559
Exp C	0.00248	1.390	1.040	0.00258
Exp D	0.664	0.503	0.564	0.374
Exp E	0.64	0.664	0.777	0.497
Exp G	0.117	0.954	1.090	0.127
Exp H	0.0368	1.400	1.160	0.0428
Exp J	0.121	1.490	1.400	0.169
Exp K	0.0196	1.270	1.840	0.0362
Exp L	0.117	1.710	1.720	0.201
Exp M	1.02×10^{-5}	1.680	1.610	1.65×10^{-5}
Exp N	1.35×10^{-6}	1.760	2.030	2.73×10^{-6}
Global	0.64	0.365	0.399	0.255
Item	0.64	0.464	0.479	0.306
Equal	0.571	0.313	0.302	0.172

Table 5.1: Calibration and information scores from the Classical Model. The bottom three rows show the decision makers.

The elicitation results for the Classical Model are shown in Table 5.1, showing the statistical accuracy and the information score. Looking at the statistical accuracies, expert E and D have a high score (≥ 0.5), experts J, G, L and B a significant score (≥ 0.05) and the rest a score below the significance level. The information scores show less variation, as is usual with the Classical Model. Note that the information scores for all items (column 2) and the seed questions (column 3) are similar, indicating that experts have answered similarly for both categories of questions. Note that experts with a high statistical accuracy tend to have a lower information score, and vice versa. The weight is the product of information and statistical accuracy, and therefore favors experts with a high statistical accuracy.

We calculated three decision makers: global, item and equal weights. The statistical accuracies for the three decision makers are all high. The information scores for the decision makers are lower than those of individual experts. This is because the DM estimates are a weighted mixture of all experts. This often makes it a much wider distribution than the individual estimates, which results in a relatively low weight for the DM. When opti-

mizing the significance level, only expert D with the highest statistical accuracy remains. This is both the case for the global and item weights. As this does not provide additional information, DMs with optimization were not presented.

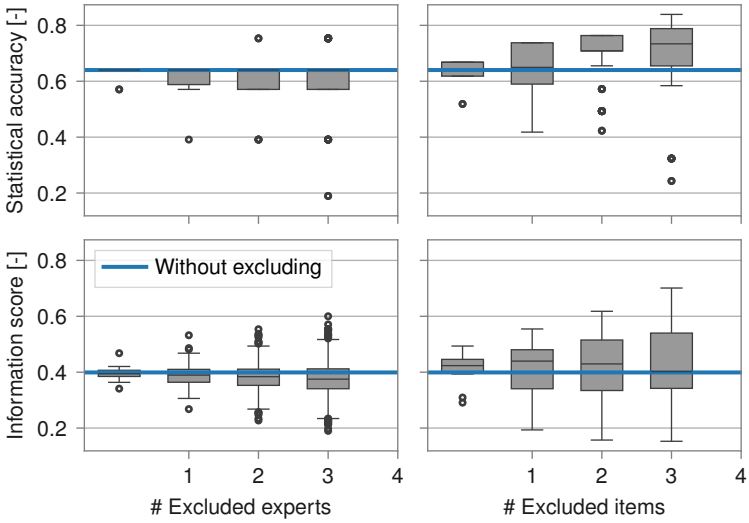


Figure 5.5: Sensitivity of calibration (top) and information score (bottom) for excluding experts (left) and items (right).

The robustness of the results for excluding experts and items are calculated for the Global DM. The variation in information and statistical accuracies when removing up to four items (experts) is shown in Fig. 5.5, which shows that the results are relatively insensitive to specific experts and items. The DM is most sensitive to item 7 (concerning discharge through a pipe), and expert D. Removing either leads to the largest reduction in DM weights. However, when excluding four items or experts at the same time, the statistical accuracy of the DM remains greater than 0.18 in any case.

5.3.2. System failure probability

The experts estimated the discharge for which at least one dike in the system fails. This question is part of the first approach in which system failure probability is estimated directly. Reference values for discharges, such as the 1995 extreme river discharge (around 12,000 m^3/s) and the design discharge of Room for the River (around 16,000 m^3/s) were discussed during the expert elicitation. Figure 5.6 shows the experts' and DMs' estimates. They are expressed as probability densities in between the assessed percentiles (i.e., 1, 5, 25, 50, 95). The thicker the squares, the higher the probability density in that range. The black diamond indicates the 50^{th} percentile. The DM estimates consist of more than four blocks (corresponding to the quantiles intervals) because they are a weighted combination of more than one expert.

For the present (at the time of elicitation) state of the dikes, most experts estimate at

least one dike to fail at a discharge in between 12,000 and 16,000 m^3/s, while for the dikes matching the norm, this is in between 14,000 and 18,000 m^3/s.

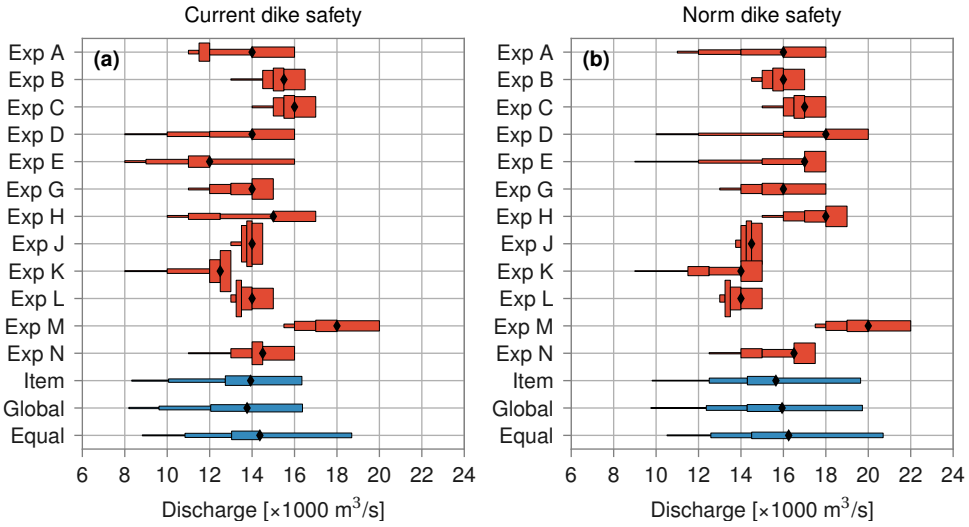


Figure 5.6: PDFs of the discharge that causes at least a single dike failure in the system, under current dike conditions (a) and norm safety conditions (b). The experts' estimates are shown on top, the DMs' in the bottom three rows.

5.3.3. Failure mechanism factors

The experts estimated the bias in the models for the relevant failure mechanisms as well (i.e., the Δk_{ij} in Eq. (5.4)). Together with the estimates of the general uncertainty (see Section 5.3.3) these are used for the more detailed second approach. Experts assessed hydraulic loads that lead to the WBI failure definition as well as a breach, which is considered a more progressed state of failure that occurs after the WBI (initial) failure state is reached. The first expresses the bias in the model (which calculates the WBI failure state), while the latter expresses the bias in the probability of flooding (resulting from a breach). The experts were asked to assume that: i) that the high water level lasts for 5 days, and ii) no emergency measures (e.g., sandbags) are used to prevent failure.

Piping

We chose a single, representative, schematization from the VNK2-dataset to estimate the bias in the 'adjusted Sellmeijer' model Knoeff et al., 2009. The cross section contained a single cohesive cover layer on top of a permeable layer (similar to Fig. 5.2). The failure probability was calculated from the experts' estimates of the water level at which they expect a breach to occur with a 1, 5, 25, 50, and 95% probability (implicitly constructing a fragility curve). The failure probability is calculated by combining this with the water level statistics. The results of this are shown in Fig. 5.7 a. The model result is a contin-

uous distribution of the water level at which the dike is expected to fail. However, for comparison to the experts' estimates, it is drawn for the same percentiles.

Figure 5.7 b shows the factor between the failure probability from the adjusted Sellmeijer model, 9.31×10^{-4}, and the failure probability from the expert. Note that the two experts with highest statistical accuracies, expert D and E, estimate a 10 to 100 times higher failure probability. This is due to the uncertainty in their estimates, which leads to a substantial conditional failure probability for water levels that frequently occur. The Global and Item DM are similar to D's and E's estimates, as these experts together have the majority of the DM-weights. Other experts, like M, only contribute to any discernible extent through the equal weights DM.

Macro-instability

The bias for the D-stability model (for macro-instability assessments) is determined in an equivalent way to the piping bias. The evaluated cross section consists of a sand dike body with a silty clay layer on the outer slope and crest (as in Fig. 5.2). The results are shown in Fig. 5.7 c. The model for this failure mechanism shows a smaller range of water levels within which failure will take place, compared to the expert assessments. In other words, the experts are more uncertain about the water level that leads to failure than the model. Contrarily to piping, most experts estimate the macro-instability model to be conservative.

Again, experts D and E estimate a more conservative outcome than the model, and again this results in assigning a significant failure probability to frequently occurring water levels. The estimates for the water levels that cause macro-instability failure show less variation than those for piping. The plenary discussion showed that it was easier to narrow the relevant water level range for macro-instability than it was for piping. This is because experts find it easier to indicate water levels that will have for both small and large failure probabilities. However, the ability to do so depends largely on the domain knowledge of the expert.

Overtopping

In addition to the water level at the dike, waves can cause failure due to overflow and overtopping. Therefore, instead of the water levels, the distribution of the critical overtopping discharge is elicited $P(H_{ov} > h_{ov})$, which was then used to determine the dike failure probability directly. Estimates were made for an open grass cover on top of a 0.5 m clay layer with a sand dike body. Figure 5.7 e shows the experts' estimates for the critical overtopping discharge that causes a breach when exposed for 6 hours. The model result is derived from the log-normal distribution for H_s less than 1.0 meter and an open grass cover. The bar heights in this panel do not represent probability density. The right graph gives the difference in failure probability compared to the model results for all 525 dike sections. According to eight of the experts, the failure probabilities from the WBI are too unconservative, for four experts they are conservative. The difference between the failure probabilities is small, most factors are between 1/5 and 2. Overtopping or overflow discharge increases rapidly with rising water levels, meaning that water levels

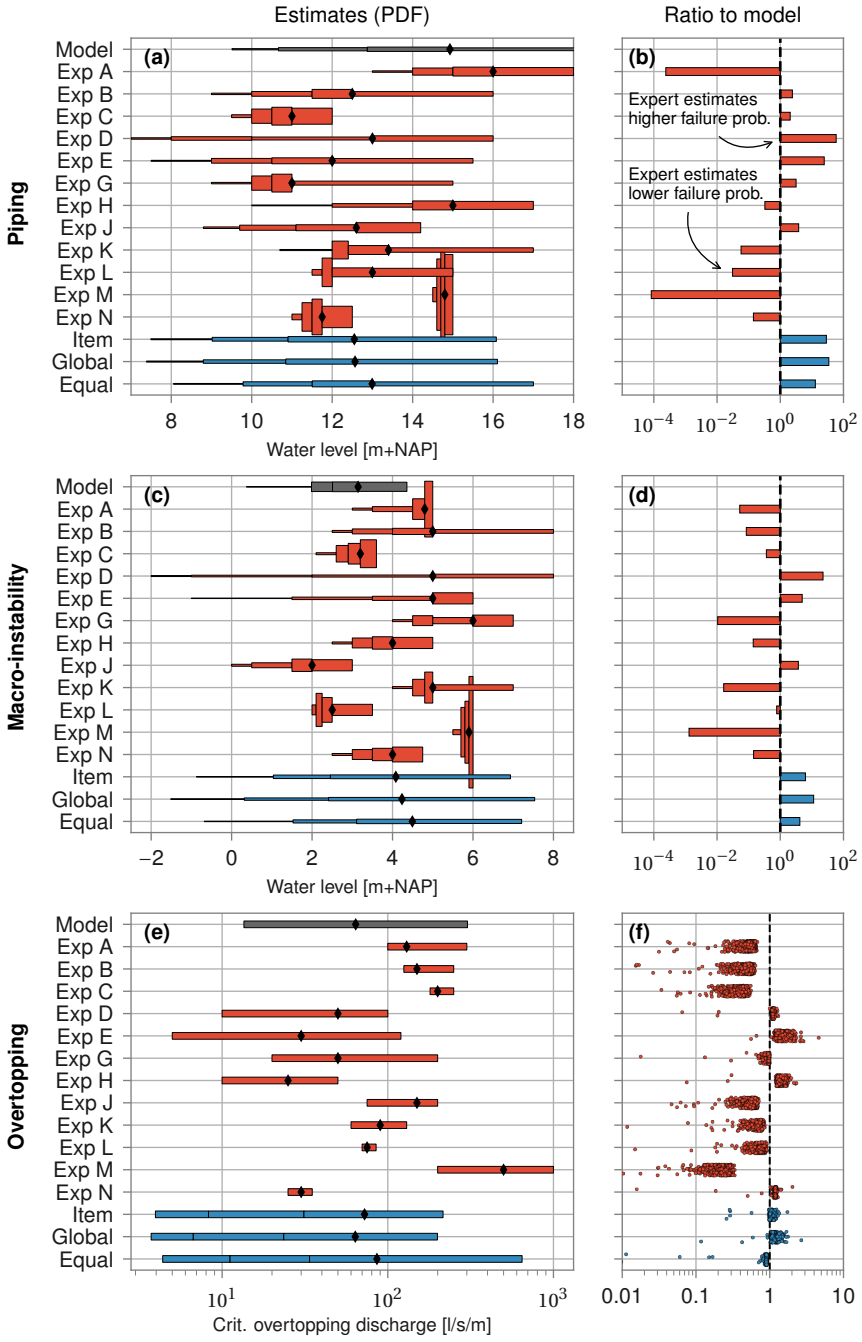


Figure 5.7: PDFs of the load at which the considered schematization is expected to fail due to piping (a), macro-instability (c) or overtopping (e), according to the model (top row in each panel), the experts (middle rows) and the DMs (bottom three rows). The ratio between expert or DM and model outcome is shown on the right (b, d, and f).

corresponding to low and high overtopping discharges are closer together than they are for piping or macro-instability, and consequently the exceedance frequencies as well.

General uncertainty in water level at failure

To calculate the total system failure probability in the second approach, we need an overall estimate of the distribution of failure probabilities conditional on the discharge (i.e., the $P(Z<0|k)$-part in Eq. (5.4)). The results of the experts' assessments for this are shown in Fig. 5.8.

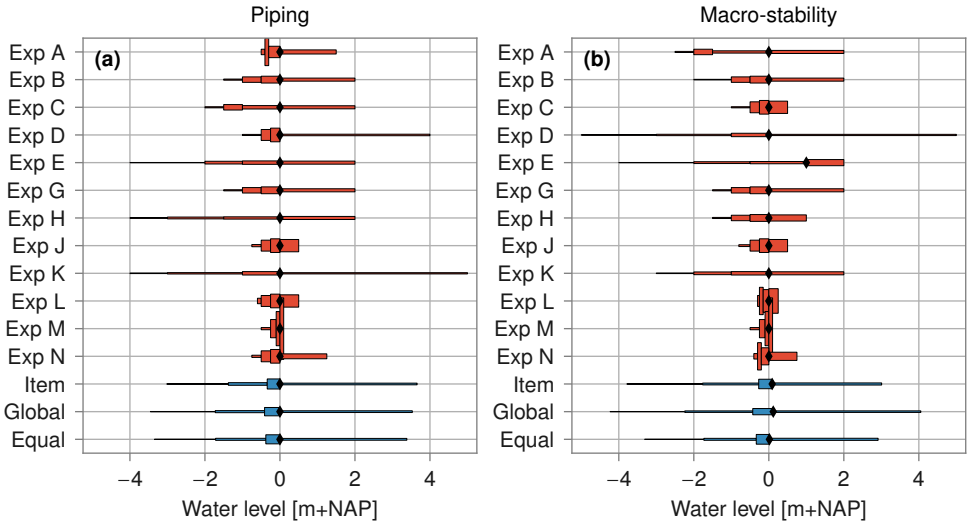


Figure 5.8: Uncertainty in the water level at which a dike fails for piping (a) and macro-instability (b), according to the experts (top rows) or DMs (bottom three).

During high water, a difference of 1000 m^3/s in the discharge (at Lobith) typically leads to a 25 cm difference in local water level. Extrapolating this, a range of 4 meters would give a 16,000 m^3/s discharge difference. This very wide range does however spans from discharges with a very small (1%) to a large failure (95%) probability.

The differences between the experts' estimates are, again, large. Experts L, M and N give a very small, almost deterministic, range of water levels that cause failure. On the contrary, experts D, E and K are very uncertain about the water levels at which a dike will fail due to piping or stability.

5.3.4. Comparing the failure probabilities from the two approaches

Recall that we estimated system failure probability in two ways: by integrating the estimates on system level (Section 5.3.2) with the discharge statistics, and by combining the estimates on dike level (bias, and uncertainty, Section 5.3.3) with model-based results through Eq. (5.4). Figure 5.9 shows the resulting system failure probabilities for the current dike safety level (a), and for the situation in which all dikes match the safety

standard (b). Each of the figures contains two bars per expert or DM. The first repre-

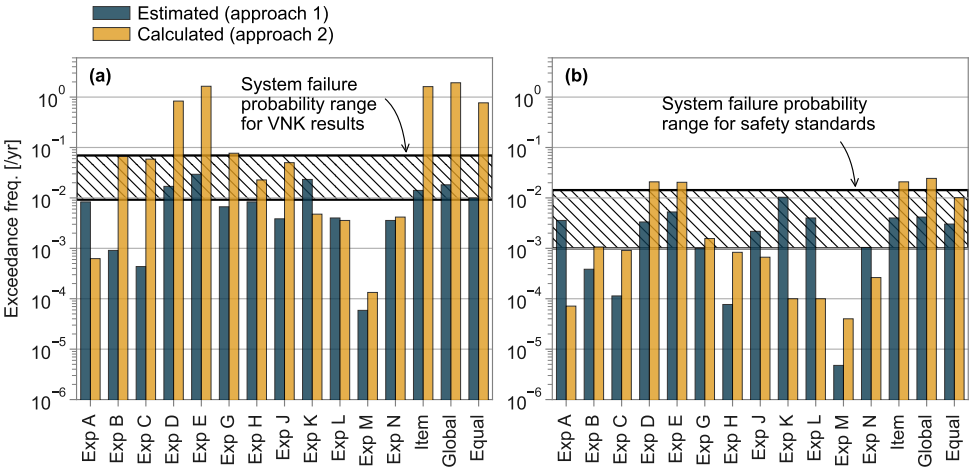


Figure 5.9: Estimated and calculated system failure probabilities, for current dike safety (a) and dikes matching safety standard (b).

sents the failure probability from the first approach (i.e., the system failure estimates), the second bar the failure probability for the second approach (i.e., updating model results with estimates). The figures also contain a horizontal hatched area that indicates the range of the reference failure probability. For the current dike safety situation, this is the range between the independent and dependent combination of the VNK2 section failure probabilities: 1/15 to 1/100 in a year. The reference value for the norm safety is the range between the independent and dependent combination of the individual dike safety levels: 1/70 to 1/1000 in a year. Note that we combined the the values displayed in Fig. 5.1 assuming full dependence for the upper bound, and the 'signal' values (about a factor 3 lower than Fig. 5.1) assuming independence to obtain the lower bound. This is the dike safety level that signals a dike should be reinforced soon (Ministry of Infrastructure and Environment, 2016), so a dike will be reinforced to at least this level (as its reliability constantly decreases due to deterioration (Chen & Mehrabani, 2019)).

All experts estimate (approach 1) a slightly smaller to much smaller failure probability, while the calculated answer (approach 2) is significantly higher, ranging from 1/10,000 in a year to larger that once in a year. For the situation in which the dikes have a safety level up to the norm, the estimates are in the range of 1/100 to 1/10,000 in a year, while the calculated answers are in the range 1/10 to 1/17,000 in a year. Looking at the estimated failure probabilities, we see that, for the current dike safety, experts assess the system failure probability to be roughly 1/100 in a year (ranging from 1/30 to 1/30,000). For the norm safety situations, experts estimate the failure probability to be 1/1000 year (ranging from 1/100 to 1/200,000).

The differences between the two approaches for an expert or DM can be quite large, up to a factor 100 in failure probability. In most cases, the calculated failure probability (approach 2) is larger than the estimated probability (approach 1), especially for the current

dike safety situation. The main cause of this difference is the uncertainty in the experts' estimates for the failure-critical water level for piping and macro-instability. Uncertain estimates assign a relatively large probability to failure during frequently occurring water levels. The exceedance probability for higher water levels decreases rapidly (more or less exponentially). This means higher conditional failure probabilities for low water levels contribute much more to the total failure probability. Experts that estimate a wide range of water levels for the different quantiles more likely get an implausible system failure probability. Note that the experts did not receive feedback on the implication of their answers during the expert session, as the goal was to observe which approach yields more credible answers.

5.4. Applicability for estimating dike failure probabilities

5.4.1. Application of the Classical Model

As explained in Section 5.2.4, the elicitation was organized in two half-day sessions. The first was used for filling out the questionnaire, the second to discuss the results. The questionnaire was not discussed with the participants in advance. From the discussion afterwards, we noticed that the questions' underlying assumptions can have a major effect on the estimated uncertainties. Despite the questionnaire being as clear as possible on the assumptions and context, experts still create their own image of the assessed dike failure. This was most evident in the question of the failure probability of the macro-instability schematization. A specific image of the failure process would be more determinative of the estimates than, for example, their thoughts on uncertainties in the model parameters.

Discussing the questions together beforehand is common for expert studies using the Classical Model but was not done in this study because of time and COVID constraints. Such a discussion can help to steer the experts into a similar way of reasoning, likely converging their estimates. This might lead to a dominant or convincing expert's, potentially wrong, viewpoint getting the upper hand in the DM through the answers of other (now influenced) experts. Still, such a discussion seems to be preferred for questions that are not a straightforward parameter estimation, or when the answer is difficult to relate to experience.

There is little empirical data for failure of Dutch dikes in their current state. Experts therefore need to think through a number of steps that lead to failure (in essence, a model) and quantify subsequent steps by indirectly assigning probabilities to them. The more steps, the greater the uncertainty in the final answer. In such situations a plenary discussion could have aided some experts. When a problem can be interpreted in several ways, experts can remind each other of conditions or ways of reasoning to reach a plausible answer. The questions about the piping and macro-instability schematizations, being examples of this. However, when referential values are available, as with the question about the failure probability of the river system, the estimates of the experts are closer together. It is common knowledge among flood risk experts that the Lobith discharge at high water in 1995 was approximately 12,000 m^3/s, and Room for the River was designed at a discharge of 16,000 m^3/s. These are values that may be used as reference

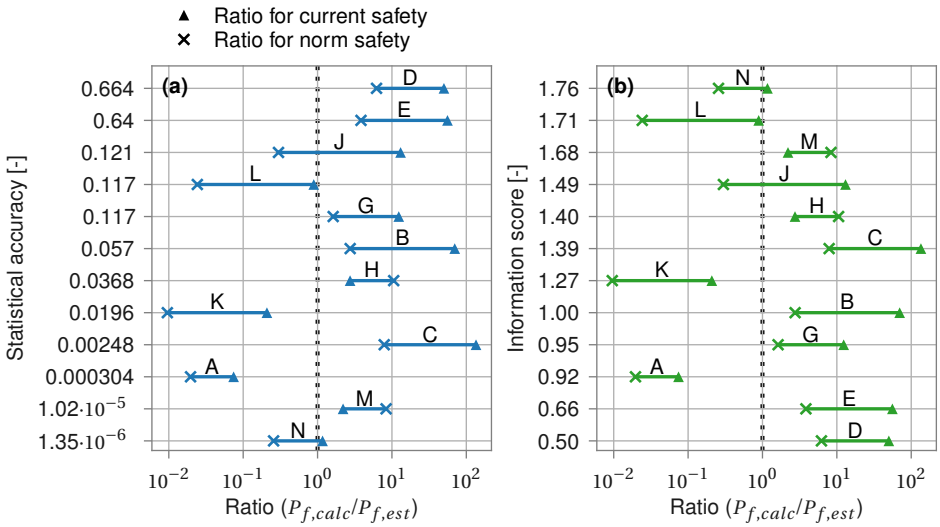


Figure 5.10: Relation between failure probability ratio and statistical accuracy (a) and information score (b).

values for estimates, which, according to the post-elicitation discussion, was done by at least some of the experts.

5.4.2. Experts score and consistency

To find out how experts perform for estimating small probabilities, we followed two approaches to derive a system failure probability. Given enough time and resources, an expert would ideally arrive at the same answer following both approaches. However, in Fig. 5.9 we observed large differences. Figure 5.10 plots these differences against the statistical accuracy and information score from the Classical Model.

Both figures contain, per expert, two markers with a connecting line. A triangle represents the current dike safety situation, while the cross represents the situation in which the dikes meet the safety standard. Expert N estimates a similar probability with the two methods, while, for example, experts K and C, have a larger deviation. We do not observe a clear relationship between statistical accuracy and consistent answers. Ratios close to 1.0 imply consistent answering because both approaches lead to the same failure probability. While consistency is not equal to correctness, a positive relationship between consistency and statistical accuracy supports the hypothesis that statistical accuracy expresses the experts' performance in giving credible estimates, independent of the method used for assessing failure probabilities. The comparison between estimated and calculated failure probabilities does not indicate a pruning bias (Fischhoff et al., 1978). The fact that i) we did not ask experts to estimate failure probabilities directly, and ii) asked experts to estimate system-level probabilities last (i.e., after being made aware of all the failure mechanisms), might contribute to this.

In the Classical Model, the statistical accuracy affects the DM solution more than the

corresponding information score. It generally favors the more uncertain experts: in 27 of the 33 studies considered in (Leontaris & Morales-Nápoles, 2018) the rank correlation between the information score and statistical accuracy is less than 0.0. The median rank correlation over all 33 studies is -0.4, indicating an inverse relationship between calibration and information score. Although the Classical Model is not infallible, the use of weight-based DMs has shown to outperform equal weighting (Clemen, 2008; Colson & Cooke, 2017b; Eggstaff et al., 2014). However, in this study a large uncertainty (meaning low information, and likely high calibration) dramatically increases the estimated failure probabilities. Consequently, we observe that the 'best' experts score as well as the DMs, estimate high failure probabilities (perhaps too high).

5.4.3. Model versus experts

This study presents expert judgment as an alternative to an approach relying on physics-based models. Reality is however less binary. A model-based approach for the Dutch flood defense system mostly means an engineer (or a group of engineers) uses a model for a safety assessment. The engineers however make several choices while doing so. First, what model to choose, and subsequently, which parameter values to use. These choices can be passive, for example when based on a standard procedure, but it is still a choice to adopt these. Similarly, experts can use models to substantiate their estimates. When viewed from this perspective, the difference between a model-based approach and expert judgment is mainly the focus: (Structured) expert judgment clearly puts this on the experts, who determine what model and parameter values to use. Standard procedures and defaults will more likely be questioned in this approach.

5.5. Conclusion

In this study we applied the Classical Model for structured expert judgments to estimate system failure probability for the Dutch river Rhine. To do so, we followed two approaches: One in which experts directly estimate the discharge at which at least one dike fails, and a second approach in which estimates of bias and uncertainty are used to adjust failure probabilities from models. The first approach results in failure probabilities from 1/30 to 1/17,000 in a year. For most experts, these probabilities contrast dramatically to the 1/15 to 1/100 in a year failure probability range resulting from the VNK2 failure probabilities. Even under the assumption that no emergency measures are used (e.g., placing sandbags), most experts assess a failure probability smaller than 1/100 in a year for the current dike state. For the situation in which the dikes are reinforced to match the required safety level, experts estimate a failure probability varying between 1/100 and 1/200,000 in a year. These estimates are more in line with the expected failure probability resulting from combining the safety standards independently (1/70 in a year) and fully dependent (1/1000 in a year). Thus, our expert pool expresses that in the current state, dikes are safer than the models say, while they estimate a level of safety that aligns the standards after the dikes have been reinforced.

Considering the more detailed results, half the experts estimated a failure probability for the piping schematization that is larger than the model suggests. This is surprising

because the experts' shared impression is that the adjusted Sellmeijer model results in failure probabilities that are too high. It would be premature to assume this impression is wrong only based on the single assessed schematization. It is however a surprising result. For macro-instability, most experts think the model is slightly conservative, but the differences are smaller than for piping. More schematizations for piping as well as macro-instability should be assessed to reach a general conclusion on a potential model bias.

The second approach, in which the failure probability was calculated by adjusting model results, gives total system failure frequencies ranging from larger than once in a year to 1/7500 in a year. The upper limits of these recurrence rates are unlikely high and are caused by the large uncertainty ranges given by some experts: Wide estimates for the failure-critical water level range assign too high failure probabilities to frequent events. Especially for the piping and macro-instability assessments, the lack of reference values leads to wide uncertainty estimates. On the other hand, the direct failure discharge estimates result in narrower uncertainty estimates as this conceptualization is easier to relate to reality. The questionnaire was not discussed with the experts before the elicitation. Especially for questions that leave room for interpretation, the absence of discussion and expert interaction allowed quite different, unmodulated, interpretations to be articulated.

Providing wide uncertainty estimates may result in a high statistical accuracy, and large expert weight in the Decision Maker. A positive relation between the consistency of each expert's estimates and the weight in the Classical Model would support the choice of using a global or item weights DM. However, there seems to be no significant relation.

To conclude, we found that experts estimated plausible probabilities of dike failure on a system level, while they struggled to answer the questions concerning dike sections accurately. Compared to a model-based approach, structured expert judgment has the advantage that uncertainties are made explicit, whereas in a model-based approach these may be hidden, latent or ill-defined.

To obtain reliable and defensible estimates of event and exceedance probabilities using expert judgment for hazards such as dike failure, it is desirable to establish a clear and, if possible, agreed framework of technical definitions, empirical observations, modeling assumptions, and established knowledge. Our study of expert judgment-based failure probability estimation for a system of river dikes in the Netherlands, explored some key aspects of this challenge, and how, quantitatively, they influenced our findings. Importantly, these point at certain aspects that merit further investigation by expert elicitation.

Appendix 5.A: Supplementary information

This appendix contains the supplementary information for the study:

- An overview of the questionnaire including realizations (between parenthesis) is shown in the list below. For questions 1 to 12 the 5th, 50th, and 95th percentiles were elicited. For questions 13 to 22 the 1st and 25th percentiles were elicited as well to increase accuracy in the lower tail.

- Table 5.A.1 lists the twelve participating experts with their affiliation and specialism. One expert wished to remain anonymous, bringing the total number of experts to the 13 mentioned in Section 5.2.4. The experts are ordered alphabetically by first name, which holds no relation to the letters used throughout the chapter.

- The experts' estimates are displayed in Fig. 5.A.1 and Fig. 5.A.2.

Overview of seed (item 1 to 10) and target questions (item 11 to 22):

1. Looking at the damage from the overtopping experiment in the photo: how long (in hours) did it take between the first visible damage and the damage in the photo? (0.75 h)
2. What is the highest discharge (m^3/s) that will occur at Lobith (where the Rhine flows into the Netherlands) in December 2020? *Note that the questionnaire was deducted before this month* (3000 m^3/s)
3. What is the wind speed (m/s) that is exceeded on average once a year at Deelen? (15.0 m/s)
4. What is the average difference on the Maas (between Venlo and Den Bosch) between the water level at a discharge at Borgharen of 3000 and 4000 m^3/s? (1.1 m)
5. Considering the provided information about the flood that followed the heavy rainfall during Typhoon Hagibis, how many of these 29 overflown dikes have failed? (10)
6. How many of the 142 dikes that failed in total during the floods after Typhoon Hagibis, where due to the failure mechanism piping? (2)
7. Considering the characteristics of the described piping experiment, what is the flow through the well when the critical gradient is reached? (2.3e-05 m^3/s)
8. What is the mean (μ) permeability (k) of the subsoil under the dikes of section 48-1? (0.00048 m/s)
9. What is the mean coefficient of variation (V = σ/μ) of the permeability (k) of the subsoil under the dikes of section 48-1? (0.679)
10. Given that a Rhine discharge of 6000 m^3/s is exceeded. What is the probability that the Meuse discharge of 1500 m^3/s is exceeded within a period of 10 days before or after the moment the Rhine discharge has been exceeded? (0.58)
11. At what overtopping discharge (l/s/m) do you expect these specific wave conditions to erode the 50 cm clay layer? In other words, what do you expect the overtopping discharge to be?
12. At what overtopping discharge (l/s/m) do you expect a breach in the dike (dike opening) to occur, again after 6 hours of wave attack?
13. For the given schematization, at what river water level (m+NAP) do you expect piping (an unstable pipe) to occur?
14. For the given schematization, at what river water level (m+NAP) do you expect a breach to occur as a result of a sand-carrying pipe?

15. For the given schematization, at what river water level (m+NAP) do you expect a deformation die to instability of the inner slope, with an entry point in the crest?
16. For the given schematization, at what river water level (m+NAP) do you expect a breach to occur as a result of an instability?
17. For river dikes in general, at what river water level (relative to 0 m+NAP) do you expect piping to start?
18. For river dikes in general, at what river water level (relative to 0 m+NAP) do you expect a breach due to piping?
19. For river dikes in general, at what river water level (relative to 0 m+NAP) do you expect macro-instability to cause a shearing of the inner slope with an entry point in the crest?
20. For river dikes in general, at what water level (relative to 0 m+NAP) do you expect a breach due to macro-instability?
21. At which peak discharge (m^3/s) do you expect at least one dike in the river system to fail for the current dike safety situation?
22. At which peak discharge (m^3/s) do you expect at least one dike to fail in the river system when all dikes meet the required safety level?

Name	Affiliation	Specialism
Carlijn Bus	Waterschap Brabantse Delta	Specialized in flood risk assessment of dikes.
Don de Bake	HKV	Senior advisor flood risk. Specialized in flood risk management, dike safety assessment, and dike restoration projects. Policy advisor to the Ministry of Infrastructure and Water Management.
Henk van Hemert	Rijkswaterstaat	over 25 years professional experience in dike projects with a geotechnical focus
Jan Blinde	Deltares	Flood risk, dike design, dike assessment
Jan Tigchelaar	HKV	Specialized in geotechnics and probability applied to dike failure and flood risk. Advisor in different national and international projects.
Jan-Kees Bossenbroek	Waterschap Hollandse Delta	Flood risk advisor, specialized in applying the flood risk approach and flood defense knowledge in the South-Holland Delta.
Leo van Nieuwenhuijzen	Waterschap Rijn en IJssel	Flood risk advisor, contact point for calamity care in case of imminent flood waves.
Marinus Aalberts	Witteveen+Bos	Senior engineer in flood risk and dike design. Member of Expertise Network for Flood Protection (ENW)
Philippe Schoonen	Waterschap Drents Overijsselse Delta	Technical manager Flood Protection, Coordinator innovation program
R.B. Jongejan	Jongejan Risk Management Consulting BV	Specialized in flood risk analysis and probabilistic design; independent engineering consultant
Stefan van den Berg	Rijkswaterstaat	Flood risk advisor in the execution phase of projects, with a focus on connecting theory and practice.
Wim Kanning	Deltares and Delft University of Technology	Expert in levee safety, geotechnical reliability, and risk.

Table 5.A.1: List of experts with their affiliation and professional interests.

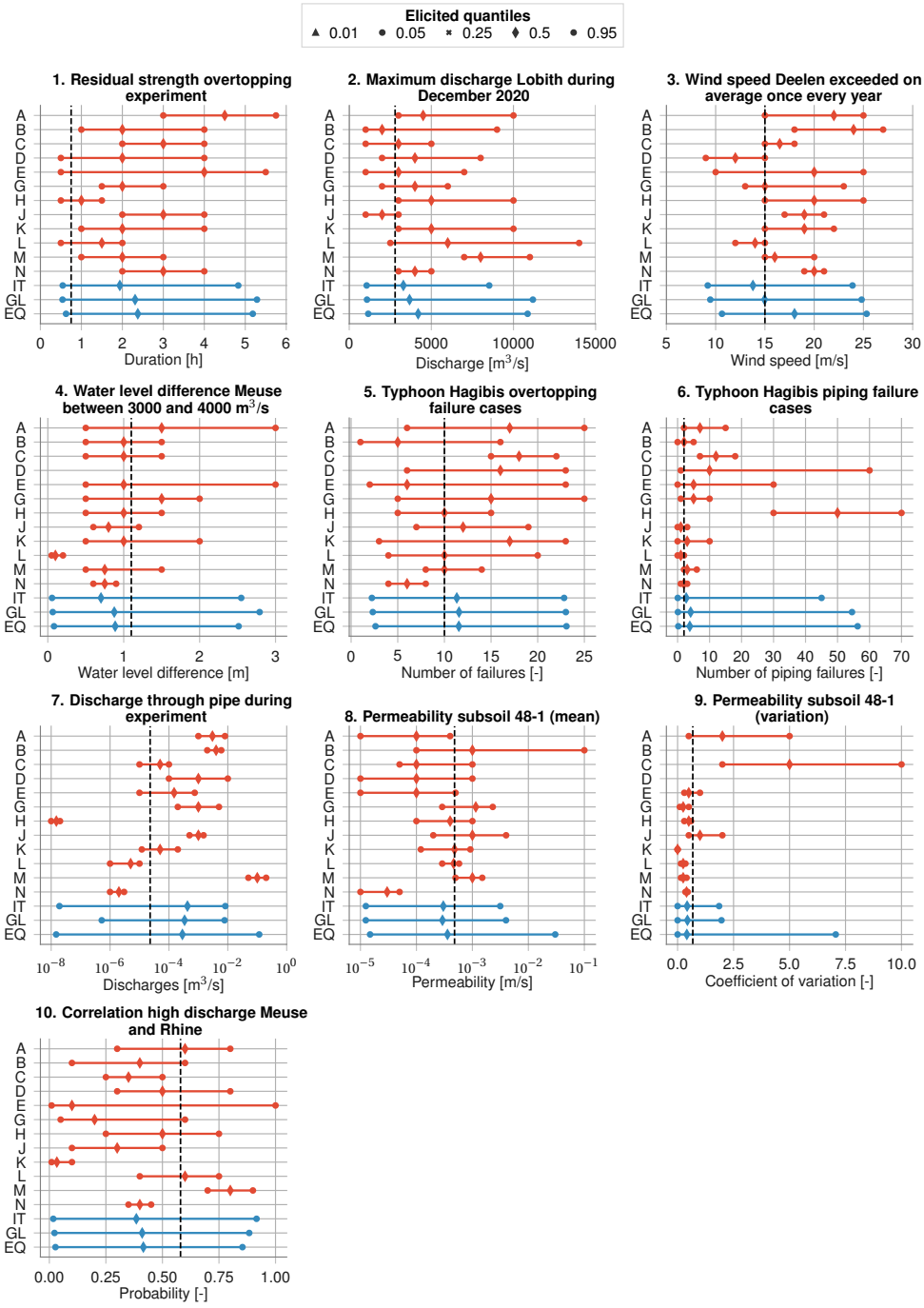


Figure 5.A.1: Expert and DM estimates for seed questions. The realization is indicated with the dashed vertical line.

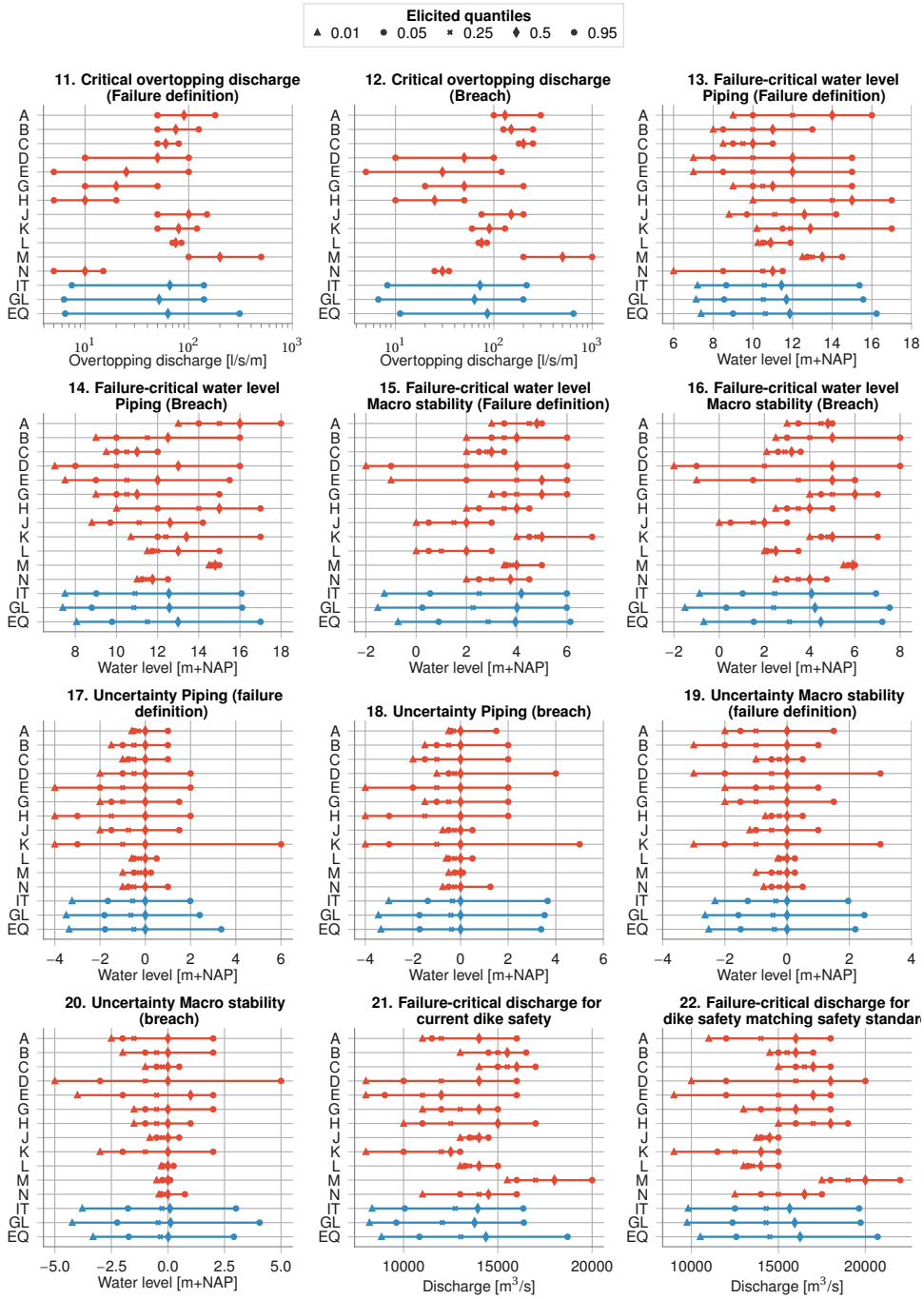


Figure 5.A.2: Expert and DM estimates for target questions.

6

Estimating extreme discharges on the Meuse's tributaries

Accurate estimation of extreme discharges in rivers, such as the Meuse, is crucial for effective flood risk assessment. However, hydrological models that estimate such discharges often lack transparency regarding the uncertainty of their predictions. This was demonstrated by the devastating flood that occurred in July 2021 which was not captured by the existing model for estimating design discharges. This article proposes an approach to obtain uncertainty estimates for extremes with structured expert judgment, using the Classical Model. A simple statistical model was developed for the river basin, consisting of correlated GEV distributions for discharges from upstream tributaries. The model was fitted to seven experts' estimates and historical measurements using Bayesian inference. Results fitted to only the measurements were solely informative for more frequent events, while fitting to only the expert estimates reduced uncertainty solely for extremes. Combining both historical observations and estimates of extremes provided the most plausible results. The Classical Model reduced the uncertainty by appointing most weight to the two most accurate experts, based on their estimates of less extreme discharges. The study demonstrates that with the presented Bayesian approach that combines historical data and expert-informed priors, a group of hydrological experts can provide plausible estimates for discharges, and potentially also other (hydrological) extremes, with a relatively manageable effort.

The contents of this chapter have been published in: Rongen, G., Morales-Nápoles, O., & Kok, M. (2024). Using the Classical Model for structured expert judgment to estimate extremes: a case study of discharges in the Meuse River. *Hydrology and Earth System Sciences, 28*(13), 2831–2848.

6.1. Introduction

Estimating the magnitude of extreme flood events comes with considerable uncertainty. This became clear once more on the 18th of July 2021: A flood wave on the Meuse River, following a few days of rain in the Eiffel and Ardennes, caused the highest peak discharge ever measured at Borgharen. Unprecedented rainfall volumes fell in a short period of time (Dewals et al., 2021). These caused flash floods with large loss of life and extensive damage in Germany, Belgium, and to a lesser extent also in the Netherlands (Mohr et al., 2022; TFFF, 2021). The discharge at the Dutch border exceeded the flood events of 1926, 1993, and 1995. Contrary to those events, this flood occurred during summer, a season that is (or was) often considered less relevant for extreme discharges on the Meuse. A statistical analysis of annual maxima from a fact-finding study done recently after the flood, estimates the return period to be 120 years based on annual maxima, and 600 years when only summer half years (April to September) are considered (TFFF, 2021). These return periods were derived including the July 2021 event itself. Prior to the event, it would have been assigned higher return periods. The season and rainfall intensity made the event unprecedented with regard to historical extremes. Given enough time, new extremes are inevitable, but with the Dutch flood safety standards being as high as once per 100,000 years (Ministry of Infrastructure and Environment, 2016) one would have hoped this type of event to be less surprising. The event underscores the importance of understanding the variability and uncertainty that comes with estimating extreme floods.

Extreme value analysis often involves estimating the magnitude of events that are greater than the largest from historical (representative) records. This requires establishing a model that described the probability of experiencing such events within a specific period, and subsequently extrapolating this to specific exceedance probabilities. For the Meuse, the traditional approach is fitting a probability distribution to periodic maxima and extrapolate from it (van de Langemheen & Berger, 2001). However, a statistical fit to observations is sensitive to the most extreme events in the time series available. Additionally, the hydrological and hydraulic response to rainfall during extreme events might be different for more frequently occurring events, and therefore be incorrectly described by statistical extrapolation.

GRADE (Generator of Rainfall And Discharge Extremes) is a model-based answer to these shortcomings. It is used to determine design conditions for the rivers Meuse and Rhine in the Netherlands. GRADE is a variant on a conventional regional flood frequency analysis. Instead of using only historical observations, it resamples these into long synthetic time series of rainfall that express the observed spatial and temporal variation. It then uses a hydrological model to calculate tributary flows and a hydraulic model to simulate river discharges (Hegnauer et al., 2014; Leander et al., 2005). Despite the fact that GRADE can create spatially coherent results and can simulate changes in the catchment or climate, it is still based on resampling available measurements or knowledge. Hence, it cannot simulate all types of events that are not present in the historical sample. This is illustrated by the fact that the July 2021 discharge was not exceeded once in the 50,000 years of summer discharges generated by GRADE.

GRADE is an example where underestimation of uncertainty is observed, but certainly not the only model. For example, Bouaziz et al. (2020) and de Boer-Euser et al. (2017) compared different hydrological modeling concepts for the Ourthe catchment (considered in this study as well) and showed the large differences that different models can give when comparing more characteristics than only stream flow. Regardless of the conceptual choices, all models have severe limitations when trying to extrapolate to an event that has not occurred yet. We should be wary to disqualify a model in hindsight after a new extreme has occurred. Alternatively, data-based approaches try to solve the shortcomings of a short record by extending the historical records with sources that can inform on past discharges. For example, paleoflood hydrology uses geomorphological marks in the landscape to estimate historical water levels (Benito & Thorndycraft, 2005). Another approach is to utilize qualitative historical written or depicted evidence to estimate past floods (Brázdil et al., 2012). The reliability of historical records can be improved as well, for example by combining this with climatological information derived from more consistent sea level pressure data De Niel et al. (2017).

In this context, structured expert judgment (SEJ) is another data-based approach. Expert Judgment (EJ) is a broad term for gathering data from judgments based on expertise in a knowledge area or discipline. It is indispensable in every scientific application as a way of assessing the truth or value of new information. *Structured* expert judgment formalizes EJ by eliciting expert judgments in such a way that judgments can be treated as scientific data. One structured method for this is the Classical Model, also known as *Cooke's method* (Cooke & Goossens, 2008). The Classical Model assigns a weight to each expert within a group (usually 5 to 10 experts) based on their performance in estimating the uncertainty in a number of seed questions. These weights are then applied to the experts' uncertainty estimates for the variables of interest, with the underlying assumption that the performance for the seed questions is representative for the performance in the questions of interest. Cooke and Goossens (2008) show an overview of the different fields in which the Classical Model for structured expert judgment is applied. In total, data from 45 expert panels (involving in total 521 experts, 3688 variables, and 67,001 elicitations) are discussed, in applications ranging from nuclear, chemical and gas industry, water related, aerospace sector, occupational sector, health, banking, and volcanoes. Marti et al. (2021) used the same database of expert judgments and observed that using performance-based weighting gives more accurate DMs than assigning weights at random. Regarding geophysical applications, expert elicitation has recently been applied in different studies aimed at informing the uncertainty in climate model predictions (e.g., Bamber et al., 2019; Oppenheimer et al., 2016; Sebok et al., 2021). More closely related to this article, Kindermann et al. (2020) reproduced historical water levels using structured expert judgment (SEJ), and Rongen et al. (2022a) applied SEJ to estimate the probabilities of dike failure for the Dutch part of the Rhine River.

While examples of using specifically the Classical Model in hydrology are not abundantly available, there are many examples of expert judgment as prior information to decrease uncertainty and sensitivity. Four examples in which a Bayesian approach, similar to this study, was applied to limit the uncertainty in extreme discharge estimates are given by (Coles & Tawn, 1996; Parent & Bernier, 2003; Renard et al., 2006; Viglione et al., 2013).

The mathematical approach varies between the different studies, but the rationale for using EJ is the same: adding uncertain prior information to the likelihood of available measurements to help achieve more plausible posterior estimates of extremes. In this study, the approach with which these prior estimates are elicited is formalized by applying the Classical Model.

Structured expert judgment is applied to estimate the magnitude of discharge events for the Meuse River up to an annual exceedance probability of, on average, once per 1,000 years. We aim to get uncertainty estimates for these discharges. Their credibility is assessed by comparing them to GRADE, the aforementioned model-based method for deriving the Meuse River's design flood frequency statistics. A statistical model is quantified both with observed annual maxima and seven experts' estimates for the 10-year and 1000-year discharge on the main Meuse tributaries. The 10-year discharges (unknown to experts at the moment of the elicitation) are used to derive a performance-based expert weight that is used to combine the 1000-year discharges. Participants use their own approach to produce uncertainty estimates. To investigate how the method that combines a) data and expert judgments compares to b) the data-only or c) the expert estimates-only approach, we quantify the model based on all three options. The differences show the added value of each component. This indicates the method's performance both when measurements are available and when they are not, for example in data scarce areas.

6.2. Study area and data used

Figure 6.1 shows an overview of the catchment of the Meuse River. The catchments that correspond to the main tributaries are outlined in orange. The three locations for which we are interested in extreme discharge estimates, Borgharen, Roermond, and Gennep, are indicated with the purple circles. We call these 'downstream locations' throughout this study. The river continues further downstream until it flows into the North Sea near Rotterdam. This part of the river becomes increasingly intertwined with the Rhine River and more affected by the downstream sea water level. Consequently, the water levels can be ascribed decreasingly to the discharge from the upstream catchment. For this reason, we do not assess discharges further downstream than Gennep in this study.

The tributary names are alongside the tributaries. The orange circles indicate the locations along the tributaries where the discharges are measured, the names of the corresponding town is shown within parenthesis. Elevation is shown with the grey scale. Elevation data were obtained from EU-DEM (Copernicus Land Monitoring Service, 2017) and used to derive catchment delineation and tributary steepness. These data were provided to the experts together with other hydrological characteristics, like:

- *Catchment overview*: A map with elevation, catchments, tributaries, and gauging locations
- *Land use*: A map with land use from Copernicus Land Monitoring Service (2018)

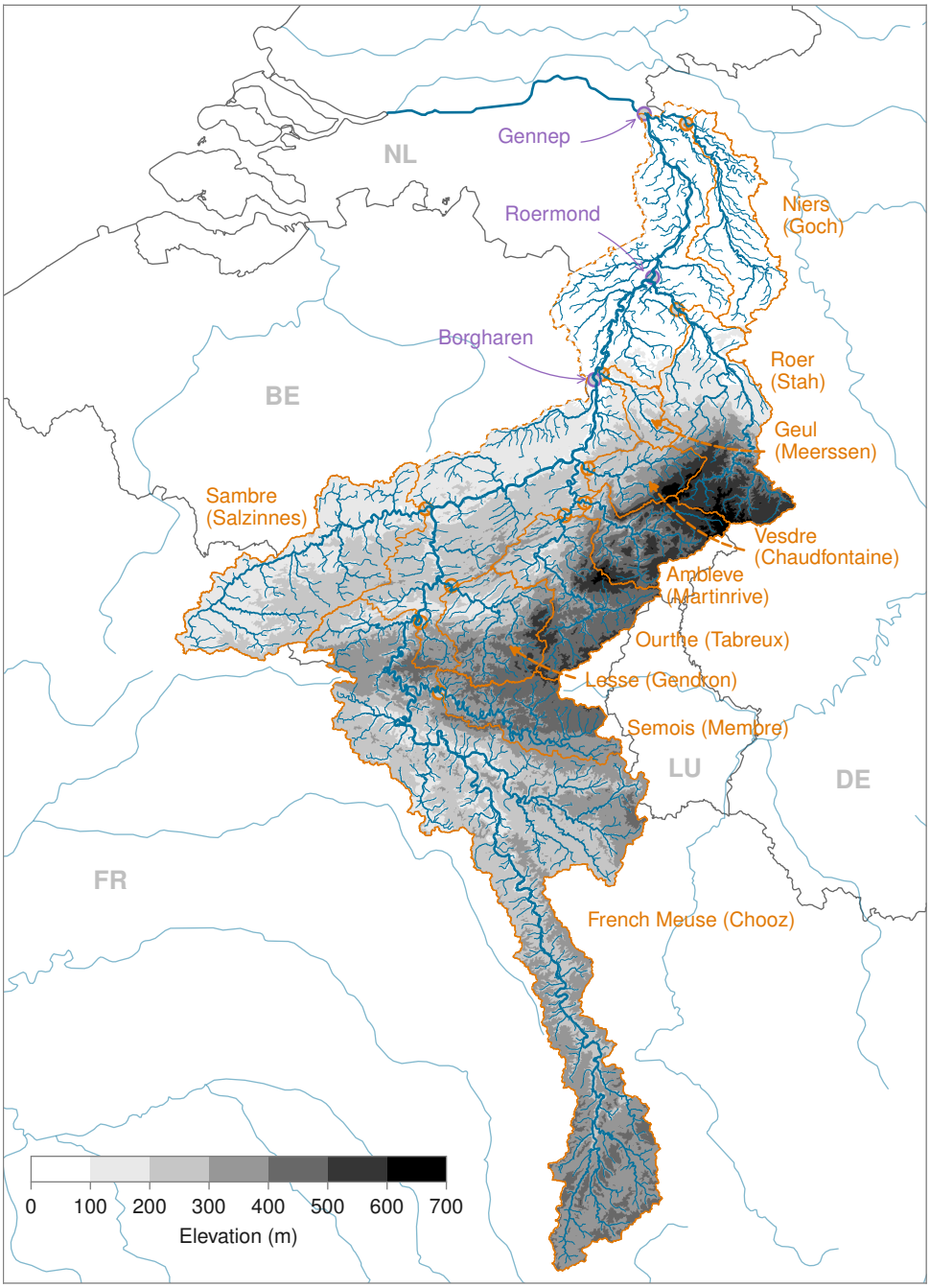


Figure 6.1: Map of the Meuse catchment considered in this study, with main river, tributaries, streams, and catchment boundaries.

- *River profiles and time of concentration*: A figure with longitudinal river profiles and a figure with time between the tributary peaks and the peak at Borgharen for discharges at Borgharen greater than 750 m^3/s.
- *Tabular catchment characteristics*, such as: Area per catchment, as well as the catchment's fraction of the total area upstream of the downstream locations. Soil composition from Food and Agriculture Organization of the United Nations (2003), specifying the fractions of sand, silt, and clay in the topsoil and subsoil. Land use fractions (paved, agriculture, forest & grassland, marshes, water bodies).
- *Statistics of precipitation*: Daily precipitation per month and catchment. Sum of annual precipitation per catchment. Intensity duration frequency curves for the average recurrence intervals (per year): 1, 2, 5, 10, 25, 50, and the maximum. All calculated from gridded E-OBS reanalysis data provided by Copernicus Land Monitoring Service (2020).
- *Hyetographs and hydrographs*: Temporal rainfall patterns and hydrographs for all catchments/tributaries during the 10 largest discharges measured at Borgharen. The sources of discharge data are described below.

This information, included in the supplementary information accompanying (Rongen et al., 2024), was provided to the experts to support them in making their estimates. The discharge data needed to fit the model to the observations were obtained from Service public de Wallonie (2022) for the Belgian gauges, Rijkswaterstaat (2022) and Waterschap Limburg (2021) for the Dutch gauges, and Land NRW (2022) for the German gauge. These discharge data are mostly derived from measured water levels and rating curves. During floods, water level measurements can be incomplete and rating curves inaccurate. Consequently, discharge data during extremes can be unreliable. Measured discharge data were not provided to the experts, except in normalized form as hydrograph shapes.

6.3. Method for estimating extreme discharges with experts

6.3.1. Probabilistic model

To obtain estimates for downstream discharge extremes, experts needed to quantify individual components in a model that gives the downstream discharge as the sum of the tributary discharges, times a factor correcting for covered area and hydrodynamics:

$$Q_d = f_{\Delta t} \cdot \sum_u Q_u, \tag{6.1}$$

where Q_d is the peak discharge of a downstream location during an event, and Q_u the peak discharge of the u'th (upstream) tributary during that event. Location d can be any location along the river where the discharge is assumed to be dependent mainly on rainfall in the upstream catchment. The random variable Q_u is modeled with the generalized extreme value (GEV) distribution (Jenkinson, 1955). We chose this family

of distributions firstly because it is widely used to estimate the probabilities of extreme events. Secondly, it provides flexibility to fit different rainfall-runoff responses by varying between Fréchet (heavy tailed), Gumbel (exponential tail) and Weibull distributions (light tailed). We fitted the GEV distributions to observations, expert estimates, or both, using Bayesian inference (described in Section 6.3.3). The 1,000-year discharge is meant to inform the tail of the tributary discharge probability distributions. This tail is represented by the GEV tail shape parameter that is most difficult to estimate from data. We chose to elicit discharges, rather than a more abstract parameter like the tail shape itself, such that experts make estimates on quantities that may be observed and at "a scale on which the expert has familiarity" (Coles & Tawn, 1996, p. 467).

The factor or ratio $f_{\Delta t}$ in Eq. (6.1) compensates for differences between the sum of upstream discharges and the downstream discharge. These result from, for example, hydraulic properties such as the time difference between discharge peaks and peak attenuation as the flood wave travels through the river (which would individually lead to a factor < 1.0), or rainfall in the Meuse catchment area that is not covered by one of the tributaries (which would individually lead to a factor > 1). When combined, the factor can be lower or higher than 1. The estimated factors are displayed in the last panels of Fig. 6.B.1.

The tributary peak discharges Q_u are correlated because a rainfall event is likely to affect an area larger than a single tributary catchment and nearby catchments have similar hydrological characteristics. This dependence is modeled with a multivariate Gaussian copula that is realized through Bayesian Networks estimated by the experts (A. Hanea et al., 2015). The details of this concern the practical and theoretical aspects of eliciting dependence with experts, which is presented in Chapter 7. The resulting correlation matrices that describe the tributary dependence are displayed in Fig. 7.3.

In summary, using the method of SEJ described in Section 6.3.2, the experts estimate

1. the tributary peak discharges Q_u that are exceeded on average once per 10 years and once per 1,000 years (for brevity called the 10-year and 1,000-year discharge hereafter),
2. the factor $f_{\Delta t}$, and
3. the correlation between tributary peak discharges.

With these, the model in Eq. (6.1) is quantified. The model was deliberately kept simple to ensure that the effect of the experts' estimates on the result remains traceable for them. Section 6.3.4 explains how downstream discharges were generated from these model components (i.e., the different terms in Eq. (6.1)), including uncertainty bounds. The model is also described in more detail in (Rongen et al., 2022b) as well, where it was used in a data-driven context.

6.3.2. Assessing uncertainties with expert judgment

The experts' estimates are elicited using the Classical Model. Refer to Section 2.1 for an elaborate description on this method of structured expert judgment.

The Classical Model combines quantiles estimates from different experts into a single probability density function per elicited item. These estimates are called the decision maker (DM). We used the Metalog distribution to model the probability density connecting the expert estimates (Keelin, 2016). This distribution was presented in Chapter 4 and is capable of exactly fitting any three-percentile estimate. Notice that for this research, the Metalog distribution represents the uncertainty distribution of each expert over a particular discharge with a given return period. While it is related to the underlying distribution of discharge it does not make any assumption about this underlying distribution other than the ones expressed by experts through their percentile estimates. For symmetric estimates, the Metalog is bell-shaped. For asymmetric estimates, it becomes left- or right-skewed. Typically, the Classical Model assumes a uniform distribution in between the percentiles (minimum information). This leads to a stepped PDF where the Metalog gives a smooth PDF. An example of using the Metalog distribution in an expert elicitation study is described by Dion et al. (2020). All calculations related to the Classical Model were performed using the open-source software ANDURYL (Leontaris & Morales-Nápoles, 2018; Rongen et al., 2020; 't Hart et al., 2019).

In this study, the seed questions involve the 10-year discharges for the tributaries of the river Meuse. An example of a seed question is: "What is the discharge that is exceeded on average once per 10 years, for the Vesdre at Chaudfontaine?" The target questions concern the 1000-year discharges, as well as the ratio between the upstream sum and downstream discharge. Discharges with a 10-year recurrence interval are exceptional but can in general be reliably approximated from measured data. Seven experts participated in the in-person elicitation that took place on the 4th of July 2022. The study and model were discussed before the assessments to make sure that the concepts and questions were clear. After this, an exercise for the Weser catchment was done in which the experts answered four questions that were subsequently discussed. In this way, the experts could compare their answers to the realizations and view the resulting scores using the Classical Model.

Apart from the training exercise, the experts answered 26 questions: 10 seed questions regarding the 10-year discharge (one for each tributary), 10 target questions, regarding the 1,000-year discharge, and 6 target questions for the ratios between upstream sum and downstream discharge (10-year and 1,000-year, for three locations). In a second part of the elicitation, the dependence between tributaries was estimated. The method and result for this are explained in Chapter 7. A list of the seven participants' names, their affiliations, and their field of expertise is shown in Table 6.1. While the participants are invited based on their expertise, experts are scored *post hoc* in terms of their ability to estimate uncertainty in the context of the study. We note that the alphabetical order of the experts in the table does not correspond to their labels in the results. An overview of the data provided to the participants is given in Section 6.2, while the data itself, as well as the questionnaire, are presented in the supplementary information accompanying (Rongen et al., 2024).

Table 6.1: List of experts with their affiliation and professional interests.

Name	Affiliation	Field of expertise
Alexander Bakker	Rijkswaterstaat & TU Delft	Risk analysis for storm surge barriers, extreme value analyses, climate change and climate scenarios.
Eric Sprokkereef	Rijkswaterstaat	Coordinator crisis advisory group Rivers. Operational forecaster for Rhine and Meuse
Ferdinand Diermanse	Deltares	Expert advisor and researcher flood risk.
Helena Pavelková	Waterschap Limburg	Hydrologist
Jerom Aerts	Delft University of Technology	Hydrologist, focused on hydrological modeling on a global scale. PhD candidate.
Nicole Jungermann	HKV consultants	Advisor water and climate
Siebolt Folkertsma	Rijkswaterstaat	Advisor in the Team Expertise for the River Meuse

6.3.3. Determining model coefficients with Bayesian inference

The model for downstream discharges (Eq. (6.1)) consists of generalized extreme value (GEV) distributions per tributary. The GEV-distribution has three parameters, the location (μ), scale (σ), and shape parameter (ξ). Consider $z = (x-\mu)/\sigma$. The probability density function (PDF) of the GEV is then,

$$f(x) = \begin{cases} \frac{1}{\sigma}\exp(-\exp(-z))\exp(-z), & \text{if } \xi = 0 \\ \frac{1}{\sigma}\exp(-(1-\xi z)^{1/\xi})(1-\xi z)^{1/\xi-1}, & \text{if } z \leq 1/\xi \text{ and } \xi > 0 \end{cases} \tag{6.2}$$

For each tributary, a (joint) distribution of the model parameters was determined using Bayesian inference, based on expert estimates and observed tributary discharge peaks during annual maxima at Borgharen. Bayesian methods explicitly incorporate uncertainty, a key aspect of this study, and provide a natural way to integrate expert judgment with observed data.

Bayes theorem gives the posterior distribution $p(\theta|\mathbf{q})$ of the (hypothesized) combination of GEV-parameters, θ, given the observed peaks $\mathbf{q}$, as a function of the likelihood $p(\mathbf{q}|\theta)$ and the prior distribution $\pi(\theta)$:

$$p(\theta|\mathbf{q}) = \frac{p(\mathbf{q}|\theta)\pi(\theta)}{p(\mathbf{q})}. \tag{6.3}$$

The likelihood can be calculated using Eq. (6.2) from the product of the probability density of all (independent) annual maxima: $p(\mathbf{q}|\theta) = \prod_i \left(f(q_i|\theta)\right)$. The calculation of the prior is discussed below. That leaves $p(\mathbf{q})$, which is not straightforward to calculate. However, the posterior distribution can still be estimated using the Bayesian sampling technique Markov-Chain Monte Carlo (MCMC). MCMC algorithms compare different propositions of the numerator in Eq. (6.3), leaving the denominator as a normalization

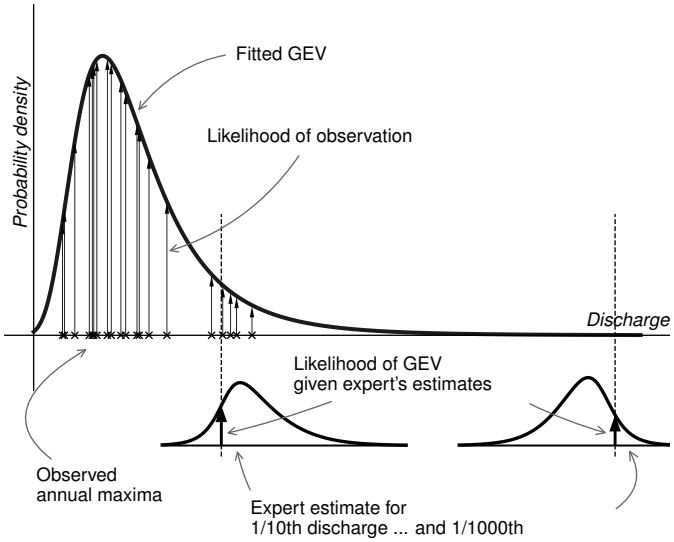


Figure 6.2: Conceptual visualization of elements in the likelihood-function of a tributary GEV-distribution.

factor that crosses out. In this study, we used the affine invariant MCMC ensemble sampler as described by Goodman and Weare (2010), available through the Python module 'emcee' (Foreman-Mackey et al., 2013). This sampler generates a trace of distribution parameters that forms the empirical joint probability distribution of, in our case, the three GEV parameters for each tributary. These are subsequently used to calculate the downstream discharges (see Section 6.3.4). For more background on Bayesian inference and MCMC, refer to Section 2.3.3.

The prior consists of two parts, the expert estimates for the 10-year and 1,000-year discharge, and a prior for the GEV tail shape parameter ξ. Since the experts do not know the values of the discharges they are estimating, their estimates can be considered prior information. The prior probability $\pi(\theta)$ of the expert's estimates is calculated in a similar way as described by Viglione et al. (2013): Given a GEV-distribution $f(Q|\theta)$, the discharge q for a specific annual exceedance probability p is calculated from the quantile function or inverse CDF (F^{-1}),

$$q_{p_j} = F^{-1}(1 - p_j|\theta), \tag{6.4}$$

with p_j being the j'th elicited exceedance probability. This discharge is compared to the expert's or DM's estimate for this 10- or 1,000-year discharge, $g(q_{p_j})$. Figure 6.2 illustrates this procedure. The top curve $f(Q|\theta)$ represents a proposed GEV-distribution for the random variable Q (tributary peak discharge) with parameter vector θ. This GEV gives discharges corresponding to the 0.9 and 0.999th quantile (i.e., the 10-year and 1,000-year discharge). These discharges can then be compared to the expert estimates, illustrated by the two bottom graphs. Additionally, the figure shows the likelihood of observations with the vertical arrows ($p(\mathbf{q}|\theta)$ in Eq. (6.3)).

Apart from the expert estimates, we prefer a weakly informative prior for θ (i.e., uninfor-

mative, but within bounds that ensure a stable simulation), such that only the data and expert estimates inform the final result. However, an informative prior was added to the shape parameter ξ because with only expert estimates and no data, two discharge estimates are not sufficient for fitting the three parameters of the GEV-distribution. Additionally, the variance in the shape-parameter decreases with increasing number of years (or other block maxima) in a time series (Papalexiou & Koutsoyiannis, 2013). The 30 to 70 annual maxima per tributary in this study are not sufficient to reach convergence. Similar observations have been presented before for extreme precipitation in (Koutsoyiannis, 2004a, 2004b) Therefore, we employ the geophysical prior as presented by Martins and Stedinger (2000); a beta distribution with hyper parameters $\alpha = 6$ and $\beta = 9$ for $x \in [-0.5, 0.5]$, for which the PDF is:

$$h(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{6.5}$$

with $x = \xi + 0.5$, and Γ being the gamma-function. This PDF is slightly skewed towards negative values of the shape parameter, preferring the heavy tailed Fréchet distribution over the light tailed reversed Weibull. In their analysis of a very large number of rainfall records worldwide, Papalexiou and Koutsoyiannis (2013) came to a similar distribution for the GEV-shape parameter. For μ and σ, we assigned equal probability to all values greater than 0. This corresponds to a weakly informative prior for μ (positive discharges), and an uninformative prior for σ (only positive values are mathematically feasible).

With both expert estimates g and the constrained tail shape, the prior distribution becomes

$$\pi(\theta) = \prod_j \left(g_j\left(F_\theta^{-1}(1-p_j)\right)\right) \cdot h(\xi + 0.5) \tag{6.6}$$

for $-0.5 < \xi < 0.5$, $\sigma > 0$, and $\mu > 0$. $\pi(\theta) = 0$ for any other combination. This gives all the components to calculate the posterior distribution in Eq. (6.3) using MCMC.

The posterior distribution comprises the prior tail-shape distribution, the prior expert estimates of the 10-year and 1,000-year discharges, and the likelihood of the observations. As described in Section 6.1 we compare the performance of using data, EJ, and the combination of both. If only data are used, the expert estimates drop out. If only expert judgments are used, the likelihood drops out and both expert estimates are used. If both data and expert judgment are used, only the 1,000-year expert estimate is used.

With the just described procedure, the (posterior) distributions for the tributary discharges (Q_u in Eq. (6.1)) are quantified. This leaves the ratio between the upstream sum and downstream discharge ($f_{\Delta t}$) and the correlations between the tributary discharges to be estimated. For the ratios, we distinguished between observations and expert estimates as well. A log-normal distribution was fitted to the observations. This corresponds to a practical choice for a distribution of positive values with sufficient shape flexibility. The ratio itself does not represent streamflow, so there is no need to assume a heavy tailed distribution as would be expected for streamflow (Dimitriadis et al., 2021). The experts estimated a distribution for the factor as well, which was used directly for the experts-only fit. For the combined model fit, the observation-fitted log-normal distribution was used up to the 10-year range, and the expert estimate (fitted with a Metalog

distribution) for the 1,000-year factor. Values of $f_{\Delta t}$ for return periods T greater than 10 were interpolated (up to 1000-years) or extrapolated,

$$f_{\Delta t}|_T = f_{\Delta t}|_{10y} + \frac{\log(T) - \log(10)}{\log(1,000) - \log(10)} \cdot (f_{\Delta t}|_{1,000y} - f_{\Delta t}|_{10y}), \tag{6.7}$$

with $f_{\Delta t}|_{10y}$ being sampled from the log-normal and $f_{\Delta t}|_{1000y}$ from the expert estimated Metalog distribution. During the expert session, one participant requested to make different estimates for the factor at the 10-year event and 1,000-year event, a distinction that initially was not planned. Following this request, we changed the questionnaire such that a factor could be specified at both return periods. One expert used the option to make two different estimates for the factors.

Regarding the correlation matrix that describes the dependence between tributary extremes, the observed correlations were used for the data-only option and the expert-estimated correlations for the expert-only option. For the combined option, we took the average of the observed correlation matrix and the expert-estimated correlation matrix. Other possibilities for combining correlation matrices are available (see for example Al-Awadhi & Garthwaite, 1998, for a Bayesian approach), however in-depth research of these options is beyond the scope of this study.

6.3.4. Calculating the downstream discharges

The three components from Eq. (6.1) needed to calculate the downstream discharges are:

- Tributary (marginal) discharges, represented by GEV-distributions derived through Bayesian inference.
- Dependence between tributaries, represented by a multivariate normal copula (for more background on this, see Chapter 7).
- The ratio between the upstream sum and downstream discharges ($f_{\Delta t}$).

In line with the objective of this article, an uncertainty estimate is derived for the downstream discharges. This section describes the method in a conceptual way. Section 6.A contains a formal step-by-step description.

To calculate a single exceedance frequency curve for a downstream location, 10,000 events (annual discharge maxima) are drawn from the 9 tributaries' GEV-distributions. Note that 10 tributaries are displayed in Fig. 6.1. The Semois catchment is however part of the French Meuse catchment and therefore only used to assess expert performance. The 9 tributary peak discharges are summed per event and multiplied with 10,000 factors (one per event) for the ratio between upstream sum and downstream discharge. The 10,000 resulting downstream discharges are assigned an annual exceedance probability through empirical plot positions, resulting in an exceedance frequency curve. This process is repeated 10,000 times with different GEV-realizations from the MCMC-trace, resulting in 10,000 curves (each based on 10,000 discharges) from which the uncertainty bandwidth is determined. This is illustrated in Fig. 6.3. The grey lines depict 50 of

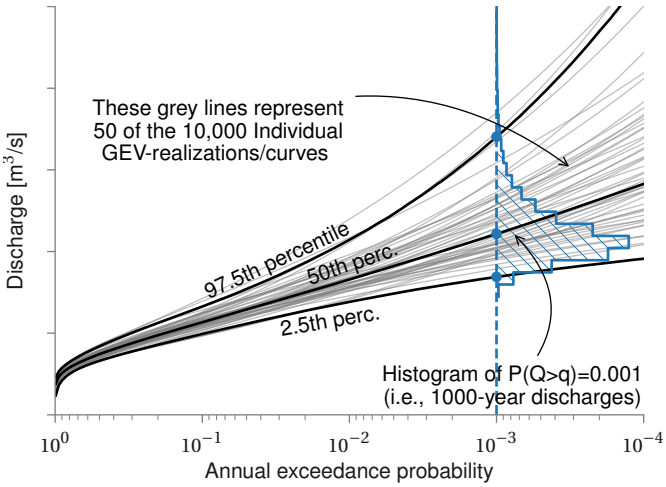


Figure 6.3: Individual exceedance frequency curves for each GEV-realization or downstream discharge, and the different percentiles derived from these.

the 10,000 curves (these can be both tributary GEV-curves, or downstream discharge curves). The horizontal histogram gives the distribution of the 1000-year discharges. The dots indicate the 2.5th, 50th, and 97.5th percentiles in this histogram. Calculating these percentiles for all annual exceedance probabilities results in the black percentile curves, creating the uncertainty interval.

The dependence between tributaries is incorporated in two ways. First, the 10.000 events underlying each downstream discharge curve are correlated. This is achieved by drawing the [9 × 10,000] sample from the (multivariate normal) correlation model, transforming these samples to uniform space (with the normal CDF), and then to each tributary's GEV-distribution space (with the GEV's quantile function). This is the usual approach when working with a multivariate normal copula. The second way of incorporating the tributary dependence is by choosing GEV-combinations from the MCMC-results while considering the dependence between tributaries (i.e., picking high or low curves from the uncertainty bandwidth for multiple tributaries). As illustrated in Fig. 6.3, a tributary's GEV-distribution can lead to relatively low or high discharges. This uncertainty is largely caused by a lack of realizations in the tail (i.e., not having thousands of years of independent and identically distributed discharges). If one tributary would fit a GEV distribution resulting in a curve on the upper end of the bandwidth, it is likely because it experienced a high discharge event that affected its neighboring tributary as well. Consequently, the neighboring tributary is more likely to also have a 'high-discharge' GEV-combination. To account for this, we first sort the GEV-combinations based on their 1,000-year discharge (i.e., the curves' intersections with the dashed line), and draw a 9-sized sample from the dependence model. Transforming this to uniform space gives a value between 0 and 1 that is used as rank to select a (correlated) GEV-combination for each tributary. Doing this increases the likeliness that different tributaries will have relatively high or low sampled discharges.

6.4. Experts' performance and resulting discharge statistics

This result section first presents the experts' scores for the Classical Model (Section 6.4.1) and the experts' rationale for answering the questions (Section 6.4.2). After this, the extreme value results for the tributaries (Section 6.4.3) and downstream locations (Section 6.4.4) are presented.

6.4.1. Results for the Classical Model

The experts estimated three-percentiles (5th, 50th and 95th) for the 10- and 1,000-year discharge for all larger tributaries in the Meuse catchment. An overview of the answers is displayed in Fig. 6.B.1. Based on these estimates, the scores for the Classical Model are calculated as described in Section 6.3.2. The resulting statistical accuracy, information score, and combined score (which, after normalizing, become weights) are shown in Table 6.2.

Table 6.2: Scores for the Classical Model, for the experts (top 7 rows) and decision makers (bottom 3 rows).

	Statistical accuracy	Information score		Comb. score
		All	Seed	
Exp A	0.000799	1.605	1.533	0.00123
Exp B	0.000456	1.576	1.633	0.000745
Exp C	$2.3 \cdot 10^{-8}$	1.900	1.868	$4.4 \cdot 10^{-8}$
Exp D	0.683	0.711	0.626	0.427
Exp E	0.192	1.395	1.263	0.242
Exp F	0.000456	1.419	1.300	0.000593
Exp G	0.00629	1.302	1.232	0.00775
GL (opt)	0.683	0.659	0.670	0.458
GL	0.683	0.648	0.661	0.452
EQ	0.493	0.537	0.551	0.271

The statistical accuracy varies between $2.3 \cdot 10^{-8}$ for expert C to 0.683 for expert D. Two experts have a score above a significance level of 0.05. The information scores show, as usual, less variation. The expert with the highest statistical accuracy (expert D) also has the lowest information score. Expert E, who has a high statistical accuracy as well, estimated more concentrated percentiles, resulting in a higher information score.

Figure 6.4 zooms in on the statistical accuracy, by showing the position of each realization (outcome) within the experts' three-percentile estimate for each of the 10-year discharges. A high statistical accuracy means realizations to these seed variables are distributed accordingly to (or as close to) the mass in each inter-quantile bin: one realization below the 5th percentile, 4 in between the 5th and the median, four between the median and the 95th and one above the 95th. Expert D's estimates closely resemble this distribution ($\frac{1}{10}, \frac{5}{10}, \frac{4}{10}, \frac{0}{10}$ for each inter-quantile respectively), hence the high statistical

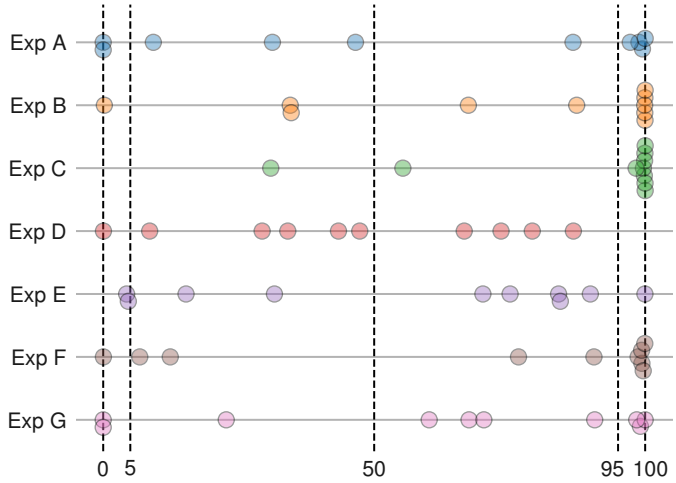


Figure 6.4: Seed question realizations compared to each expert's estimates. The position of each realization is displayed as percentile point in the expert's distribution estimate.

accuracy score. A concentration of dots on both ends indicates overconfidence (too close together estimates, resulting in realizations outside of the 90% bounds). We observe that most experts tend to underestimate the measured discharges, since most realizations are higher than their estimated 95th percentile. Note that the highest score is not received for the (median) estimates closest to the realization but to evenly distributed quantiles, as the goal is estimating uncertainty rather than estimating the observation (see Section 6.3.2).

The variation between the three decision makers (DMs) in Table 6.2 is limited. Optimizing the DM (i.e., excluding experts based on statistical accuracy to improve the DM-score) has a limited effect. In this case, only expert D and E would have a non-zero weight, resulting in more or less the same results compared to including all experts, even when some of them contribute with 'marginal' weights. The equal weights DM in this case results in an outcome that is comparable to that of the performance-based DM, i.e., a high statistical accuracy with a slightly lower information score compared to the other two DMs.

We present the model results as discussed earlier through three cases a) only data, b) only expert estimates, and c) the two combined as described in Section 6.3.3. We used the global weights DM for the data and experts option (c). This means the experts' estimates for the 10-year discharges were used to assess the value of the 1,000-year answer. For the experts-only option, we used the equal weights DM, because using the global weights emphasizes estimates matching the measured data in the 10-year range. This would indirectly lead to including the measured data in the fit. By using equal weights, we ignore the relevant seed questions and the corresponding differential weights.

6.4.2. Rationale for estimating tributary discharges

We requested the experts to briefly describe the procedure they followed for making their estimates. Overall, three approaches were distinguished. The first was using a simple conceptual hydrological model, in which the discharge follows from catchment characteristics like (a subset of) area, rainfall, evaporation and transpiration, rainfall-runoff response, land-use, subsoil, slope, or the presence of reservoirs. Most of this information was provided to the experts, and if not, they made estimates for it themselves. A second approach was to compare the catchments to other catchments known by the expert, and possibly adjusting the outcomes based on specific differences. A third approach was using rules of thumb, such as the expected discharge per square kilometer of catchment or a 'known' factor between an upstream tributary discharge and a downstream discharge (of which the statistics are better known). For estimating the 1,000-year discharge, the experts had to do some kind of extrapolation. Some experts scaled with a fixed factor, while others tried to extrapolate the rainfall, for which empirical statistics where provided. The hydrological data (described in Section 6.2) was provided to the experts in spreadsheets as well, making it easier for them to do computations. However, the single day that was available for the full elicitation limited the possibilities for making detailed model simulations.

Figure 6.5 shows how the different approaches led to different answers per tributary. It compares the 50th percentile of the discharge estimates per tributary of each expert, by dividing them through the catchment area.The 10-year and 1,000-year discharges from

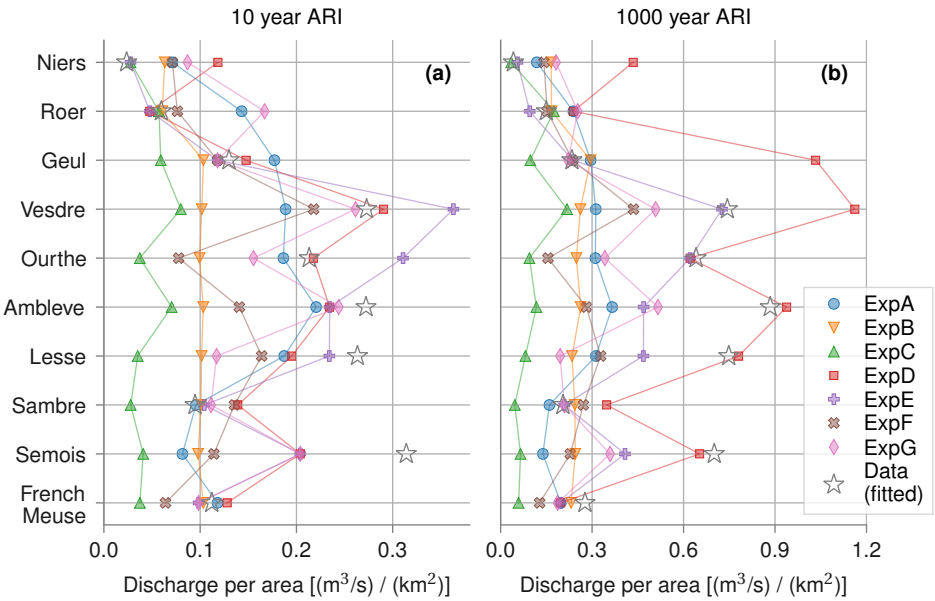


Figure 6.5: Discharge per area for each tributary and experts, based on the estimate for the 50th percentile. **(a)** for the 10-year, and **(b)** for the 1,000-year discharge. The lines are displayed to help distinguish overlapping markers.

fitting the observations (i.e., the data only approach) are indicated with the stars. Note that the discharge per area varies between tributaries. This is the result of catchment properties, such as land use (e.g., a forest can retain more water than an urban area) or topographical properties (e.g., catchment steepness, size, or shape). Figure 6.5 shows that most experts estimated higher discharges for the steeper tributaries (Ambleve, Vesdre, Lesse). The experts estimated the median 1,000-year discharges to be 1.7 to 3.8 times as high as the median 10-year discharge, with an average of on average 2.3 for all experts and tributaries. The statistically most accurate expert, Expert D, estimated factors in between 1.6 and 7.0. Contrarily, expert E, with the second highest score, estimated a ratio of 2.0 for all tributaries. For estimating the factor between the tributaries' sum and the downstream discharge ($f_{\Delta t}$ in Eq. (6.1)), experts mainly took into consideration that not 100% of the area is covered by the tributary catchments for which the discharge-estimates were made, and that the tributary hydrograph peaks have different lag times. Additional aspects noted by the experts were the effects of flood peak attenuation and spatial dependence between tributaries and rainfall.

6.4.3. Extreme discharges for tributaries

We calculated the extreme discharge statistics for each of the tributaries based on the procedures described in Section 6.3.3. Figure 6.6 shows the results for Chooz and Chaudfontaine (left and middle column). Chooz is a larger not too steep tributary, while Chaudfontaine is a smaller steep tributary (see Fig. 6.1). The right column shows the discharges for Borgharen, the location of interest, estimated through Eq. (6.1), which is further discussed in Section 6.4.4. The results for the other tributaries for all experts and DMs are shown in the supplementary information accompanying (Rongen et al., 2024).

The top row (a, d, g) in Fig. 6.6 shows the uncertainty interval of these distributions when fitted only to the discharge measurements. The outer colored area is the 95% interval, the opaquer inner area the 50% interval, and the thick line the median value. The second row (b, e, h) shows the fitted distributions when only expert estimates are used. The bottom row (c, f, i) shows the combination of expert estimates and data. The data-only option closely matches the data in the range of return periods that can be estimated with reasonable accuracy from data, but the uncertainty interval grows for larger return periods. Contrarily, the experts-only option shows much more variation in the frequency range corresponding to return periods of observed events, while the return periods that exceed the record lengths are more constrained. The combined option is accurate in the observed range, while the constraining influence of the DM estimates is visible in the extrapolated range as well.

6.4.4. Extreme discharges for Borgharen

Combining all the marginal (tributary) statistics with the factor for downstream discharges and the correlation models estimated by the experts, we get the discharge statistics for Borgharen. The results for this are shown in Fig. 6.6 (g, h, i). As with the statistics of the tributaries, we observe high accuracy for the data-only estimates in the 'in sample' range, constrained uncertainty bounds for EJ-only in the range with higher return

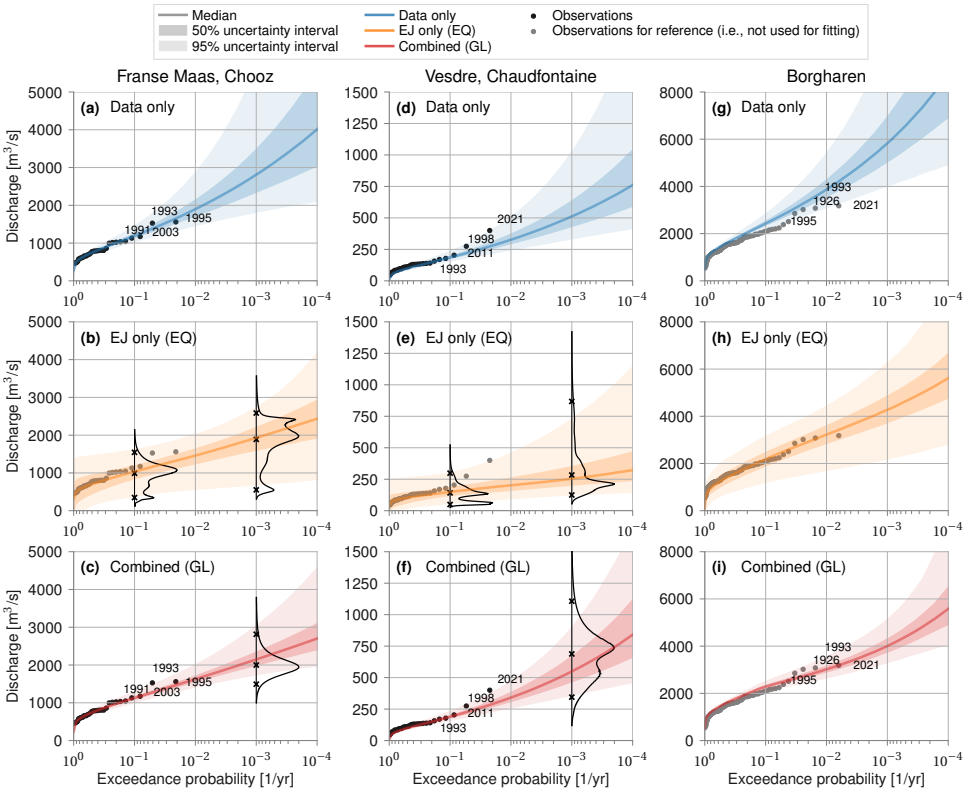


Figure 6.6: Extreme discharge statistics for Chooz **(a, b, c)**, Chaudfontaine **(d, e, f)** and Borgharen **(g, h, i)**. **(a, d, g)** represent data only, **(b, e, h)** expert judgment only, and **(c, f, i)** the data and expert judgment combined.

periods, and both when combined. The combined results match the historical observations well. Note that this is not self-evident as the distributions were not fitted directly to the observed discharges at Borgharen but rather obtained through the dependence model for individual catchments and Eq. (6.1). Contrarily, the data-only results deviate from the observations in the 10- to 100-year range. Sampling from the fitted model components (GEVs, dependence model, and factors) does not accurately reproduce the downstream discharges in this range because they were individually fitted and not as a whole. We do not consider this a problem, as the study is oriented towards showing the effects of expert quantification in combination with more traditional hydrological modeling. The EJ-only estimates give a much wider uncertainty estimate. The experts' combined median matches the observations surprisingly well, but the large uncertainty within the observed range cautions against drawing general conclusions on this.

Zooming in on the discharge statistics for the downstream location Borgharen, we consider the 10, 100, and 1,000-year discharge. Figure 6.7 shows the (conditional) probability distributions (smoothed with a kernel density estimate) for these discharges at the location of interest.

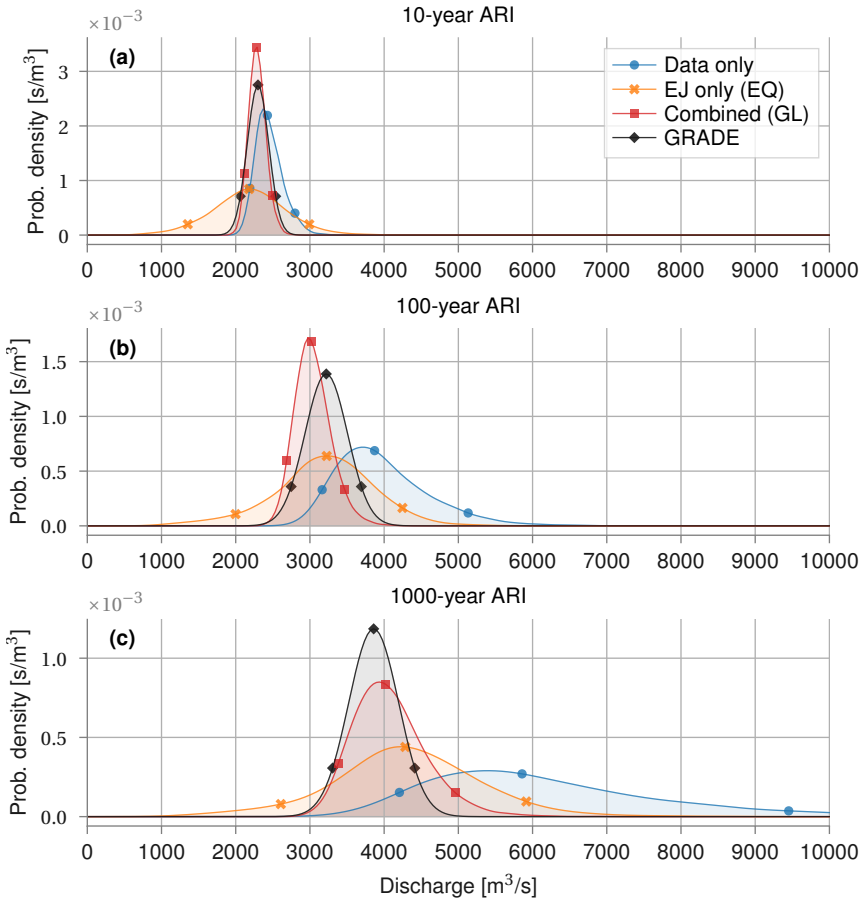


Figure 6.7: Kernel density estimates for the 10-year **(a)**, 100-year **(b)**, and 1,000-year **(c)** discharge for Borgharen. The dots indicate the 5th, 50th and 95th percentile.

Comparing the three modeling options discussed thus far, we see that the data-only option is very uncertain, with a 95% uncertainty interval of 4,000 to around 9,000 m^3/s for the 1,000-year discharge. A Meuse-discharge of 4,000 m^3/s will likely flood large stretches along the Meuse in the Dutch province Limburg, while a discharge of 5,000 m^3/s also floods large areas further downstream (Rongen, 2016). For discharges higher than 6,000 m^3/s the applied model (Eq. (6.1)) should be reconsidered, as the hydrodynamic properties of the system change due to upstream flooding.

The combined results are remarkably close to the currently used GRADE-statistics for dike assessment; the uncertainty is slightly larger, but the median is very similar. The EJ-only results are less precise, but the median values are similar to the combined results and GRADE-statistics. The large uncertainty is mainly the results of equally weighting all experts instead of assigning most weight to experts D and E (as done for the global weight DM). For the combined data and EJ approach, the results for the tributary discharges

roughly cover the intersection of the EJ-only and data-only results (see Fig. 6.6 a-f). Figure 6.7 does not show this pattern, with the EJ-only results positioned in between the data-only and combined results. This is mainly due to equal weight DM used for the EJ-only results, which gives a higher factor between upstream and downstream discharges ($f_{\Delta t}$ in Eq. (6.1)), and therefore higher resulting downstream discharges. Overall, the combined effect of data and EJ is more difficult to identify in the downstream discharges (Fig. 6.6 g-i) than it is in the tributary discharge GEVs (Fig. 6.6 a-f). This is due to the additional model components (i.e., the factor between upstream and downstream, and the correlation model) affecting the results. Additional plots similar to Fig. 6.6 that illustrate this are presented in the supplementary information accompanying (Rongen et al., 2024). There, the results for the other two downstream locations, Roermond and Gennep, are presented as well. These results behave similar to those for Borgharen and are therefore not presented here.

6.5. Discussion

This study proposed a method to estimate credible discharge extremes for the Meuse River (1,000-year discharges in the case of this research). Observed discharges were combined with expert estimates through the GEV-distribution, using Bayesian inference. The GEV-distribution has typically less predictive power in the extrapolated range. Including expert estimates, weighted by their ability to estimate the 10-year discharges, improved the precision in this range of extremes.

Several model choices were made to obtain these results. Their implications warrant further discussion and substantiation. This section addresses the choice for the elicited variables, the predictive power of 10-year discharge estimates for 1000-year discharges, the overall credibility of the results, and finally, some comments on model choices and uncertainty.

6.5.1. Method and model choices

We chose to elicit tributary discharges, rather than the downstream discharges (our ultimate variable of interest) themselves. We believe that experts' estimates for tributary discharges correspond better to catchment hydrology (rainfall-runoff response). Additionally, this choice enables us to validate the final result with the downstream discharges. With the chosen set-up we thus test the experts' capabilities for estimating system discharge extremes from tributary components, while still considering the catchment hydrology, rather than just informing us with their estimates for the end results. However, this does not guarantee that the downstream discharges calculated from the experts' answers match the discharges they would have given if elicited directly.

We fitted the GEV-distribution based on the elicited 10-year and 1000-year discharges. In particular the GEV's uncertain tail shape parameter is informed through this, as the location and scale parameter can be estimated from data with relative certainty. Alternatively, we could have estimated the tail shape parameter directly or estimated a related parameter such as the ratio or difference between discharges. The latter was done by

Renard et al. (2006) who elicited the 10-year discharge and the *differences* between the 10- and 100-year and 100- and 1,000-year discharges. This approach reduces the dependence between expert estimates for different quantiles, and therefore between the priors (when more than one quantile is used) (Coles & Tawn, 1996). Additionally, it shifts the experts' focus to assessing how surprising or extreme rare events can be. Because we were ultimately interested in the 1000-year discharges, we chose eliciting this discharge directly. This will give a more accurate representation of this specific value than composing it of two random variables with a dependence that is unknown to us. We appreciate however that if experts would have estimates ratios or differences, and been evaluated by this, different weights would have resulted than the ones presented in this research (refer to the markedly different ratios between the 10-year and 1,000-year discharge for the two best experts D and E in Fig. 6.5). A study focusing on how surprising large events can be, and whether one method renders consistently larger estimates than the other, would make an interesting comparison. Finally, we note that Renard et al. (2006) combine different extreme value distributions with non-stationary parameters in a single Bayesian analysis, which makes their method a good example of incorporate climate change effects (often considered a driver of for new extremes) in the method as well. This was however out of the scope of our research, which shows that extreme discharge statistics can be improved when combining them with structured expert judgment procedures.

Regarding the goodness-of-fit of the chosen GEV distribution, we note that some of the experts estimated 1,000-year discharges much higher of lower than would be expected from observations. This might indicate that the GEV-distribution is not the right model to observations and expert estimates. However, a significantly lower estimate indicates that the estimated discharge is wrong, as it is unlikely that the 1,000-year discharge is lower than the highest on record. A significantly higher estimate, on the other hand, might be valid, due to a belief in a change in catchment response under extreme rainfall (e.g., due to a failing dam). This would violate the GEV-distribution's 'identically distributed' assumption. However, the GEV has sufficient shape flexibility to facilitate substantially higher 1,000-year discharges, so we do not consider this a realistic shortcoming. Accordingly, rather than viewing the GEV as a limiting factor for fitting the data, we use it as a validation for the Classical Model scores, as described in Section 6.5.2.

Finally, we note the model's omission of seasonality. The July 2021 event was mainly extraordinary because of its magnitude *in combination with* the fact that it happened during summer. Including seasonality would have been a valuable addition to the model but it would also have (at least) doubled the number of estimates provided by each expert, which was not feasible for this study. The exclusion of seasonality from our research does not alter our main conclusion, which is the possibility of enhancing estimation of extreme discharges through structured expert judgments.

6.5.2. Validity of the results

The experts participating in this study were asked to estimate 10-year and 1000-year discharges. While both discharges are unknown to the expert, the underlying processes leading to the different return period estimates can be different. An implicit assumption is that the experts' ability to estimate the seed variables (a 10-year discharge) reflects

their ability to estimate the target variables (a 1000-year discharge). This assumption is in fact one of the most crucial assumptions in the Classical Model. The objective of this research is not to investigate this assumption. For an example of a recent discussion on the effect of seed variables on the performance of the Classical Model the reader is referred to (Eggstaff et al., 2014). The representativeness of the seed variables for calibration variables has extensively been discussed in, for example, (Cooke, 1991). Seed questions have to be as close as possible to the variables of interest, and mostly concern similar questions from different cases or studies. Precise 1000-year discharge estimates are however unknown for any river system, making this option infeasible for this study. In comparison, with a conventional model-based approach, the ability of a model to predict extremes is also estimated from (and tailored to) the ability to estimate historical observations (through calibration). Advantages of relying in the extrapolation of a group of experts are that they can explicitly consider uncertainty and are assessed on their ability to do so through the Classical Model. In Section 6.5.1 we described how inconsistencies between the observations and expert estimates can lead to a sub-optimal GEV-fit. The fact that this is most prevalent in the low-scoring experts and least for experts D and E supports the credibility of the results. Moreover, this means that the 'bad' fits have little weight in the final global weight DM results, and secondly that the GEV is considered a suitable statistical distribution to fit observations and expert estimates.

The GRADE results from (Hegnauer & Van den Boogaard, 2016) were used to validate the 1,000-year downstream discharge results. These GRADE-statistics at Borgharen (currently used for dike assessment) give a lower and less uncertain range for the 1,000-year discharge than the estimates obtained through our methodology. The estimates obtained in this study present larger uncertainty bands and indicate higher extreme discharges. This might be a consequence of the fact that we did not show the measured tributary discharges to the experts, such that we could clearly distinguish the effect of observations and 'prior' expert judgments. Moreover, GRADE (at the time) did not include the July 2021 event. If the GRADE statistics had been derived with the inclusion of the July 2021 event, it would likely assign more probability to higher discharges. The experts' estimates on the contrary were elicited after the July 2021 event which likely did affect their estimates. Therefore, the comparison between GRADE and the expert estimates should not be used to assess correctness, but as an indication of whether the results are in the right range. Finally, note that the full GRADE-method is not published in a peer-reviewed journal (the weather generator is, (Leander et al., 2005)). However, because the results are widely used in the Dutch practice of flood risk assessment (and known to the experts as well) we considered them the best source for comparing the results in the present study.

To evaluate the value of the applied approach that uses data combined with expert estimates, we compared the results that were fitted to only data or only expert judgment to the results of the combination. For the last option we used an equal weight decision maker, a conservative choice as the experts' statistical accuracy could potentially still be determined based on a different river where data for seed questions are available. While the marginal distributions of the EJ-only case present wide bandwidths (see Fig. 6.6 b and e), the final results for Borgharen still gave a statistically accurate result but with a

few caveats, namely that the uncertainty is very large and that the 10-year and 1,000-year estimates in itself are insufficient to inform the GEV without adding prior information (otherwise we have 2 estimates for 3 parameters). Consequently, when only using expert estimates, eliciting the random variable (discharges) directly through a number of quantiles of interest, might be a suitable alternative.

6.5.3. Final remarks on model choices

Finally, we note that using expert judgment to estimate discharges through a model (like we did) still gives the analyst a large influence in the results. We try to keep the model transparent and provide the experts with unbiased information, but by defining the model on beforehand and providing specific information we steer the participants towards a specific way of reasoning. Every step in the method, such as the choice for a GEV-distribution, the dependence model, or the choice for the Classical Model, affects the end result. By presenting the method and providing background information explicitly, we hope to have made this transparent and show the usefulness of the method for similar applications.

6.6. Conclusions

This study sets out to establish a method for estimation of statistical extremes through structured expert judgment and Bayesian inference, in a case-study for extreme river discharges on the Meuse River. Experts' estimates of tributary discharges that are exceeded in a once per 10 year and once per 1,000-year event are combined with high river discharges measured over the past 30-70 years. We combine the discharges from different tributaries with a multivariate correlation model describing their dependence and compare the results for three approaches, a) data only, b) expert judgment only, and c) the combination. The expert elicitation is formalized with the Classical Model for structured expert judgment.

The results of applying our method show credible extreme river discharges resulting from the combined expert-and-data approach. A comparison to GRADE, the prevailing method for estimating discharge extremes on the Meuse, gives similar ranges for the 10-, 100-, 1,000-year discharges as GRADE. Moreover, the two experts with the highest scores from the Classical Model had discharge estimates that correspond well with those discharges that might be expected from the observations. This indicates that using the Classical Model to assess expert performance is a suitable way of using expert judgment to limit the uncertainty in the "out of sample" range of extremes. The experts-only approach performs satisfactory as well, albeit with a considerably larger uncertainty than the EJ-data option. The method may also be applied to river systems where measurement data are scarce or absent, but adding information on less extreme events is desirable to increase the precision of the estimates.

On a broader level, this study has demonstrated the potential of combining structured expert judgment and Bayesian analysis in informing priors and reducing uncertainty in statistical models. When estimates on uncertain extremes are needed, which cannot sat-

isfactorily be derived (exclusively) from a (limited) data-record, the presented approach provides a means (not the only mean) of supplementing this information. Structured expert judgment provides an approach of deriving defensible priors, while the Bayesian framework offers flexibility for incorporating these into probabilistic results by adjusting the likelihood of input or output parameters. In our application to the Meuse River, we successfully elicited credible extreme discharges. However, case studies for different rivers should verify these findings. Our research does not discourage the use of more traditional approaches such as rainfall-runoff or other hydrodynamic or statistical models. Considering the credible results and the relatively manageable effort required, the approach (when well implemented) can present an attractive alternative to models that approach uncertainty in extremes in a less transparent way.

Appendix 6.A: Downstream discharge calculation

Section 6.3.4 explained the method applied and choices made for calculating downstream discharges. This appendix explains this in more detail, including the mathematical equations.

Three model components are elicited from the experts and data:

- Marginal tributary discharges, in the form of a MCMC GEV-parameter trace. Each combination θ consists of a location (μ), scale (σ), and tail-shape parameter (ξ).
- A ratio between the sum of upstream peak discharges and the downstream peak discharge, represented by This is a single probability distribution.
- The interdependence between tributary discharges, in the form of a multivariate normal distribution.

The exceedance frequency curves for the downstream discharges are calculated based on 9 tributaries (N_T), a trace of 10,000 MCMC parameter combinations (N_M), and 10,000 discharge events (N_Q) per curve.

The N_M parameter combinations for each tributary are sorted based on the (1,000-year) discharge with an exceedance probability of 0.001: $F^{-1}_{GEV}(1-0.001|\theta)$, in which F^{-1}_{GEV} is the inverse cumulative density function, or percentile point function, of the tributary GEV. Sorting the discharges like this enables us to select parameter combinations that lead to low or high discharges in multiple tributaries, and in this way express the tributary correlations. The sorting order might be different for the 10-year discharge than it is for the 1000-year discharge. The latter is however chosen as it is most interesting for this study.

For calculating a single curve, N_T realizations are drawn from the dependence model. These normally distributed realizations ($\mathbf{x}$) are transformed to the $[1, N_M]$ interval, and are then used as index $\mathbf{j}$ to select a GEV-parameter combination for each of the N_T tributaries:

$$\mathbf{j} = Round(F_{norm}(\mathbf{x}) \cdot (N_M - 1) + 1)). \tag{6.8}$$

This is the first of two ways in which the interdependence between tributary discharges is expressed. The second is the next step, drawing a $(N_T \times N_Q)$ sample $\mathbf{Y}$ from the dependence model. These events (on a standard normal scale) are transformed to the discharge realizations $\mathbf{Q}$ for each tributary's GEV parameter combination:

$$\mathbf{Q} = F^{-1}_{GEV,\mathbf{j}}(F_{norm}(\mathbf{Y})) \tag{6.9}$$

An N_Q sized sample for the ratio between upstream sum and downstream discharges ($\mathbf{f}$) is drawn as well. The $(N_T \times N_Q)$ discharges $\mathbf{Q}$ are summed per event (for all tributaries), and multiplied with the factor $\mathbf{f}$,

$$\mathbf{q} = \mathbf{f} \cdot \sum(\mathbf{Q}). \tag{6.10}$$

Note that this notation corresponds to Eq. (6.1). The N_Q discharges $\mathbf{q}$ are subsequently sorted and assigned a plot positions:

$$\mathbf{p} = \frac{\mathbf{k} - a}{N_Q + b}, \tag{6.11}$$

with a and b being the plot positions, 0.3 and 0.4, respectively (from Bernard & Bos-Levenbach, 1955). $\mathbf{k}$ indicates the order of the events in the set (1 being the largest, N_Q the smallest), The plot positions ($\mathbf{p}$) are the 'empirical' exceedance probabilities of the model. With 10,000 discharges and our exceedance probability of interest of 1/1,000, the results are insensitive to the choice of plot positions.

This procedure results in one exceedance frequency curve for the downstream discharge. The procedure is repeated 10,000 times to generate an uncertainty interval for the discharge estimate. Note that the full Monte Carlo simulation comprises $10,000 \times 10,000 = 100,000,000$ 'events' for the 9 tributaries.

Appendix 6.B: Expert estimates

The expert estimates for all items are shown in Fig. 6.B.1. Each panel shows the seven experts' and two decision maker's estimates, through their 5th, 50th, and 95th percentile.

Figure 6.B.1: Expert and DM estimates for all items.

7

Dependence elicitation using non-parametric Bayesian Networks

In absence of sufficient data, structured expert judgment is a suitable method to estimate uncertain quantities. While such methods are well-established for individual variables, eliciting their dependence in a structured manner is a less explored field of research. We tested the performance of experts in constructing and quantifying a non-parametric Bayesian network, describing the correlation between river tributary discharges. Specialized software was provided to assist the experts with this task. Expert performance was investigated using the dependence calibration score (a correlation matrix distance metric) and the likelihood of the joint distribution. Desirable properties of the dependence calibration score are investigated theoretically. Individual expert judgments are combined based on performance into a group opinion or decision maker (as usually called in the expert judgment literature). All experts were able to create and quantify a correlation matrix between 10 variables that resembled the observed correlations well. The decision makers performed similarly to the best expert. Based on the metrics investigated, it mattered little which expert opinions and with what weight, were combined in a decision maker. This is partly because all experts performed well in this study. While this good performance is encouraging, it does eliminate the need of scoring experts and developing scoring rules for dependence elicitation. The results are nonetheless promising: The research shows that experts are able to quickly create and quantify dependence structures, especially when aided by specialized graphical software.

The manuscript related to this chapter is under review at the time of writing. Authors: Rongen, G., Morales-Nápoles, O., Worm, D., & Kok, M.

7.1. Introduction

Scientific models can involve substantial uncertainty, especially when used to predict unprecedented situations. In absence of data or resources to quantify these uncertainties, for example, because of the unfeasibility of large experiments or data collection campaigns, structured expert judgment is a good alternative for quantifying parameters of interest. When sources of uncertainty are related, these dependencies should be assessed in a structured way as well.

Estimating uncertainty, especially multivariate uncertainty, has been a challenge in science and engineering. Methods for estimating univariate uncertainties with expert judgment are well established and include the Delphi method (B. B. Brown, 1968) and the Classical Model, also known as Cooke's method (Cooke, 1991). Most studies on the utilization of expert judgments in science and engineering concentrate on obtaining univariate probability distributions. However, determining multivariate uncertainty (i.e., the joint probability distribution) is a more challenging task that requires not only the evaluation of one-dimensional marginal distributions but also the assessment of the relationships between these distributions. Hence, it poses a larger challenge on experts. To simplify the representation of a joint distribution, various dependence models can be used, each having different characteristics and underlying assumptions. For example, the Bayesian Belief Net (BBN) or Bayesian Network (BN) is a graphical model that depicts the relationship between random variables (the graph's nodes) and their dependence (the graph's arcs) (Darwiche, 2009; Pearl, 2000). Another approach is to assume the dependence follows a multivariate distribution, such as a multivariate normal, t, or Dirichlet distribution. If there are only two dependent random variables, a copula can be used (Nelsen, 2007), which offers greater flexibility in specifying, for example, tail dependence, than the three above-mentioned multivariate distributions.

There are also several methods for eliciting dependence from experts. The choice of method may depend on the type of dependence model being used, and the specifics of the study. Daneshkhah and Oakley (2010) outline several methods for quantifying multivariate distributions and copulas. Morales et al. (2008) explores eliciting conditional rank correlations from experts, while examples of elicitation of non-parametric Bayesian networks (i.e., a specific form of a BN) by experts may be found in (Delgado-Hernández et al., 2014; Morales-Nápoles, Delgado-Hernández, et al., 2014), and (A. M. Hanea et al., 2022). An example of a Delphi based method for eliciting BNs is given by (Nyberg et al., 2022). In Bayesian probability, expert elicitation can also be used to create informed (multivariate) priors when insufficient data are available to specify the posterior without using (expert) informed priors (Al-Awadhi & Garthwaite, 1998; Garthwaite & Al-Awadhi, 2001; Moala & O'Hagan, 2010). For a comprehensive overview of dependence model and their elicitation, see Werner et al. (2017).

While a considerable body of research is available on dependence elicitation, the conclusions on the suitability of different methods for eliciting and scoring results are not straightforward. Additionally, dependence elicitation in structured form (i.e., creating defendable decision makers from experts estimates), requires a procedure for measuring performance, which is a largely unexplored field of research.

We conducted an expert elicitation to determine if expert judgment can be used to accurately elicit multivariate dependence in extreme river discharges for the Meuse River. Seven experts estimated a correlation matrix by specifying a non-parametric Bayesian network (NPBN). They first estimated the tributary discharges (marginals) and then their correlations. Both were then used to calculate extreme river discharges. The experts used software that was provided to help them draw their NPBN and calculate correlations. They were given examples to understand the relationship between data properties and correlation coefficients. The correlation matrices were scored using the *dependence calibration score* or *d-calibration score* (Morales Nápoles & Worm, 2013). These were then used as weights to create decision makers (DMs). We analyzed the performance of these DMs compared to the performance of individual experts and did several sensitivity analyses to test the potential effect of individual expert on the result. Additionally, a significance level for the d-calibration score is calculated to indicate whether an expert's estimate is significantly better than an uninformed guess. Finally, we show theoretical properties of the dependence-calibration (or d-calibrations) score as a desirable metric of expert performance when eliciting dependence. The methods of estimating the marginals, as well as the discharge statistics resulting from the elicited marginals and dependencies, are described in Chapter 6.

7.2. Methods

In this study, experts estimate dependence between discharge peaks of tributaries within a catchment by quantifying a Non-Parametric Bayesian Network (NPBN) network using supporting software. The method description includes the expert elicitation process (Section 7.2.1), the discharge data used (Section 7.2.2), and the method for scoring the experts' estimates (Section 7.2.3). Background information on (Non-Parametric) Bayesian Networks and copulas is found in Section 2.2.3. For this study, the most relevant feature about NPBNs to remember is that it may be characterized by a rank correlation matrix, which is used to quantify the dependence in a multivariate normal copula.

7.2.1. Expert elicitation of correlated tributary discharges

The dependence elicitation presented in this study was conducted as part of a larger expert elicitation focused on extreme discharges of the river Meuse, which runs through parts of France, Luxembourg, Belgium, and the Netherlands. Seven experts participated in the elicitation that took place on 4 July 2022. During this session, the experts estimated the discharge that is exceeded on average once per 10 and 1,000 years. These estimates were then combined with data to form extreme value distributions (Rongen et al., 2024). For calculating extreme discharges along the Dutch part of the Meuse, also the statistical dependencies between tributary discharges were elicited. The results of this, the estimates of statistical dependence, are presented in this article.

Participants were tasked with estimating a correlation matrix representing the dependence between 10 tributaries. A non-parametric Bayesian network (NPBN) was deemed an appropriate tool for this task, for three reasons: Firstly, experts can intuitively consider a 'causal' structure when specifying correlations. Secondly, a NPBN reduces the

number of coefficients to be specified, as only the (conditional) rank correlations for the arcs of the NPBN are needed instead of between each pair of nodes (see Section 2.2.3). Bivariate correlations not directly specified on the arcs of the NPBN are calculated from the specified ones and the conditional independence statements embedded in the graph of the BN. Finally, all specified conditional rank correlations in the NPBN will result in a valid (i.e., positive semi-definite) correlation matrix, while the specification of numbers in $[-1,1]$ for every element of a squared matrix would not necessarily result in a valid correlation matrix. The simplest example for this is a correlation matrix with three variables X_1, X_2, and X_3, in which both pairs (X_1, X_2) and (X_2, X_3) are fully positively correlated. In this case the pair (X_1, X_3) must then be fully dependent as well, as they are related through X_2, with which they are both fully dependent. Any other value than 1.0 between X_1 and X_3 will thus result in an invalid correlation matrix. In case the correlations are strong, but not perfect, such conditions become less clear, but they still need to be satisfied to create a valid correlation matrix.

To assist the experts in creating the NPBN, we developed a GUI-based program called Matlatzinca. This program, based on (Koot et al., 2023; Paprotny et al., 2020), enables experts to easily draw a NPBN by adding nodes and edges, and specifying correlations between them. The program also imposes limits on the correlations that can be assessed by experts, such as the example above of the random vector (X_1, X_2, X_3), helping them in creating valid correlation matrices. Matlatzinca also provides a visualization tool to show the impact of a certain rank correlation coefficient on conditional probabilities, similar to (Morales et al., 2008, Fig. 3 and 4), which are intended to clarify the effect of a specific correlation coefficient.

In structured expert judgment, seed questions are used to determine the experts' performance. Performance-based weights are derived from these seed variables which are then used to obtain the answers for the (unknown) target variables. In this study, experts estimate a correlation matrix that is also calculated from the observations. This enables us to test the experts' performance. We do not separately define tail-dependence for the correlations (i.e., different dependence for the extremes) since the used dependence model does not facilitate the possibility to model these in detail.

7.2.2. Discharge data and peak selection

We obtained the discharge data needed for testing the experts' performance from Service public de Wallonie (2022) for the Belgian gauges, from Waterschap Limburg (2021) and Rijkswaterstaat (2022) for the Dutch gauges, and from Land NRW (2022) for the German gauge. These discharge data are mostly derived from measured water levels and rating curves. During floods, water level measurements can be incomplete and rating curves inaccurate. For our application, this matters less as we elicited rank correlations; measurement errors and errors in the rating curves are less likely to change the ranks (the order of magnitude) than the absolute values.

Figure 7.1 shows the availability of data for the elicited tributaries and Borgharen. Events were selected based on the discharge at Borgharen. Peak over Threshold (PoT) was applied to select every event with a discharge larger than 750 m^3/s within a centered time

window of 15 days (7 days before the peak, the day of the peak, and 7 days after).

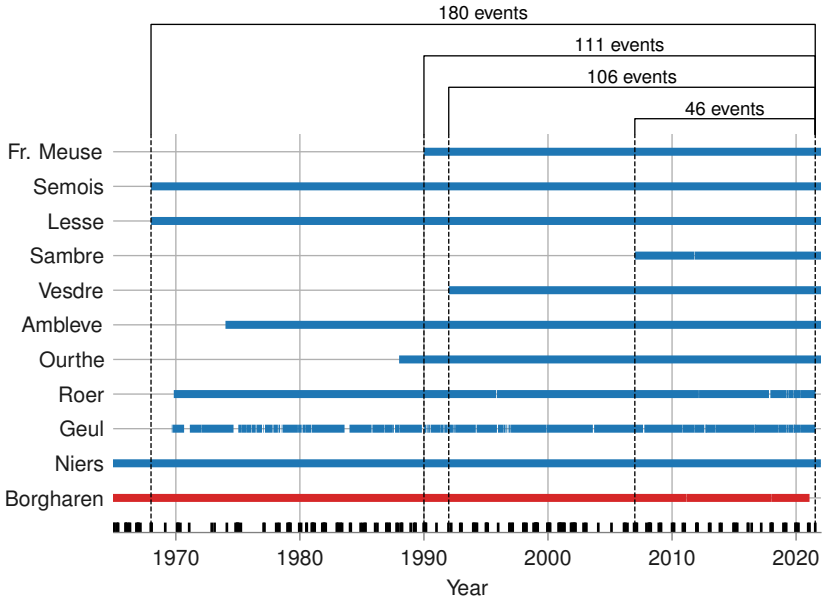


Figure 7.1: Availability of measured discharge data for different tributaries and the main river branch at Borgharen

The time ranges for which data are available varies between different stations. Creating a valid correlation matrix requires complete records, which is why we decided to exclude the time series for the river Sambre when comparing the estimated dependencies and observed dependencies. This resulted in 106 events instead of 46. Omitting the river Vesdre would further increase the number of events to 111 but we consider it to be a more significant tributary and the 5 extra events not worth excluding it. After excluding the Sambre, 9 of the 10 elicited tributaries remain in the correlation matrices.

The 106 events are used to evaluate the performance of experts and DMs in estimating dependence. This is small number of events, considering the 36 unique correlation coefficients that are present in a 9-variable correlation matrix. In several analyses, we account for the uncertainty that results from the specific set of observations by using a non-parametric bootstrap. This involves drawing a random sample with replacement from the observed discharge peaks and calculate the results for that set of events. Some events may appear multiple times in the resampled set, while others may not appear at all.

7.2.3. Scoring the experts' performance

We apply performance-based weighting to combine the different experts' estimates into a decision maker. For the (univariate) tributary discharges, we combined the estimates using the Classical Model (Cooke & Goossens, 2008). The underlying idea is that a (perfor-

mance-based) weighting of expert estimates gives a better estimate than a single expert or an equally weighted combination. We continue on this assumption but need a different score to assess an expert's performance, because the Classical Model is not suited for scoring dependence. We use the d-calibration score instead (Morales Nápoles & Worm, 2013; Morales-Nápoles, Hanea, & Worm, 2014). This score uses the Hellinger distance d_H to compare two multivariate probability distributions. For the case of NPBNs the Hellinger distance is a function of two correlation matrices:

$$d_H(R_1, R_2) = \sqrt{1 - \frac{|R_1|^{\frac{1}{4}} |R_2|^{\frac{1}{4}}}{|\frac{1}{2}R_1 + \frac{1}{2}R_2|^{\frac{1}{2}}}} \tag{7.1}$$

R_1 and R_2 are the two correlation matrices being compared. Notice that if $R_1 = R_2$, $d_H = 0$, while the maximum value d_H may take is 1. The d-calibration score for expert e, $dCal(e)$, is defined as:

$$dCal(e) = 1 - d_H(R_q, R_e). \tag{7.2}$$

This score consequently varies on a scale from 0 to 1 and can be used as weights (after normalization) to calculate decision makers (DMs) similar to the Classical Model (Cooke, 1991). In Eq. (7.2), R_q denotes the observed correlation matrix to be used for calibration purposes and R_e the expert estimated correlation matrix. The d-calibration score has the following properties: a) an expert will receive the maximum score when and only when she/he captures exactly the observed dependence structure; b) an expert may get a low calibration score if, for example, a high correlation between a pair of variables was expressed by the expert while this was not expressed by the true dependence structure R_q (or vice-versa); and c) a necessary condition for an expert to be highly calibrated is to sufficiently approximate the dependence structure of interest entry-wise. A formal treatment of the d-calibration score and proofs of the properties discussed are presented in Section 7.A. Other scores may be used as well. However, their properties have not been investigated by the authors to a similar extent as the d-calibration score (Section 7.A) and are therefore not considered in this research.

We did however consider the likelihood to check if the d-calibration performs as expected. Likelihood is a measure to compare a probabilistic model with observations. The probability density function of the used MVN-distribution is:

$$f(\mathbf{q}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{q}-\mu)^T \Sigma^{-1} (\mathbf{q}-\mu)\right). \tag{7.3}$$

The discharge observations $\mathbf{q}$ is a vector with a realization for each of the k tributaries. Σ is the covariance matrix. By transforming the observations to standard normal space (i.e., $\mathbf{x} = \Phi^{-1}(\text{rank}(\mathbf{q}))$) the covariance matrix Σ becomes the correlation matrix R, and the mean μ drops out. The log-likelihood then becomes:

$$\ell(R|\mathbf{x}) = \log\left(\frac{1}{\sqrt{(2\pi)^k |R|}}\right) - \frac{1}{2}\mathbf{x}^T R^{-1} \mathbf{x}. \tag{7.4}$$

The log-likelihood is not a probability and does not range from 0 to 1. With a 9-variable MVN-distribution, the likelihoods are in general very small and will vary greatly (more

or less exponentially) between experts. This means that a single expert will almost always have close to 100% of the weight making it too strict to use as performance-based weight. We did however use it to further investigate the performance and consistency of the d-calibration score. Note that the log-likelihood compares the observations to the (chosen) MVN-distribution that corresponds to the estimated correlation matrix, while the d-calibration score compares the observed rank correlation and estimated matrix directly.

7.3. Results

7.3.1. Bayesian networks and correlation matrices

The Bayesian networks quantified by individual experts are shown in Fig. 7.2. We underscored that the primary goal of the expert judgment exercise was accurately obtaining the correlation coefficients of interest. The general approach for quantifying the correlations was that experts chose to connect neighboring tributaries and assigned (conditional) rank correlations to the arcs such that the resulting non-conditional correlation matches their estimate. Expert C and expert G adopted an approach in which different catchments are linked through hierarchical nodes presenting precipitation. Expert C additionally connected the tributaries upstream to downstream, while expert G created three fully connected groups connected through parent precipitation nodes.

The Bayesian networks in Fig. 7.2, together with the experts' assessments of (conditional) rank correlations, give the correlation matrices shown in Fig. 7.3. The observed correlation matrix, which is the one against which experts' performance will be evaluated, is shown in the top left matrix. Expert A estimated generally high correlations (higher than observed), experts C, F, and G present lower correlation coefficients than A, while the lowest correlation coefficients are estimated by experts E, D, and B. The hierarchical approach used by expert C and G did not result in distinctly different matrices. The hierarchical grouping of variables is more visible in Expert C's matrix compared to Expert G, although it is also present in Expert A's matrix who did not adopt a hierarchical approach.

7.3.2. Experts' and decision makers' performance

Scores

Table 7.1 shows the d-calibration scores (higher is better) and log-likelihoods (less negative is better) calculated from the expert correlation matrices. For comparison, the statistical accuracy scores according to the Classical Model from (Rongen et al., 2024) are shown in the last column. Note that these are calculated from the expert's univariate estimates. Morales-Nápoles, Hanea, and Worm (2014) found that statistical accuracy (the Classical Model) and d-calibration score are generally, but not always, well correlated, meaning that the experts that estimate univariate random variables accurately also perform generally good for estimating multivariate uncertainties.

To put the experts' performance into context, two additional d-calibration scores are presented: The first are the scores for the observed correlation matrix. Estimating this

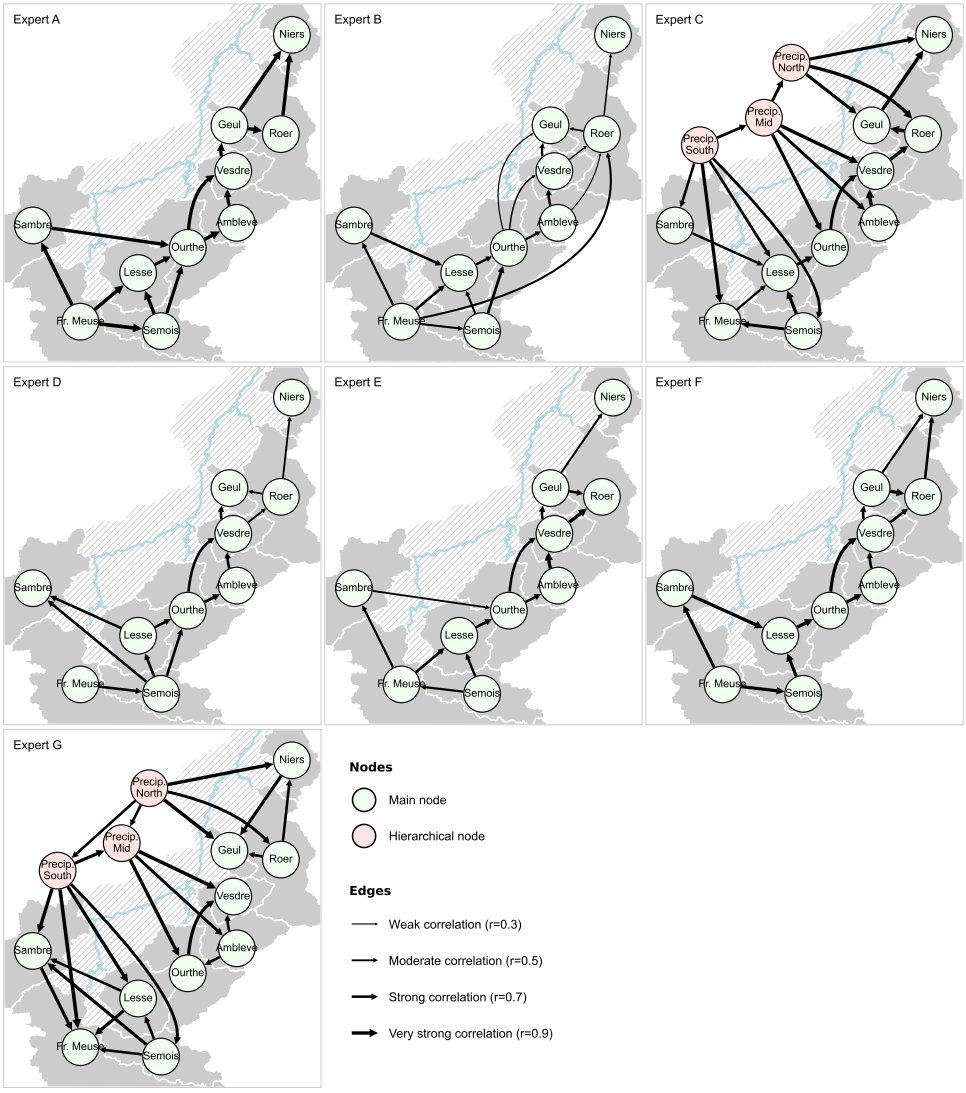


Figure 7.2: Bayesian networks as drawn by the experts. The thickness of the arrows show the strength of the *non-conditional* correlation. The grey areas on the background represent a map of the catchments between which the dependence is elicited, with the blue line showing the main branches of the Meuse River.

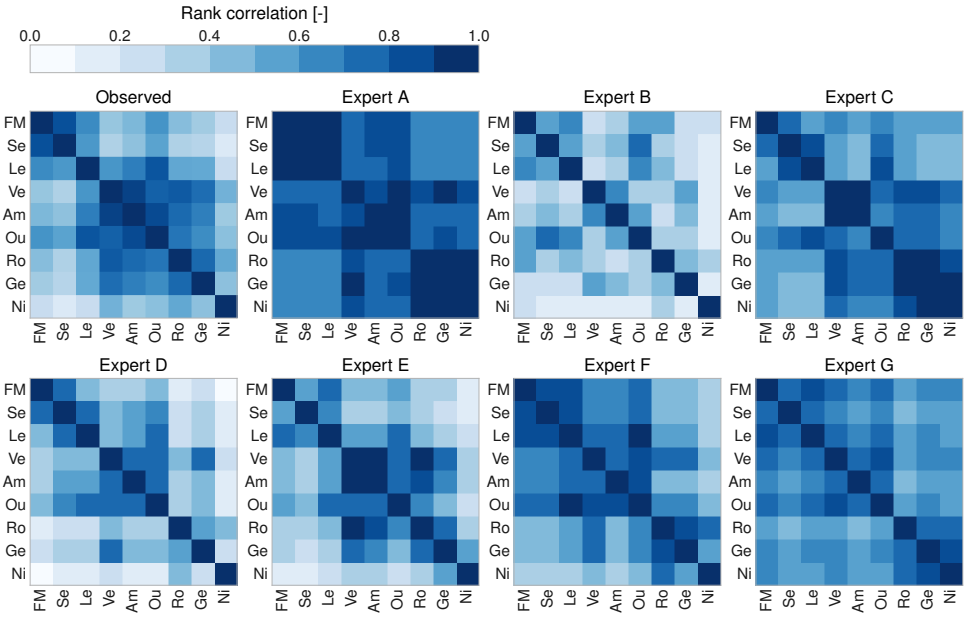


Figure 7.3: Correlation matrices corresponding to the NPBNs drawn and quantified by the experts.

would give the best possible d-calibration score and log-likelihood. The second result is a 5% significance level. A score above this level indicates that it is unlikely (<5% probability) that the expert's matrix is uninformed, i.e., part of the population of randomly drawn NPBNs. As there is no well established method for deriving such a criterion in the context of dependence elicitation, we derived one ourselves. This was done by randomly sampling NPBNs with uniform, non-negative, (conditional) rank correlations on the edges, and calculated the resulting d-calibration scores. The 95^{th} percentile of these scores, which is 0.15, is the significance level. This value depends on the number of variables and the assumptions for sampling the matrices. The method and results for this approach are explained in Section 7.B.2.

Based on both the d-calibration score and the log-likelihood shown in Table 7.1, Expert E's correlation matrix is best, closely followed by experts F and D. Expert A has the lowest score, but it is still higher than the 5% significance level. Experts B, C, and G have a score roughly in between the scores of A and E.

The global weights (GL) DM is a weighted average of the experts' correlation matrices, in which the normalized d-calibration scores are the weights. The equal weights (EQ) DM is the average of the matrices, without differentiating weights between experts. Both DMs have a high d-calibration score compared to most experts, but slightly lower than the best expert (GL has a closer to zero log-likelihood, which implies better performance than the best expert). EQ has a slightly higher d-calibration score than GL, but GL has a better log-likelihood. The global weight DM with optimization (GL opt.) is calculated by selecting experts based on a minimum required d-calibration score and calculating

Table 7.1: d-calibration scores and likelihood for experts' and DMs' correlation matrices

	dependence calibration score	Log-likelihood	Statistical accuracy (Classical Model)
Expert A	0.165	-2442.6	$7.99 \cdot 10^{-4}$
Expert B	0.308	-1016.5	$4.56 \cdot 10^{-4}$
Expert C	0.284	-1396.5	$2.3 \cdot 10^{-8}$
Expert D	0.371	-961.7	0.683
Expert E	0.444	-933.7	0.192
Expert F	0.411	-993.4	$4.56 \cdot 10^{-8}$
Expert G	0.268	-1184.6	$6.29 \cdot 10^{-3}$
EQ DM	0.439	-937.1	
GL DM	0.437	-932.5	
GL opt. DM ($dCal > 0.411$)	0.468	-923.3	
Observed	1.000	-812.2	
95% significance level (α)	0.150		

the weighted average of the included experts' matrices. Using a minimum d-calibration score of 0.411 results in the optimum (i.e., highest d-calibration score) by giving a non-zero weight to experts E and F. The result is a slightly higher score than the GL and EQ DMs and better than any of the experts.

Finding the 'best' decision maker

While it is encouraging to see that all DMs score similarly to the best expert, the resulting scores are not distinctively better. The equal and global weights are practically equal for the case under investigation (this is not always the case, see for example: Morales-Nápoles, Hanea, & Worm, 2014). This is partly due to the experts' d-calibration scores being close together, especially in comparison the scores from the Classical Model. Because the DMs are hardly distinctive, it is interesting to see what the ideal weight distribution (i.e., the 'best' DM) would look like. To further examine this, we optimized the weights by maximizing both the d-calibration score and log-likelihood. The approach is different from the GL opt. DM, where the weights are restricted to the d-calibration scores in combination with a cut-off level. Instead, we optimized while allowing all experts' weights to vary freely. To ensure stability of the optimum, we used different starting points. Figure 7.4 shows the results for this. The left bars show the optimized weights when optimizing the log-likelihood, the right bars when optimizing the d-calibration score. The maximum log-likelihood is -910.5 and the maximum d-calibration score 0.506, both higher than the DM result calculated directly. The observations were bootstrapped to check the sensitivity of the optimum to the specific set of observed events. The thin lines, a kernel density estimate of the resulting weights from bootstrapping, illustrate the uncertainty in the factor under re-sampling.

Surprisingly, the results of the weights optimization are different from the d-calibration scores in Table 7.1. Experts B, F, and G are given almost zero weight, despite having well-

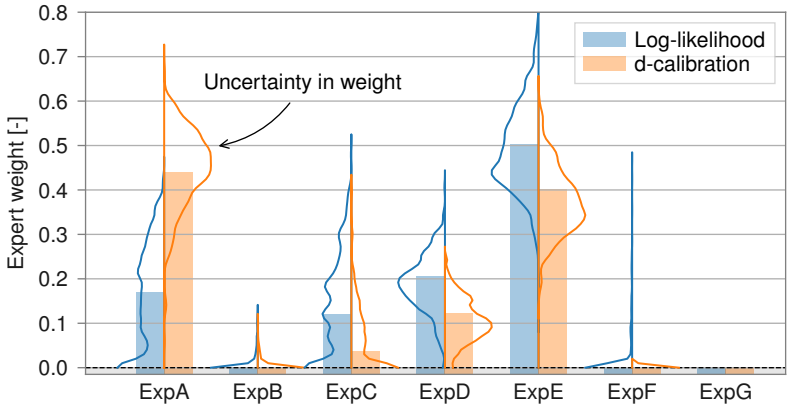


Figure 7.4: DM weights optimized based on the log-likelihood and d-calibration score, including uncertainties from bootstrapping.

approximated correlation matrices (e.g., F had the second-best d-calibration score). At the same time, expert A, with the lowest score, is assigned a very large weight. This is due this experts' high estimated correlations (see Fig. 7.3), which compensate for the weaker than observed correlation estimates from the other experts. This effect is stronger for the d-calibration score than when using log-likelihood. The opposite happens for expert B, who estimated the lowest correlations. Experts F and G do not seem to add a unique contribution to the weighted sum, which leads to their low weights. Whether this inconsistency between optimal weights on one side and d-calibration scores and likelihoods on the other side is a systematic feature of the weighing scheme or a feature of this particular exercise on dependence elicitation, remains an open question.

Robustness of the decision makers

The optimized results from last section suggest that a few specific experts should be included when constructing the DM. To assess the sensitivity of the DM to specific experts, the d-calibration scores were calculated for different combinations of experts. This is similar to checking the expert robustness in the Classical Model. The results for using equal weights are visualized in Fig. 7.5, with (a) showing the DM score for each combination including a specific expert and (b) by showing the score per combination of 1 to 7 experts. The results for the global weight DM are very similar and therefore presented in Section 7.B.1, including the robustness of the GL DM with optimization.

The results show that the performance of the DMs is relatively insensitive to the individual experts in this specific set. The differences in average scores for each expert are less than 0.05 (as Fig. 7.5 shows). Surprisingly, it matters little which experts are combined, the number of experts is more important for a good score (as Fig. 7.5 b shows). The average of the covariance matrices of multiple experts tends to result in a better performance than the individual matrices, both in terms of the d-calibration score and the (log-)likelihood. This pattern is also observed when comparing the average of a sample

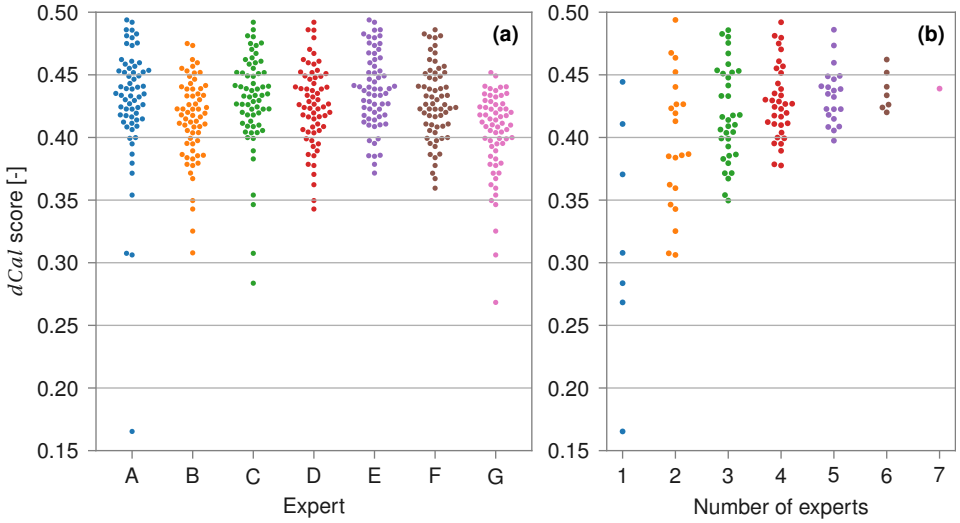


Figure 7.5: Robustness of the equal weight decision maker. Left (a), the score for all combinations including the expert. Right (b), the score for each combination of 1 to 7 experts.

of random correlation matrices to another random correlation matrix. Section 7.B shows the details of this.

While a combination with 2 experts gives the highest score (this is the GLopt DM in Table 7.1), including more experts is a more robust option as every combination of 4 experts gives a d-calibration score varying between 0.35 and 0.50. Using the average of a few experts' matrices represents the observed correlations better than most of the individual matrices. This is however closely tied to the experts' good individual scores. When a low-scoring expert is in the pool, the results do become more sensitive to individual experts. Section 7.B.3 shows this, by adding a hypothetical low-scoring expert to the pool. Doing this does make the results sensitive to individual experts, reinstating the importance of scoring, especially because the expert weights are relatively close together (a 'bad' expert is still likely to have a substantial weight). After filtering experts based on the significance level the pattern, that the mean matrix performs on average better than the individual matrices, reappears.

7.4. Discussion and final remarks

In this study, we elicited the dependence of a river's tributaries' peak discharges from experts, by making them construct and quantify a non-parametric Bayesian network. The experts were scored by the d-calibration score and likelihood of their matrix. Sensitivity analyses were done on their results, to see what combination of experts' and what scores give the best result. Dependence elicitation is much in its infancy still. By sharing our findings and insights on the elicitation process (Section 7.4.1) and the more theoretical aspect of scoring (Section 7.4.2), we hope to contribute to the progression of the field of

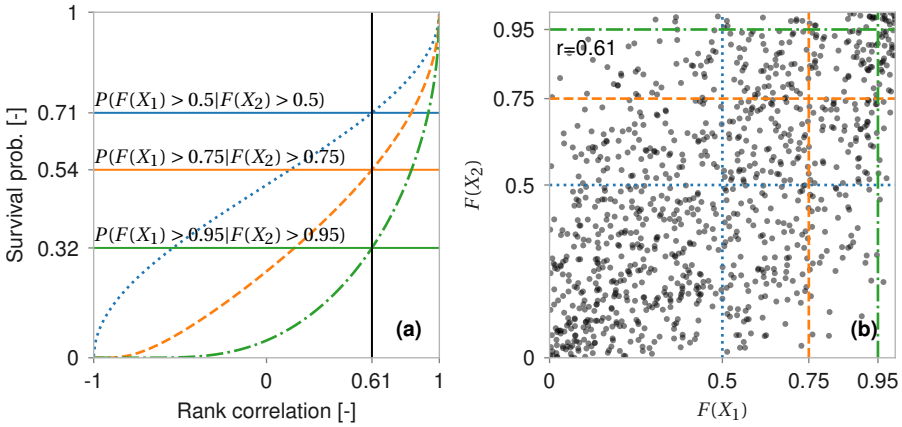


Figure 7.6: Example of information provided to the experts to aid them in estimating rank correlations.

dependence elicitation.

7.4.1. Practice of expert dependence elicitation

Non-parametric Bayesian networks are not uncommon in hydrological modeling (e.g. Paprotny & Morales-Nápoles, 2017; Ragno et al., 2022) but unknown to most hydrologists. The concept of a NPBN is for a non-statistician difficult to master within the short period of time that is usually available in the preparation of an expert elicitation. This did however not limit most experts in creating a NPBN that represented the observed correlations well with relative ease. The experts were done within half an hour, while we were initially doubting whether a 10-node network would be too much of a strain for the experts. The quick results for a) building and b) quantifying the NPBN contrasts with for example (Barons et al., 2022; A. M. Hanea et al., 2022), who explicitly split the two phases in the elicitation process. In our study, the experts needed to estimate correlations between a single physical quantity (river tributary discharges), reducing the burden of building the network. Additionally, the graphical interface (see Fig. 3.4) in which experts can directly see the effect of their structure and estimates rank correlations, likely makes the elicitation process easier for the experts.

Experts estimated rank correlations directly but were given graphical aid to inform these coefficients, for which an example is given in Fig. 7.6 a. However, during the elicitation, the experts suggested using scatter plots for relating correlation coefficients. This information was provided to them by generating samples from bivariate normal distributions with specific correlation coefficients. These were transformed to ranks, to make it easier for experts to relate it to a fraction of the discharge peaks (e.g., the correlation between the highest 10% discharges for tributary A and B). Figure 7.6 b. shows an example. For the participants, this representation was more intuitive than the accurate but abstract relation between conditional probability and rank correlation coefficient (i.e., estimating a correlation coefficient based on Fig. 7.6 b was perceived easier than based on Fig. 7.6 a).

7.4.2. Dependence scoring

The d-calibration score was used to score the experts performance and provide weights for the decision makers. A scoring rule should help with selecting the most accurate experts and ideally also assign weights such that the (weighted) combination of experts performs better than the best expert. Comparing the more established log-likelihood to the d-calibration score shows that the d-calibration score does indeed select the most accurate estimates. However, it did not result in a DM that performs significantly better than the individual experts, as this requires a set of weights that gives a greater weight to the worst performing expert (as shown in Section 7.3.2). This is inconsistent with the d-calibration scores as well as the (log-)likelihoods and is therefore unlikely to be an indicating of what a better scoring rule would be. Note that we were able to find the optimal weight because the ideal answer (i.e., the observed correlation matrix) was known. However, a score should also be trusted upon when the elicited dependencies are unknown.

Regarding the sensitivity of the d-calibration scores, we observed the following:

- All experts scored better for estimating dependencies than the derived significance level of (a d-calibration score of 0.15), while only 2 of the 7 experts scored above the significance level for the Classical Model (for estimating univariate uncertainties).
- The DM results are relatively insensitive to weights (derived from the d-calibration score) assigned to the experts; equal weights (EQ) scores similar to global weights (GL), and it does not matter much for the score which experts are combined into a DM. On top of that, increasing the number of experts in the decision maker increases the resulting DM-score for most combinations of experts.
- This changes when a hypothetical low-scoring expert is added to the pool. This makes the GL DM perform better than the EQ DM, and the results become more sensitive to the specific experts included in the DM.

These findings underscore the importance of scoring the experts and using a significance level or optimization to filter out 'bad' results. This is especially the case for the d-calibration score because the score is less rigorous, such that 'bad' results will still get a significant weight (in contrary to the Classical Model). What a generally suitable cut-off level or significance level is, is yet to be determined. The random matrix sampling might give a good indication as the expert-insensitive results indicate that the significance level, a d-calibration score of 0.15, higher than the hypothetical low-scoring expert, rightly included all experts.

It is encouraging to see that all 7 experts were able to provide good estimates for dependence, while only 2 experts had univariate estimates that scored above the significance level. We are aware that this is a comparison between two different scoring rules, and an observation from only a single study. Future research should therefore focus on cross-checking these results with past dependence elicitation studies, and if needed, on performing extra studies to generate more empirical data. This would help the research on dependence scoring rules and methods for combining dependence estimates.

7.4.3. Conclusions

This study set out to a) estimate multivariate dependencies with expert judgment and b) analyze the behavior of the d-calibration score that is used for joining different experts' results into a single decision maker. Experts estimated the dependence between peak discharges of tributaries within a river catchment, using a non-parametric Bayesian network and graphical software for support. The experts were well able to reproduce the observed dependencies in data, with all experts performing significantly better than a 5% significance level calculated from generated random networks.

The decision maker, a (weighted) combination of experts, scored similar to the best expert. It succeeded in picking out the best experts but did, in this case study, not generate a significantly 'better' expert. We observed that the more experts are included in the weighing pool, the higher the DM-score becomes on average. It does not significantly exceed the best expert's score, but the score is consistently higher than the average of the included scores and relatively insensitive to the specific included experts. This observation is closely tied together with the fact that all experts scored above the significance level for the d-calibration score. Adding a (hypothetical) low-scoring expert to the pool does make the results sensitive to individual results, thereby underscoring the relevance of expert weighting. The good expert estimates are an encouraging result for the field of dependence elicitation and contrast to the scores for their univariate estimates, in which only two experts exceeded the significance level.

This research shows promising results for eliciting dependence structures using graphical software and combining the experts' estimates. We advice comparing the results to a larger set of studies, including dependence structures with a) more physical quantities in a more complex structure and b), that include more divergent or negative correlations. While the d-calibration score has useful properties and performs satisfactorily, a comparison to other dependence scoring rules is needed to see if this can be improved. This does however not compromise the outcome of this study, which is that experts were able to quickly create and quantify dependence structures for river tributary discharges, that well represent observed dependencies.

Appendix 7.A: Proofs for d-calibration score properties[1]

In Moustafa et al., 2010 several measures of distance between Gaussian densities are discussed. We consider the Hellinger distance $d_H(N_1, N_2) = \sqrt{1-\eta(N_1, N_2)}$ where $N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$ are two Gaussian densities with covariance matrices Σ_1, Σ_2, and vector means μ_1, μ_2, and η is as in Eq. (7.5). Notice that the notation used in this appendix is slightly different from that used in the main body of the paper to make this section more self contained.

$$\eta(N_1, N_2) = \frac{|\Sigma_1|^{\frac{1}{4}}|\Sigma_2|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{\frac{1}{2}}} \times \exp\{-\frac{1}{8}(\mu_1-\mu_2)^T \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2(\mu_1-\mu_2)\} \tag{7.5}$$

[1] This appendix was authored by Worm, D. & Morales-Nápoles, O.

The dependence structure of a multivariate random vector as modeled by a copula is not disturbed by monotone transformations of the marginal distributions. In other words, by transforming the marginal distributions to standard normal because of the normal copula assumption in NPBNs, we may work out all calculations on a joint normal distribution with standard normal margins. The advantage of modeling dependence with copulas is that no assumptions need to be placed on the marginal distributions and all calculations can be performed using their transformed form. After such transformation, we can rewrite Eq. (7.5) for the transformed variables. Then the exponent term vanishes and Σ_1, Σ_2 correspond to correlation matrices. Subsequently, we will write $d_H(\Sigma_1, \Sigma_2)$ to denote the Hellinger distance between two normal copulas with correlation matrices Σ_1 and Σ_2 as follows:

$$d_H(\Sigma_1, \Sigma_2) = \sqrt{1 - \frac{|\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{\frac{1}{2}}}} \tag{7.6}$$

As discussed in Moustafa et al., 2010, the Hellinger distance satisfies the axioms of a metric: it is non-negative, it equals zero if and only if $\Sigma_1 = \Sigma_2$, it is symmetric and it satisfies the triangle inequality. Observe that its maximum value is 1, which is attained if $|\Sigma_1| = 0$ (there is some linear combination between pairs of variables) and $|\Sigma_2| > 0$ or vice versa. Another property that makes d_H interesting for our purposes is that if the d_H metric between two matrices is small enough, the pairwise differences between the entries of these matrices must be small as well. This property follows from Theorem 1 below. $||\cdot||_\infty$ denotes the supremum norm, that is, $||B||_\infty = \max_{i,j}(b_{i,j})$. Note that the pairwise differences between the entries of two matrices A and B are bounded from above by $\|A - B\|_\infty$.

Theorem 1. *Let Σ be in C_n with $|\Sigma| > 0$, where C_n denotes the space of n-dimensional correlation matrices. For all $\epsilon > 0$ there exist a $\delta > 0$, such that for each Σ_1 in C_n with $d_H(\Sigma, \Sigma_1) < \delta$ the relation $||\Sigma - \Sigma_1||_\infty < \epsilon$ holds.*

Proof. Let $a > 0$. We define $X = (C_{n,a}, \|\cdot\|_\infty)$, the metric space of n-dimensional correlation matrices whose determinant is larger than or equal to a ($a > 0$), endowed with the supremum norm. Then X is compact, since it is a closed subset of the compact set C_n. Let Y be the metric space $(C_{n,a}, d_H)$. Since it is a metric space, it is Hausdorff.

Finally, let $f : X \to Y$ be the identity map sending matrix A to A. Then it is a bijection from X to Y. It is also continuous: If $(A_k)_k$ converges to A in supremum norm, then it converges entry-wise to A. Since the determinant is a polynomial of the entries of a matrix, and hence continuous, we see that $|A_k|$ must converge to $|A|$. From this, it follows that $d_H(A_k, A) \to 0$ as well.

A basic theorem from topology (i.e. Armstrong, 1983, Theorem 3.7), implies that f is a homeomorphism. Therefore, the identity map from $Y = (C_{n,a}, d_H)$ to $X = (C_{n,a}, \|\cdot\|_\infty)$ is continuous.

Let Σ in C_n be such that $|\Sigma| > 0$. Then in particular $b := \frac{|\Sigma|^{1/2}}{|\Sigma|^{1/4}} > 0$.

From the Minkowski determinant theorem (see Marvin and Henryk, 1992) it follows that for all positive semi-definite matrices A and B, $|A+B|^{1/2} \geq |A|^{1/2} + |B|^{1/2}$. Applying this equality on the Hellinger distance, we can compute that $d_H(\Sigma, \Sigma_1) \geq \sqrt{1-|\Sigma_1|^{1/4} b}$.

From this it follows that if $d_H(\Sigma, \Sigma_1) < \gamma$ for some $\Sigma_1 \in C_n$, then $|\Sigma_1| \geq \left[\frac{1-\gamma^2}{b}\right]^4 =: c$.

Let us choose $\gamma = 1/2$, then $c > 0$ and thus $\Sigma_1 \in C_{n,c}$ for all $d_H(\Sigma, \Sigma_1) < \gamma$. Let $a = \min(c, |\Sigma|)$. Then $a > 0$. For all $\epsilon > 0$ there is a $0 < \delta < 1/2$ such that for all $\Sigma_1 \in C_n$ with $d_H(\Sigma_1, \Sigma) < \delta$, $\|\Sigma_1 - \Sigma\| < \epsilon$, since the identity map from $Y = (C_{n,a}, d_H)$ to $X = (C_{n,a}, \|\cdot\|_\infty)$ is continuous.

□

Theorem 1 implies that if the Hellinger distance from an arbitrary correlation matrix Σ_1 to the given correlation matrix Σ is close to zero, then the correlation matrices must be entry-wise close to each other as well. This is an essential important property in our context.

Based on the Hellinger distance we propose the *dependence-calibration* or *d-calibration* score to be defined as follows:

Definition 1. *Let Σ_T be the true (target) and known correlation matrix of an n-dimensional distribution used for calibration purposes. Let Σ_e be the correlation matrix elicited from expert e. Then the d-calibration of expert e is:*

$$dCal(e) = 1 - d_H(\Sigma_T, \Sigma_e).$$

Analogous to Cooke's classical model, the d-calibration score would in general be computed using a set of seed questions regarding known parameters (correlations) from the dependence structure Σ_T. The values of these parameters would only be known by the analyst, and not by the experts at the moment of the elicitation. The questions used to elicit the correlations used for calibration purposes should be as close as possible to the context of the unknown dependence estimates of interest. In this appendix Σ_T is the generic notation for the target correlation matrix realized by the appropriate NPBN (calibration) model. For example, the observed correlation matrix shown in Fig. 7.3.

The following properties of the d-calibration score hold:

Theorem 2. *Let the d-calibration score be defined as in Definition 1. Assume that the target correlation matrix Σ_T satisfies $|\Sigma_T| > 0$. Then the following properties hold:*

a) *$dCal(e) = 1$ if and only if $\Sigma_e = \Sigma_T$.*

b) *Let $(e_m)_m$ be a sequence of experts. Then $dCal(e_m) \to 0$ as $m \to \infty$ if and only $|\Sigma_{e_m}| \to 0$ as $m \to \infty$.*

c) *Let $(e_m)_m$ be a sequence of experts. Then if $dCal(e_m) \to 1$ as $m \to \infty$, then $(\Sigma_{e_m})_{i,j} \to (\Sigma_T)_{i,j}$ as $m \to \infty$.*

Proof. Property a) follows from the fact that d_H is a metric (2010). From the Minkowski determinant theorem (see Marvin and Henryk, 1992) it follows that

$$\frac{|\Sigma_T|^{\frac{1}{4}}|\Sigma_{e_m}|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_T+\frac{1}{2}\Sigma_{e_m}|^{\frac{1}{2}}} \leq \frac{|\Sigma_T|^{\frac{1}{4}}|\Sigma_{e_m}|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_T|^{\frac{1}{2}}+|\frac{1}{2}\Sigma_{e_m}|^{\frac{1}{2}}} \to 0$$

as $|\Sigma_{e_m}| \to 0$. For the converse direction, note that $d_H(e_m) \geq |\Sigma_T|^{\frac{1}{4}}|\Sigma_{e_m}|^{\frac{1}{4}}$, since the determinant of a correlation matrix is less than or equal to 1. Therefore, if $d_H(e_m) \to 0$, $|\Sigma_{e_m}| \to 0$ as well. This proves property b).
Property c) follows directly from Theorem 1.

□

Remark 1. *Each property from Theorem 2 can be understood as a characterization of a desirable propriety of an elicited correlation matrix. Property **a)** means that an expert will receive the maximum d-calibration score when and only when they capture exactly the true/target dependence structure; property **b)** indicates that an expert may get a low calibration score if, for example, a high correlation between a pair of variables was expressed by the expert while this was not expressed by the true dependence structure Σ_T (or vice-versa); and property **c)** implies that a necessary condition for an expert to be highly calibrated is to sufficiently approximate the dependence structure of interest entry-wise.*

We want to use the d-calibration score to decide whether an expert has approximated sufficiently well the true/target correlation matrix. We do this by constructing the empirical distribution of $dCal(T)$ using a sample of given size from the normal copula with correlation matrix Σ_T. Then we observe whether the value of $dCal(e)$ falls below a particular percentile (significance level) of the empirical distribution of $dCal(T)$. Thus, we test the following hypothesis:

H_0: $dCal(e)$ comes from the distribution of $dCal(T)$.

Rejecting H_0 would give grounds to believe that the difference between the target (calibration) correlation matrix and the expert's assessments may not be exclusively due to sampling fluctuation.

Appendix 7.B: Behavior of the d-calibration score

In Section 7.3.2, the robustness of the decision makers was tested by evaluating the d-calibration score for different combinations of decision makers. The results show that, on average, the mean of the covariance matrices performs better than the individual matrices from which the mean is calculated. This appendix shows more details of that analysis (Section 7.B.1), and investigates if these findings hold for randomly sampled correlation matrices as well (Section 7.B.2). This random matrix sampling is used to define a significance level for the d-calibration score. Finally, Section 7.B.3 shows the effect of adding a low-scoring expert to the pool.

7.B.1. Robustness of mean matrices for global DM

Where Fig. 7.5 shows the robustness for equal weights decision maker to the individual experts (a) and number of experts (b), Fig. 7.B.1 shows this for the global weights decision maker. The weights are calculated by normalizing the experts d-calibration scores. The difference between the EQ and GL robustness results is negligible. To show the robustness of the global optimized decision maker as well, the expert combination with the highest *individual* weights is circled in Fig. 7.B.1 b. Note that the actual global optimized DM is the circled dot for two experts, as the method for determining the global optimized DM is calculating all the circled dots, and selecting the one with the highest score. Interestingly, the circled dot for two experts is not the highest scoring two-expert combination, neither is this the case for the three, four, five, and six-expert combinations. This is consistent with our findings in Section 7.3.2, which showed that the 'best' combination is not necessarily a combination of experts with the highest d-calibration scores.

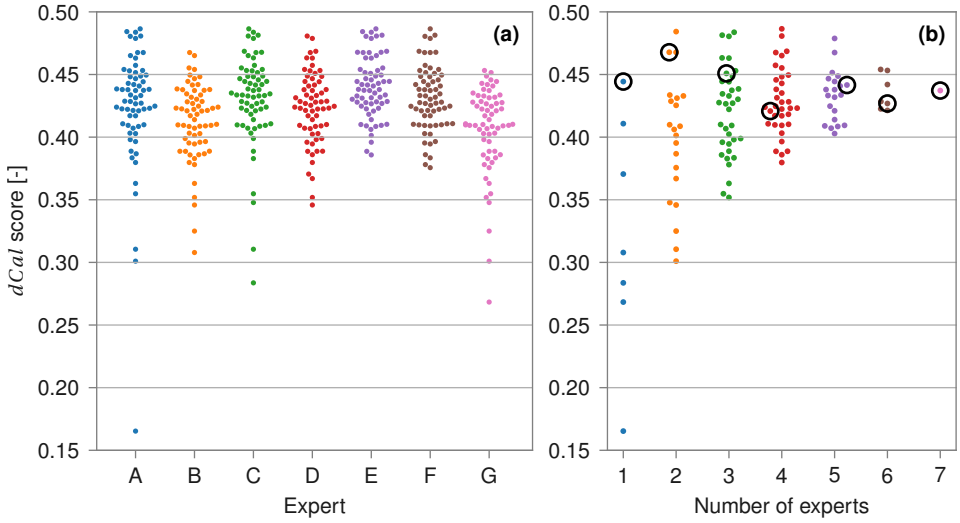


Figure 7.B.1: Robustness of the global weight decision maker. Left (a), the score for all combinations including the expert. Right (b), the score for each combination of 1 to 7 experts.

The d-calibration score (i.e., global DM weight before normalization) of the highest scoring expert is about three times as high as the lowest scoring expert's score. This variation between scores is smaller than what is usually observed in the Classical Model for univariate uncertainty (see for example Cooke & Goossens, 2008). If differences between experts' d-calibration scores would be larger, we would observe variations in the GL DM score like the results for 1 or 2 experts in Fig. 7.B.1 b, as this is the number of experts that usually share the majority of the weight in the Classical Model.

A different representation of the effect of averaging the matrices is shown in Fig. 7.B.2. Every marker in this scatter plot represents a combination of experts (the color indicating the number of experts in the combination). The x-position shows the mean of the

individual experts d-calibration scores (in that combination), and the y-position the d-calibration score of the experts' mean matrix. In other words, the further the marker is located to the upper-left corner, the greater the improvement in score from averaging the individual matrices. The diagonal line gives the average increase of the mean matrix's score to the mean score of the individual matrices, for each number of experts in the combination. The average increase in score is listed in the figure's legend as well. Notice that there is a consistent gain by combining experts estimates. However, after combinations with 3 experts, the average gain is minimal for the case under investigation.

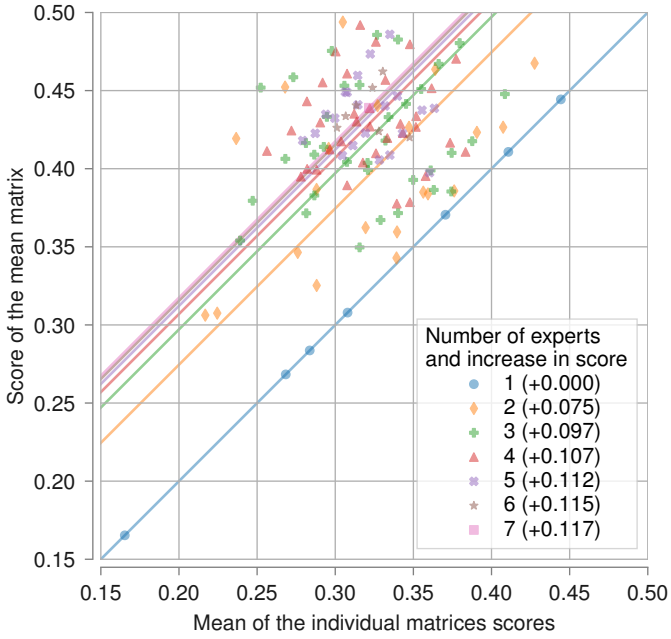


Figure 7.B.2: d-calibration scores of the mean matrix, compared to the average scores of the individual contributing matrices.

7.B.2. d-calibration scores for random matrices

Last section's analysis showed that combining the correlation matrices that were elicited from the experts consistently results in a better d-calibration score. To test this increase in scores in a more general setting, random correlation matrices were sampled, averaged (using equal weights), and compared to another random matrix.

Random correlation matrices were sampled by generating a saturated graph of 9 nodes (similar to the number of nodes in this study) and 36 edges. Each edge was then assigned a random conditional rank correlation sampled from a uniform $[-1,1]$ distribution $U(-1,1)$. This approach is known as the Vine-method (Joe, 2006). We chose this specific method because a), it is consistent with what an expert would do when ran-

domly quantifying a BN (under the assumption that the graph is saturated and the coefficients are drawn from $U(-1,1)$) and b), it is easy to create matrices with constraints on the distribution of the conditional rank correlations. The sampling procedure is:

1. Generate 100,000 correlation matrices using the Vine-method.
2. Pick a 'true' matrix and a matrix guessed by each of the N experts from the set.
3. Calculate the mean matrix of 1, 2, 3, 7 and all (experts') matrices, and compare each to the 'true' matrix by calculating the d-calibration score.

Figure 7.B.3 shows the results for this, for 1, 2, 3, 7, and an 'infinite' number of experts. To simulate the average matrix of infinite experts, the average of the 100,000 sampled matrices was used. Note that this closely matches a correlation matrix representing independence (i.e., all zeros except for ones on the diagonal) for the Vine-method under the mentioned preconditions.

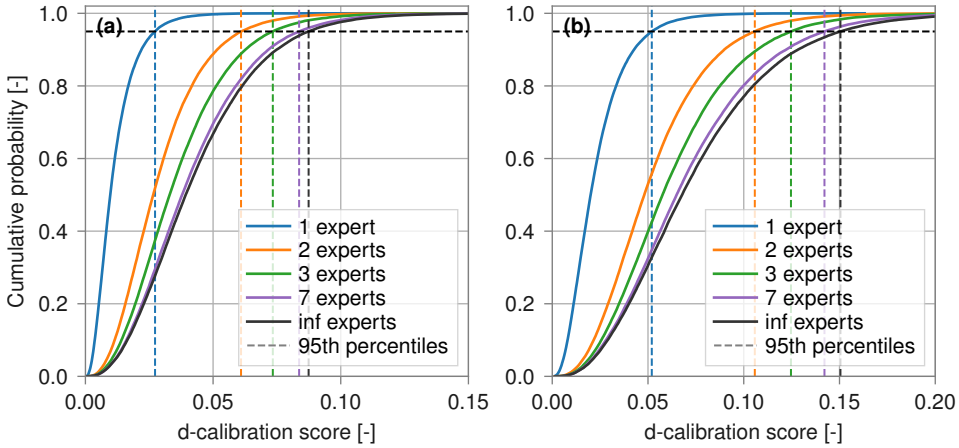


Figure 7.B.3: d-calibration scores of the average of experts' random matrices with conditional rank-correlations sampled from (a) $U(-1,1)$, or (b) $U(0,1)$, compared to a random 'true' matrix.

The consistent improvement in d-calibration score when averaging random matrices is similar to the pattern observed in Section 7.B.1, although when considering the averages (represented by the vertical dashed lines) the absolute differences are two to three times smaller for the random matrices.

The results from a random sampling exercise like this can be used to determine a significance level for the experts estimates. For example, the 95th percentile of the (converged) infinite expert solution. This means an expert needs to score higher than 95% of the uninformed (i.e., independence) guesses for random matrices. In the case presented in Fig. 7.B.3 (a), this is a d-calibration score of just under 0.09. However, this result varies for a) different number of random variables, b) methodological differences such as the sampling method used for drawing random matrices and c) the assumptions for the distribution from which the rank correlations are drawn. For example, in this case study,

it is only a small step from guessing completely uninformed to decide the rank correlations should be drawn from $U(0,1)$ instead of $U(-1,1)$ (no expert estimated a negative rank correlation coefficient during the elicitation). This would result in a significance level of 0.15, as shown in Fig. 7.B.3 (b). Note that this result is used as the 5% significance level for judging expert performance in Section 7.3.2. If an expert scores higher than this, the chance is less than 5% that the expert was making (almost) uninformed guesses.

7.B.3. Effect of adding a low-scoring expert

This study shows a relative insensitivity of the results to which individual experts are included in the DM. Section 7.3.2 showed that the DMs do not perform significantly better than the best experts, and Section 7.3.2 showed that the mean matrix of a pool scores, on average, better than the mean of the individual scores in the pool. This is partly due to all results being generally good, better than the significance level derived in last section. This section illustrates this by showing the effect of adding a low-scoring expert to the pool.

For this, we added a correlation matrix representing independence to the pool (i.e., all zeros, except for ones on the diagonal) and calculated the scores, displayed in Table 7.B.1. This gives a d-calibration score of 0.107, which is lower than the 5% significance level (0.150), and lower than the lowest scoring expert (A, 0.165). Including this estimate lowers the EQ DM score from 0.439 to 0.353, and the GL DM score from 0.437 to 0.403. The GL opt. is unaffected, as it still only uses the two highest scoring experts. In the original analysis, the EQ and GL DM had similar scores. Here we observe that the EQ is more affected than the GL, as the low score has a larger weight in the EQ DM. Note the relatively high log-likelihood for the 'low-scoring' expert. Compared to d-calibration score, an estimate with high correlations gives significantly worse log-likelihood than an estimate of independence (compare the scores of expert A and C to the 'low-scoring' expert).

Table 7.B.1: d-calibration scores and likelihood for experts' and DMs' correlation matrices, when a low-scoring expert is added to the pool.

	dependence score	calibration	Log-likelihood
Low-scoring expert	0.107		-1348.0
Other experts (A - G)	[0.165, 0.444]		[-2442.6, -933.7]
EQ DM	0.353		-971.0
GL DM	0.403		-945.4
GL opt. DM ($dCal > 0.411$)	0.468		-923.3
95% significance level (α)	0.150		

In the robustness analysis where are all combination of experts are calculated, the effect of the low-scoring expert becomes clearer. Figure 7.B.4 shows score for each combination of 1 to 8 (i.e., 7 + the hypothetical low-scoring expert) experts. For the EQ DM (panel

a), there are two clusters distinguishable, the bottom one including the low-scoring expert and the top one excluding it. For the GL DM (panel b) the influence of the low-scoring expert on the total score is smaller, because its weight is smaller.

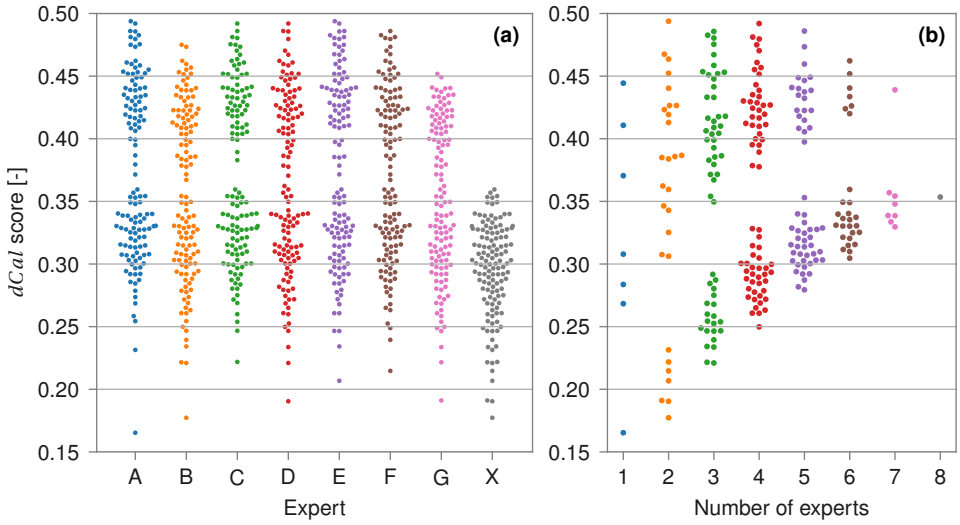


Figure 7.B.4: Score for each combination of 1 to 8 experts when a low-scoring expert is included. Equal weight DM above (a) and global weights DM below (b)

The same clustering can be observed in Fig. 7.B.5 where, similar to Fig. 7.B.2, the d-calibration scores of the mean matrices are compared to the means of the individual scores. Where in the previous analysis with the actual experts the scores were consistently better, there are now two clusters as well. On the upper right, the group without the low-scoring expert is still scoring consistently better. On the lower right, with the low-scoring expert, this effect is no longer present.

This analysis shows that a significance level, as well as global weight DM are important tools for reducing the potential impact of less accurate experts on the results. In this analysis we included a single expert. The effects will be greater when two or more experts score low. Note that the low-scoring expert had a d-calibration score of 0.107, while expert A, with a generally positive contribution to the DM pool, only had a slightly higher score of 0.165. This might give the false impression that the significance level (0.150) is a great cut-off level, while it is the high correlation estimates of expert A versus the independence estimates of the low-scoring expert that causes the difference in effect.

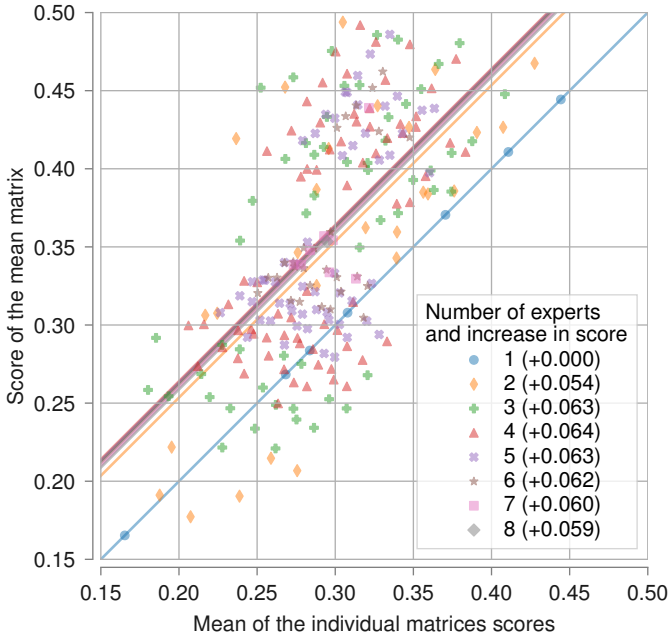


Figure 7.B.5: d-calibration scores of the mean matrix, compared to the average scores of the individual contributing matrices, when a low-scoring expert is included.

8

Conclusions and recommendations

8.1. Conclusions

The main goal of this research is assessing the value of structured expert judgment for Dutch flood risk assessment; by exploring in what way expert judgment can contribute to credible failure probabilities for engineered flood defense systems, in the context of rare non-experienced events and their probabilistic dependence. This research objective was addressed through four research questions:

1. What measure should be used to score experts, to obtain the best results in structured expert judgment studies following the Classical Model?
2. Which variables' uncertainties are most accurately estimated by experts?
3. How can expert judgment reduce uncertainty in the tails of probability distributions?
4. How do experts perform in estimating probabilistic dependence?

Each of these questions used a different approach to answer part of the objective. In the following paragraphs, the results from the individual chapters are presented in the context of these questions.

Through the case-studies, this study has demonstrated that structured expert judgment can indeed be used to obtain credible probabilities and quantify the dependence that plays an important role in flood risk assessments. However, the accuracy of these outcomes strongly depends on the types of uncertainties that are estimated, and the methods used to elicit or process them. The contributions in theoretical research and software development should help researchers and practitioners to apply structured expert judgment successfully in flood risk and related fields.

Physics-based models will always have a central role in a field that, eventually, is described by physical laws. However, as long as uncertainties around them remain unresolved, practitioners will require other methods to achieve credible outcomes. We hope to have inspired the reader of this thesis by the options that structured expert judgment can provide to aid in this goal.

What measure should be used to score experts, to obtain the best results in structured expert judgment studies following the Classical Model?

The standard approach in the Classical Model (also known as Cooke's Method) for structured expert judgment is to assess an expert's statistical accuracy by comparing expected and observed ratios of realizations within each quantile interval (see Section 2.1 for more background). These ratios are evaluated using the chi-square distribution, yielding a p-value that expresses the probability of an expert being statistically accurate. This value is combined with a second score to distinguish more informative from less informative experts but the expert weight is predominantly determined by the statistical accuracy.

The Classical Model focuses on estimating uncertainties. Instead of considering the distance between a point estimate and an outcome, it views the answer to a seed question as a realization from a probability distribution. Experts that correctly assesses this distribution and all other elicited distributions would find the realizations of all questions spread evenly throughout these distributions (in a very large sample). The conventionally chi-square distribution based test does not distinguish where an answer is within the quantile interval (e.g., where it is in between the 5th and 50th percentile). Other statistical tests, such as Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), Anderson-Darling (AD), and the Continuously Ranked Probability Score (CRPS) do make such a distinction, which is why these tests were considered as an alternative measure of statistical accuracy.

Chapter 4 showed that all five measures of statistical accuracy have different effects on the resulting decision maker (DM), such as the number of experts that are included with significant weight and the sensitivity to different biases. Weights calculated using one test were evaluated using the other tests to find out whether one of the tests calculates better weights than others. Apart from the fact that CRPS weights performed significantly worse, the comparison did not indicate a clear winner.

The five measures or scoring rules respond differently to overconfidence, the most predominant bias in structured expert judgment studies. An overconfident expert's outer percentile estimates (mostly the 5th and 95th) are too narrowly defined such that they do not enclose the realization (as you would expect with 90% probability for an accurate expert). CM, AD, and CRPS are more sensitive to overconfidence, while the statistical accuracies from KS and CvM are less affected. This means that choosing one of the first three concentrates the weight on mostly one or two experts, where KS or CvM will spread the weight over a more experts. While there are good reasons to choose a measure that is stricter on overconfidence (e.g., AD is considered an improvement of CvM for addressing the lack of sensitivity to tail realizations), an analyst may still choose a less sensitive measure of statistical accuracy, for example, to include more experts in the decision maker pool.

In addition to the different scoring rules, the probability distribution assumed around the estimated percentiles affects the scores as well. With the chi-square based test, this distribution is only used to calculate the informativeness. When using one of the alternative measures of statistical accuracy, it is also used to determine the position of the realization's quantile. The smooth Metalogistic distribution, a bell-shaped curve with more concentrated probability mass, was considered as an alternative to the stepped piecewise uniform distribution. The Metalog generally has lighter tails, meaning over- and underestimates are assigned a higher percentiles. This makes overconfidence-sensitive test like AD and CRPS behave more strictly when using the Metalog. While not specifically related to the scoring experts, the smooth Metalog seems more representative of a distribution that an expert would envisage. Removing percentiles and estimating their position showed that neither the piecewise uniform nor the Metalog performed well in estimating the location of the missing percentiles. The Metalog might look better, it does not give a substantially better (or worse) representation of expert estimates than the piecewise uniform distribution. Therefore, to increase the detail of expert estimates, more percentiles should be elicited.

Which variables' uncertainties are most accurately estimated by experts?

Expert judgment can be used to elicit estimates directly for the variable of interest, but it can also provide estimates of input parameters for a modeling approach. Both approaches were tested in a case study of Dutch river dikes to assess which of the two gives more credible dike failure probability estimates. The first approach involved eliciting the probability of (at least a single) dike failure in a year, i.e., the system-level failure probability. The second approach was more elaborate and involved scaling existing model results based on estimates of bias and uncertainty, which provides a calculated system-level failure probability as well.

The estimates for system level failure probability (the first approach) could quite easily be related to historic events. For example, 1995 was a major flood event on the Rhine River with a peak discharge of approximately 12,000 m^3/s. This information, which can be considered common knowledge amongst the experts, can serve as a referential value in the estimates. The second approach elicited failure for detailed dike cross sections. The experts were asked to estimate the water level at which a dike fails due to piping, slope instability, or overtopping. Historical reference values were not readily available, meaning the accuracy of the estimates was more dependent on the experts' technical knowledge. Consequently, a lack of knowledge about specific mechanisms caused some experts to give very wide estimates, assigning significant failure probability to frequently exceeded water levels, resulting in very high system failure probability. Overall, experts estimated credible probabilities of dike failure on a system level, while they struggled to accurately answer the questions concerning detailed dike sections.

With respect on the suitability of the Classical Model, the technical knowledge needed on several different subjects contributed to expert judgments resulting in infeasible result (e.g., potential stability failure at water levels below the dike's inner toe, or dike failure frequencies in the order of once per year). A format that is based on discussion, such as the DOT-guidelines (RWS-WVL, 2020) prescribe, might be more suitable for the specifics

of this case study. Structured expert judgment using the Classical Model can still play a role here, as it makes uncertainties explicit, but would be more suitable for estimating final, high-level, probabilities than replacing detailed model estimates.

How can expert judgment reduce uncertainty in the tails of probability distributions?

One of the difficulties with using expert judgment to estimate very small probabilities, is that these typically relate to events that have not been experienced by the expert. This is an issue for both data- and model-based approaches because the lack of extreme observations obstructs model fitting or calibration. In these cases, experts can use their knowledge from (when considering hydrological applications) different rivers, catchments, or events to inform their estimates. A natural approach of doing this is using Bayesian statistics. Without knowledge of the observed data, expert judgments can be considered prior information that can be updated with observations to obtain a posterior estimate.

This approach was applied in a case study for the river Meuse, in which the extreme value distributions for ten large tributaries were estimated. When fitting a distribution to data, the largest uncertainty is typically observed in the extrapolated range corresponding to the extreme events. To limit this, expert estimates were made for the (on average) one in 10- and one in 1000-year discharge, with the first used to evaluate expert performance, and the second being the variable of interest.

Using expert judgment as priors in this way, worked successfully in limiting the uncertainty in extremes. Moreover, applying *structured* expert judgment provides confidence in the used priors. Bayesian sampling techniques such as Markov Chain Monte Carlo provide great flexibility in including expert information, as long as the prior can be related to a model input (e.g., a distribution parameter) or a model output (e.g., the 1000-year discharge). As discussed in the second research question, the latter would likely give more reliable expert estimates.

Combining prior estimates of extremes with observations in the frequent range proved to be a good approach for limiting uncertainty in both the frequent and extreme range. Using only observations gave precise results (i.e., narrow uncertainty) exclusively in the frequent range, while using only expert estimates gave precise results exclusively in the extreme range.

How do experts perform in estimating probabilistic dependence?

Conditions that cause a risk of flooding often contain dependent elements. For example, a storm surge is dangerous due to its combination of high water levels, strong winds, and pounding waves. Similarly, high rainfall intensities might coincide with high sea levels due to an atmospheric depression. The complexity of the underlying phenomena makes the causal relationships between these factors difficult to simulate. Alternatively, such dependencies can be modeled using statistical models. This provides a natural way of treating the uncertainty in the causal relationships. However, whether it is successful relies on the accuracy of the dependence quantification.

Several statistical models exist to express dependence. Each has different properties regarding the possible number of dependent variables or the dependence pattern that is described. In the case study on the Meuse tributary discharges, the Gaussian copula was used to model the correlation between tributaries, which was quantified by the experts. This requires a correlation matrix in which each coefficient expresses the dependence between a tributary pair. Quantifying this is a complex task given the number of coefficients and the fact that not all combinations are valid. However, it can be simplified by using a Non-Parametric Bayesian Network (NPBN), a network of nodes representing the one dimensional marginal distributions and edges representing the correlation between them.

The dependence between the Meuse's tributaries was elicited by assigning each expert the task to draw and quantify a NPBN. Similar to the Classical Model, performance-based weights were used to combine different experts' results into a single expert estimate (i.e., the decision maker). The d-calibration score was used for this, which is a measure of statistical accuracy for dependence estimates. The experts performed well in estimating correlation coefficients, as everyone scored better than the 95% significance level of the d-calibration score. This threshold value was calculated from randomly generated networks, indicating the level above which the probability that an expert is 'randomly' assigning correlations, is less than 5%.

Compared to the weights for univariate estimates from the Classical Model, the d-calibration score's weights are much more evenly distributed. The weight-ratio between the highest and lowest scoring expert was only three, where a ratio of thirty million was observed with the univariate uncertainties. Moreover, we observed a limited sensitivity to which specific experts were including into the decision maker (both for equal and performance-based weights) and including more experts in the combination generally improved the result. This finding did however not hold after adding a low-scoring hypothetical expert to the pool.

How these results compare to other applications of dependence elicitation is a subject for further research. However, the NPBN presented itself convincingly as a suitable model to simplify the complex task of estimating statistical dependence.

8.2. Limitations and recommendations

The conclusions in this thesis are based on two case-studies related to the field of flood risk and additional research on the Classical Model and dependence elicitation. Each study comes with it caveats, as discussed in the individual discussion sections. It would be presumptuous to conclude that we now know how expert judgment works for flood risk, perhaps mostly because the quality of the results primarily depends on the preparation and execution of individual studies and not the field of study. Nonetheless, the case-studies extend a large body of research on structured expert judgment, which was used in this research and will be used in future research alike.

Detailed recommendations related to the individual research chapters were described there. Apart from those, a few suggestions for potentially impactful further research are:

1. One way to enhance the accuracy of Dutch flood defense safety assessments, is ensuring that the combination of all individual dike failure probabilities matches what is observed on a system-level: observed failures or the lack thereof. A calculated once per 10- or 100-year failure probability for an individual dike section seems high but could be accurate. However, if there are multiple of such dike sections, the total failure probability would quickly add up to a higher probability that related to events that should recently have been observed. A statistical model that describes the dependence between said sections (for example, like the model used in Chapter 5), can be used to decrease that potential 10- to 100-year probability to more credible levels, by using observations or the lack thereof on a system scale. Components of this model can be based on statistics, expert judgment, or physics, but should be aimed primarily at connecting the system probabilities to section probabilities. Through this, it could provide a framework to test the accuracy of the failure-mechanism models that are used in safety assessments.

2. As demonstrated in the case study for the river Meuse, expert judgment can effectively be used to reduce the uncertainty in the extrapolated range, where the predictive value of a model decreases. This concept makes the link between a model (which can either a statistical or physics-based model) and the expert's or modeler's judgment explicit, and should be explored further with models that are more comprehensive than the GEV-distribution applied in Chapter 6.

3. The non-parametric Bayesian Network is a relatively simple model to describe dependence between flood risk factors. The possibility of conditioning the Bayesian Network can be useful during floods or in preparatory emergency management. It provides a simple tool to assess the likelihood of "what-if?" scenarios. For example, what discharge can we expect on this tributary A, given the observed discharge on tributary B? Or, what are the chances of a high discharges in all these catchments?

4. The TU Delft structured expert judgment database contains many expert judgment studies and is of great use for developing the methodology of structured expert judgment (Cooke & Goossens, 2008; Cooke et al., 2021). A similar collection of dependence elicitation studies should be created and maintained, to aid development of methods to evaluate and combine expert's dependence estimates.

And finally, the last recommendation is to do (structured) expert judgment studies. In the first place because it is a valuable tool that can be used in addition to other approaches or when other approaches are not feasible. Secondly, in a field of work that has a central role for probability and uncertainty, practitioners should be skilled in assessing these. The combination of producing probability estimates without a predefined approach can be a true challenge as it forces the participant to combine models (simple or complex) with domain knowledge and common sense. Being evaluated based on the estimates gives valuable feedback on potential biases and provides the opportunity to discuss views and believes with other experts. The case studies showed that while participants found it challenging to produce estimates, they also found it a refreshing and enjoyable experience.

Acknowledgements

My PhD studies started in 2019. Between my graduation and then, Matthijs Kok had periodically reminded me that I still had plenty of years to work, so spending a few of the early ones in freedom on a PhD might be worth considering. At some point, Oswaldo had obtained a TKI grant to do research on "Elicitation of Multivariate Uncertainty for assessment of Flood Defense standards", which is a TKI EMU-FD project that funded this research, and to which Rijkswaterstaat, Deltares, and HKV contributed. Additional to the interesting title, which sounded puzzling at the time, the focus on expert judgment would be a nice change from all the model-based work I was doing on a day-to-day basis. Now, five years later, words like "Multivariate Uncertainty" and "Elicitation" are no longer puzzling but have become standard words in my (professional) jargon. I guess that is what a PhD does to you, but it might just be a spurious correlation.

First of all, I'd like to thank Oswaldo for his supervision. You are passionate, fun, and your enthusiasm is contagious. I left every catch-up inspired and with good spirit to continue the work, which is arguably the best trait for a supervisor. Matthijs, thank you as well for your guidance. You always see the larger context, and despite your reputation and respect you refrain from imposing your ideas upon others. Instead, you helped me, and others, to develop ideas and come to their own conclusions.

What certainly helped me along the way were the collaborations and discussions with Roger Cooke and Tina Nane, true experts of expert judgment. Thank you for teaching me the ins and outs of the subject and connecting me to the expert judgment community. I have greatly enjoyed working with you and it taught me a lot. Applied sciences also benefits from input from the field of practice. Therefore, I thank Bas Kolen, Annemargreet de Leeuw, Durk Riedstra, and Marieke Visser, who represented the contributing parties of the TKI-project and provided guidance on how to make the research project relevant and a contribution to the field of Flood Risk in the Netherlands.

The first prerequisite of an expert judgment study is to have experts. These studies need to be done consistently, which takes time and effort, also from the participants. Therefore, I would like to thank the experts that participated in the two studies. For the dike safety assessment: Carlijn Bus, Don de Bake, Henk van Hemert, Jan Blinde, Jan Tigchelaar, Jan-Kees Bossenbroek, Leo van Nieuwenhuijzen, Marinus Aalberts, Philippe Schoonen, R.B. Jongejan, Stefan van den Berg, Wim Kanning, and for the Meuse discharges: Eric Sprokkereef, Ferdinand Diermanse, Helena Pavelková, Jerom Aerts, Nicole Jungermann, and Siebolt Folkertsma. Without your willingness to donate some time and to let your uncertainty judgments be scrutinized by the Classical Model, the studies would not have been possible.

This PhD research was a part-time project, supported by HKV for whom I initially worked during the remainder of the work week. I am thankful to my former colleagues at this

wonderful company who supported me in this process. Firstly, by making the project and part-time combination possible but also with technical knowledge when things went outside my field of expertise, preparing the expert questionnaires, and last but not least, by being a fun distraction from the sometimes solitary PhD work. You are a talented and innovative bunch and create a great environment to bloom. Particular mention goes to Chris Geerse, whose ideas and innovations on probabilistic analysis have had great influence on me and our field of work. Unfortunately, you are no longer here with us, but I am sure your legacy will outlast us.

Halfway through 2022, I moved to New Zealand for a change of scenery and taste of the antipodean lifestyle. This move was a big step personally. It also meant quitting HKV and joining PDP in Christchurch New Zealand. The PhD studies I continued part-time. A big thanks to everyone who made this possible, accommodated it, and did not see the almost 20,000 kilometers as an obstruction to continue the work.

And last but not least, thanks to everyone who supported me personally throughout these years. In particular my family, for whom it might not be easy to have a son or brother living on the other side of the world, but who make me feel supported nonetheless. And most importantly, Enza, thanks for being in my life and taking on the big and small adventures with me, whatever crosses our path.

Bibliography

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., et al. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, *9*, e1516.

Adamczewski, T. (2023). The Metalog distribution. https://github.com/tadamcz/metalogistic

Al-Awadhi, S. A., & Garthwaite, P. H. (1998). An elicitation method for multivariate normal distributions. *Communications in Statistics-Theory and Methods*, *27*(5), 1123–1142.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, 193–212.

Armstrong, M. (1983). *Basic topology*. Springer-Verlag (New York).

Ayyub, B. M. (2000). Methods for expert-opinion elicitation of probabilities and consequences for corps facilities. *US Army Corps of Engineers, IWR Report No. 00-R-10*.

Bamber, J. L., Oppenheimer, M., Kopp, R. E., Aspinall, W. P., & Cooke, R. M. (2019). Ice sheet contributions to future sea-level rise from structured expert judgment. *Proceedings of the National Academy of Sciences*, *116*(23), 11195–11200.

Barons, M. J., Mascaro, S., & Hanea, A. M. (2022). Balancing the elicitation burden and the richness of expert input when quantifying discrete Bayesian networks. *Risk Analysis*, *42*(6), 1196–1234.

Benito, G., & Thorndycraft, V. (2005). Palaeoflood hydrology and its role in applied hydrological sciences. *Journal of Hydrology*, *313*(1-2), 3–15.

Bernard, A., & Bos-Levenbach, E. (1955). The plotting of observations on probability-paper. *Stichting Mathematisch Centrum. Statistische Afdeling*, (SP 30a/55).

Bouaziz, L. J., Thirel, G., de Boer-Euser, T., Melsen, L. A., Buitink, J., Brauer, C. C., De Niel, J., Moustakas, S., Willems, P., Grelier, B., et al. (2020). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences Discussions*, *2020*, 1–38.

Brázdil, R., Kundzewicz, Z. W., Benito, G., Demarée, G., Macdonald, N., & Roald, L. A. (2012). Historical floods in Europe in the past millennium. *Changes in Flood Risk in Europe, edited by: Kundzewicz, ZW, IAHS Press, Wallingford*, 121–166.

Brown, B. B. (1968). *Delphi process: a methodology used for the elicitation of opinions of experts* (tech. rep.). Rand Corp Santa Monica CA.

Brown, T. A. (1974). Admissible Scoring Systems for Continuous Distributions. *RAND Corporation*.

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., & Betancourt, M. (2015). The Stan math library: Reverse-mode automatic differentiation in C++. *arXiv preprint arXiv:1509.07164*.

Chbab, H. (2017). *Basisstochasten WBI-2017 - Statistiek en statistische onzekerheid* (tech. rep.). Deltares.

Chen, H.-P., & Mehrabani, M. B. (2019). Reliability analysis and optimum maintenance of coastal flood defences using probabilistic deterioration modelling. *Reliability Engineering & System Safety*, *185*, 163–174.

Clemen, R. T. (2008). Comments: Comment on Cooke's classical method. *Reliability Engineering and System Safety*, *93*(5), 760–765. https://doi.org/10.1016/j.ress.2008.02.003

Coles, S. G., & Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *45*(4), 463–478.

Colonna, K. J., Nane, G. F., Choma, E. F., Cooke, R. M., & Evans, J. S. (2022). A retrospective assessment of COVID-19 model performance in the USA. *Royal Society open science*, *9*(10), 220021.

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

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

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

Cooke, R. M. (2014). Validating expert judgment with the classical model. *Experts and consensus in social science*, 191–212.

Cooke, R. M., & Goossens, L. L. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety*, *93*(5), 657–674. https://doi.org/10.1016/j.ress.2007.03.005

Cooke, R. M., Marti, D., & Mazzuchi, T. (2021). Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *International Journal of Forecasting*, *37*(1), 378–387.

Cooke, R. M., & Slijkhuis, K. A. (2002). Expert judgment in the uncertainty analysis of dike ring failure frequency. *Case Studies in Reliability and Maintenance*, 331–350.

Copernicus Land Monitoring Service. (2017). EU-DEM [© European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA). Date accessed: 12 October 2021]. https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1/view

Copernicus Land Monitoring Service. (2018). CORINE Land Cover [© European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA). Date accessed: 16 March 2022]. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download

Copernicus Land Monitoring Service. (2020). E-OBS [© European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA). Date accessed: 19 May 2022]. %7Bhttps://cds.climate.copernicus.eu/cdsapp#!/dataset/insitu-gridded-observations-europe?tab=overview%7D

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, *1928*(1), 13–74.

CSöRgő, S., & Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-von Mises statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 221–234.

Curran, A., De Bruijn, K., Domeneghetti, A., Bianchi, F., Kok, M., Vorogushyn, S., & Castellarin, A. (2020). Large-scale stochastic flood hazard analysis applied to the Po River. *Natural Hazards*, *104*. https://doi.org/10.1007/s11069-020-04260-w

Daneshkhah, A., & Oakley, J. (2010). Eliciting multivariate probability distributions. *Rethinking risk measurement and reporting*, *1*, 1–43.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge university press.

De Niel, J., Demarée, G., & Willems, P. (2017). Weather Typing-Based Flood Frequency Analysis Verified for Exceptional Historical Events of Past 500 Years Along the Meuse River. *Water Resources Research*, *53*(10), 8459–8474. https://doi.org/10.1002/2017WR020803

de Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., Fenicia, F., Grelier, B., Nossent, J., Pereira, F., et al. (2017). Looking beyond general metrics for model comparison–lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, *21*(1), 423–440.

de Bruijn, K., van den Hurk, B., Slager, K., Rongen, G., Hegnauer, M., & van Heeringen, K. J. (2023). Storylines of the impacts in the Netherlands of alternative realizations of the Western Europe July 2021 floods. *Journal of Coastal and Riverine Flood Risk*, *2*, 8.

Delgado-Hernández, D.-J., Morales-Nápoles, O., De-León-Escobedo, D., & Arteaga-Arcos, J.-C. (2014). A continuous Bayesian network for earth dams' risk assessment: an application. *Structure and Infrastructure Engineering*, *10*(2), 225–238. https://doi.org/10.1080/15732479.2012.731416

Dewals, B., Erpicum, S., Pirotton, M., & Archambeau, P. (2021). Extreme floods in Belgium. The July 2021 extreme floods in the Belgian part of the Meuse basin. *Hydrolink*, 104–107. https://static.iahr.org/library/HydroLink/HL2021_4/Hydrolink_2021_4_Extreme_Flooding_Events.pdf

Dimitriadis, P., Koutsoyiannis, D., Iliopoulou, T., & Papanicolaou, P. (2021). A Global-Scale Investigation of Stochastic Similarities in Marginal Distribution and Dependence Structure of Key Hydrological-Cycle Processes. *Hydrology*, (2). https://doi.org/10.3390/hydrology8020059

Dion, P., Galbraith, N., & Sirag, E. (2020). Using expert elicitation to build long-term projection assumptions. In *Developments in demographic forecasting* (pp. 43–62). Springer, Cham. https://doi.org/10.1007/978-3-030-42472-5

Druzdel, M. J., & Van Der Gaag, L. C. (2000). Building probabilistic networks:" Where do the numbers come from?" *IEEE Transactions on knowledge and data engineering*, *12*(4), 481–486.

Dupuits, E., Klerk, W., Schweckendiek, T., & de Bruijn, K. (2019). Impact of including interdependencies between multiple riverine flood defences on the economically

optimal flood safety levels. *Reliability Engineering & System Safety*, *191*, 106475. https://doi.org/10.1016/j.ress.2019.04.028

Dupuits, E., Schweckendiek, T., & Kok, M. (2017). Economic optimization of coastal flood defense systems. *Reliability Engineering & System Safety*, *159*, 143–152.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's lassical model. *Reliability Engineering & System Safety*, *121*, 72–82.

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

Food and Agriculture Organization of the United Nations. (2003). Digital Soil Map of the World [Source: Land and Water Development Division, FAO, Rome. Date accessed: 20 June 2022]. https://data.apps.fao.org/map/catalog/srv/eng/catalog.search?id=14116#/metadata/446ed430-8383-11db-b9b2-000d939bc5d8

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. (2013). emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, *125*(925), 306. https://doi.org/10.1086/670067

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, *51*(9), 1417–1432.

French, S. (2023). Reflections on 50 years of MCDM: Issues and future research needs. *EURO Journal on Decision Processes*, *11*, 100030. https://doi.org/10.1016/j.ejdp.2023.100030

Garthwaite, P. H., & Al-Awadhi, S. A. (2001). Non-conjugate prior distribution assessment for multivariate normal sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(1), 95–110.

Geerse, C. (2011). *Hydra-Zoet for the fresh water systems in the Netherlands - Probabilistic model for the assessment of dike heights* (tech. rep.). HKV consultants.

Goodman, J., & Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, *5*(1), 65–80.

Grace, A. W., & Wood, I. A. (2012). Approximating the tail of the Anderson–Darling distribution. *Computational Statistics & Data Analysis*, *56*(12), 4301–4311.

Hanea, A., Morales Napoles, O., & Ababei, D. (2015). Non-parametric Bayesian networks: Improving theory and reviewing applications. *Reliability Engineering & System Safety*, *144*, 265–284. https://doi.org/10.1016/j.ress.2015.07.027

Hanea, A. M., Hilton, Z., Knight, B., & P. Robinson, A. (2022). Co-designing and building an expert-elicited non-parametric Bayesian network model: demonstrating a methodology using a Bonamia Ostreae spread risk case study. *Risk Analysis*, *42*(6), 1235–1254.

Hathout, M., Vuillet, M., Carvajal, C., Peyras, L., & Diab, Y. (2019). Expert judgments calibration and combination for assessment of river levee failure probability. *Reliability Engineering & System Safety*, *188*, 377–392. https://doi.org/10.1016/j.ress.2019.03.019

Hegnauer, M., Beersma, J., Van den Boogaard, H., Buishand, T., & Passchier, R. (2014). *Generator of Rainfall and Discharge Extremes (GRADE) for the Rhine and Meuse*

basins. Final report of GRADE 2.0 (tech. rep.) (Last access date: 21 June 2024). Deltares. Delft. https://publications.deltares.nl/1209424_004_0018.pdf

Hegnauer, M., & Van den Boogaard, H. (2016). *GPD verdeling in de GRADE onzekerheidsanalyse voor de Maas* (tech. rep.). Deltares. Delft.

Hokstada, P., Øien, K., & Reinertsen, R. (1998). Recommendations on the use of expert judgment in safety and reliability engineering studies. Two offshore case studies. *Reliability Engineering & System Safety*, *61*(1-2), 65–76.

ILT – Informatiehuis Water. (2024). Waterveiligheidsportaal - Landelijk Veiligheidsbeeld [Online; accessed 4 October 2024]. https://waterveiligheidsportaal.nl/nss/prognosis

IPCC. (2023). IPCC, 2023: Summary for Policymakers. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland., 1–34. https://doi.org/10.59327/IPCC/AR6-9789291691647.001

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, *81*(348), 158–171. https://doi.org/10.1002/qj.49708134804

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of multivariate analysis*, *97*(10), 2177–2189.

Jongejan, R., & Maaskant, B. (2015). Quantifying flood risks in the Netherlands. *Risk Analysis*, *35*(2), 252–264.

Jongejan, R. B., Diermanse, F., Kanning, W., & Bottema, M. (2020). Reliability-based partial factors for flood defenses. *Reliability Engineering & System Safety*, *193*, 106589.

Jonkman, S. N., Jongejan, R., & Maaskant, B. (2011). The use of individual and societal risk criteria within the Dutch flood safety policy—Nationwide estimates of societal risk and policy applications. *Risk Analysis: An International Journal*, *31*(2), 282–300.

Keelin, T. W. (2016). The metalog distributions. *Decision Analysis*, *13*(4), 243–277.

Kind, J. M. (2014). Economically efficient flood protection standards for the Netherlands. *Journal of Flood Risk Management*, *7*(2), 103–117.

Kindermann, P. E., Brouwer, W. S., van Hamel, A., van Haren, M., Verboeket, R. P., Nane, G. F., Lakhe, H., Prajapati, R., & Davids, J. C. (2020). Return level analysis of the hanumante river using structured expert judgment: a reconstruction of historical water levels. *Water*, *12*(11), 3229. https://doi.org/10.3390/w12113229

Klerk, W., Kok, M., De Bruijn, K., Jonkman, S., & Van Overloop, P. (2014). Influence of load interdependencies of flood defences on probabilities and risks at the Bovenrijn/IJssel area, The Netherlands. *Proceeding of the 6th international conference on flood management–ICFM6*, 1–13.

Knoeff, H., Sellmeijer, J., Lopez, J., & Luijendijk, S. (2009). *SBW Piping - Hervalidatie piping* (tech. rep.). Deltares. Delft.

Kok, M., Jongejan, R., Nieuwjaar, M., & Tanczos, I. (2017). *Fundamentals of flood protection* (tech. rep.). Expertise Netwerk Waterveiligheid. Ministerie van Infrastructuur en Milieu.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, *4*, 89–91.

Koot, P., Mendoza-Lugo, M. A., Paprotny, D., Morales-Nápoles, O., Ragno, E., & Worm, D. T. (2023). PyBanshee version (1.0): A Python implementation of the MATLAB toolbox BANSHEE for Non-Parametric Bayesian Networks with updated features. *SoftwareX*, *21*, 101279. https://doi.org/10.1016/j.softx.2022.101279

Koutsoyiannis, D. (2004a). Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation / Statistiques de valeurs extrêmes et estimation de précipitations extrêmes: I. Recherche théorique. *Hydrological Sciences Journal*, *49*(4), 575–590. https://doi.org/10.1623/hysj.49.4.575.54430

Koutsoyiannis, D. (2004b). Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records / Statistiques de valeurs extrêmes et estimation de précipitations extrêmes: II. Recherche empirique sur de longues séries de précipitations. *Hydrological Sciences Journal*, *49*(4). https://doi.org/10.1623/hysj.49.4.591.54424

Land NRW. (2022). ELWAS-WEB [Last date accessed: 21 June 2024]. https://www.elwasweb.nrw.de

Leander, R., Buishand, A., Aalders, P., & Wit, M. D. (2005). Estimation of extreme floods of the River Meuse using a stochastic weather generator and a rainfall. *Hydrological Sciences Journal*, *50*(6), 1089–1103. https://doi.org/10.1623/hysj.2005.50.6.1089

Leontaris, G., & Morales-Nápoles, O. (2018). ANDURIL: A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments. *SoftwareX*, *7*, 313–317. https://doi.org/10.1016/j.softx.2018.07.001

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Marsaglia, G., & Marsaglia, J. (2004). Evaluating the anderson-darling distribution. *Journal of statistical software*, *9*, 1–5.

Marti, D., Mazzuchi, T. A., & Cooke, R. M. (2021). Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis. *Expert Judgement in Risk and Decision Analysis*, *293*, 53–82. https://doi.org/10.1007/978-3-030-46474-5_3

Martins, E. S., & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*(3), 737–744.

Marvin, M., & Henryk, M. (1992). *A survey of matrix theory and matrix inequalities* (Vol. 14). Courier Dover Publications.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Ministry of Infrastructure and Environment. (2016). Regeling veiligheid primaire waterkeringen 2017 no IENM/BSK-2016/283517 [https://wetten.overheid.nl/BWBR0039040/2017-01-01].

Moala, F. A., & O'Hagan, A. (2010). Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference*, *140*(7), 1635–1655. https://doi.org/10.1016/j.jspi.2010.01.004

Mohr, S., Ehret, U., Kunz, M., Ludwig, P., Caldas-Alvarez, A., Daniell, J. E., Ehmele, F., Feldmann, H., Franca, M. J., Gattke, C., et al. (2022). A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe. Part 1: Event description and analysis. *Natural Hazards and Earth System Sciences Discussions*, 1–44.

Morales, O., Kurowicka, D., & Roelen, A. (2008). Eliciting conditional and unconditional rank correlations from conditional probabilities [Expert Judgement]. *Reliability Engineering & System Safety*, *93*(5), 699–710. https://doi.org/10.1016/j.ress.2007.03.020

Morales Nápoles, O., & Worm, D. (2013). *Hypothesis testing of multidimensional probability distributions*.

Morales-Nápoles, O., Hanea, A., & Worm, D. Experimental results about the assessments of conditional rank correlations by experts: Example with air pollution estimates. In: 2014, 1359–1366.

Morales-Nápoles, O., Delgado-Hernández, D. J., De-León-Escobedo, D., & Arteaga-Arcos, J. C. (2014). A continuous Bayesian network for earth dams' risk assessment: methodology and quantification. *Structure and Infrastructure Engineering*, *10*(5), 589–603. https://doi.org/10.1080/15732479.2012.757789

Moustafa, A., Karim, T., De La Torre, F., & Ferrie, F. (2010). *Designing a Metric for the Difference between Gaussian Densities* (J. Angeles, B. Boulet, J. Clark, J. Kövecses, & K. Siddiqi, Eds.; Vol. 83). Springer.

Nane, G. F., & Cooke, R. M. (2024). Scoring rules and performance, new analysis of expert judgment data. *Futures & Foresight Science*, e189.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Nofal, O. M., van de Lindt, J. W., & Do, T. Q. (2020). Multi-variate and single-variable flood fragility and loss approaches for buildings. *Reliability Engineering & System Safety*, *202*, 106971. https://doi.org/10.1016/j.ress.2020.106971

Nyberg, E. P., Nicholson, A. E., Korb, K. B., Wybrow, M., Zukerman, I., Mascaro, S., Thakur, S., Oshni Alvandi, A., Riley, J., Pearson, R., et al. (2022). BARD: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *Risk Analysis*, *42*(6), 1155–1178.

Oppenheimer, M., Little, C. M., & Cooke, R. M. (2016). Expert judgement and uncertainty quantification for climate change. *Nature climate change*, *6*(5), 445–451.

Papalexiou, S. M., & Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, *49*(1), 187–201.

Paprotny, D., & Morales-Nápoles, O. (2017). Estimating extreme river discharges in Europe through a Bayesian network. *Hydrology and Earth System Sciences*, *21*(6), 2615–2636.

Paprotny, D., Morales-Nápoles, O., Worm, D. T., & Ragno, E. (2020). BANSHEE–A matlab toolbox for non-parametric bayesian networks. *SoftwareX*, *12*, 100588.

Parent, E., & Bernier, J. (2003). Encoding prior experts judgments to improve risk analysis of extreme hydrological events via POT modeling. *Journal of Hydrology*, *283*(1-4), 1–18.

Pearl, J. (2000). Causality: models, inference and reasoning.

Projectbureau VNK2. (2010). *De veiligheid van Nederland in kaart* (tech. rep. No. November). Rijkswaterstaat Projectbureau VNK.

Ragno, E., Hrachowitz, M., & Morales-Nápoles, O. (2022). Applying non-parametric Bayesian networks to estimate maximum daily river discharge: potential and challenges. *Hydrology and Earth System Sciences*, *26*(6), 1695–1711.

Ramousse, B., Mendoza-Lugo, M. A., Rongen, G., & Morales-Nápoles, O. (2024). Elicitation of Rank Correlations with Probabilities of Concordance: Method and Application to Building Management. *Entropy*, *26*(5), 360.

Ren, X., Nane, G. F., Terwel, K. C., & van Gelder, P. H. (2024). Measuring the impacts of human and organizational factors on human errors in the Dutch construction industry using structured expert judgement. *Reliability Engineering & System Safety*, *244*, 109959. https://doi.org/10.1016/j.ress.2024.109959

Renard, B., Lang, M., & Bois, P. (2006). Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data. *Stochastic environmental research and risk assessment*, *21*(2), 97–112.

Renooij, S. (2001). Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review*, *16*(3), 255–269. https://doi.org/10.1017/S0269888901000145

Rentschler, J., Salhab, M., & Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. *Nature communications*, *13*(1), 3527.

Rijkswaterstaat. (2022). Waterinfo [Date accessed: 11 March 2022]. https://waterinfo.rws.nl/#!/kaart/Afvoer/Debiet___20Oppervlaktewater___20m3___2Fs/

Rongen, G., Morales-Nápoles, O., & Kok, M. (2022a). Expert judgment-based reliability analysis of the Dutch flood defense system. *Reliability Engineering & System Safety*, *224*, 108535.

Rongen, G., Morales-Nápoles, O., & Kok, M. (2022b, August). Extreme Discharge Uncertainty Estimates for the River Meuse Using a Hierarchical Non-Parametric Bayesian Network. In M. C. Leva, E. Patelli, L. Podofillini, & S. Wilson (Eds.), *Proceedings of the 32th european safety and reliability conference (esrel 2022)* (pp. 2670–2677). Research Publishing. https://doi.org/10.3850/978-981-18-5183-4_S17-04-622-cd

Rongen, G., Morales-Nápoles, O., & Kok, M. (2024). Using the Classical Model for structured expert judgment to estimate extremes: a case study of discharges in the Meuse River. *Hydrology and Earth System Sciences*, *28*(13), 2831–2848. https://doi.org/10.5194/egusphere-2023-39

Rongen, G., & Morales-Nápoles, O. (2024). Matlatzinca: A PyBANSHEE-based graphical user interface for elicitation of non-parametric Bayesian networks from experts. *SoftwareX*, *26*, 101693. https://doi.org/10.1016/j.softx.2024.101693

Rongen, G., 't Hart, C. M. P., Leontaris, G., & Morales-Nápoles, O. (2020). Update (1.2) to ANDURIL and ANDURYL: Performance improvements and a graphical user interface. *SoftwareX*, *12*, 100497. https://doi.org/10.1016/j.softx.2020.100497

Rongen, G. (2016). *The effect of flooding along the Belgian Meuse on the discharge and hydrograph shape at Eijsden* [Master's thesis, Delft University of Technology].

RWS-WVL. (2020, July). *Handreiking DOT, Deskundigen Oordeel voor de Toets op maat – Groene versie* (tech. rep.). Rijkswaterstaat WVL, Unie van Waterschappen, Deltares.

Sayers, P., & Meadowcroft, I. (2005). RASP-A hierarchy of risk-based methods and their application. *Defra Flood and Coastal Management Conference 2005*, 1–18.

Sebok, E., Henriksen, H. J., Pastén-Zapata, E., Berg, P., Thirel, G., Lemoine, A., Lira-Loarca, A., Photiadou, C., Pimentel, R., Royer-Gaspard, P., et al. (2021). Use of expert elicitation to assign weights to climate and hydrological models in climate impact studies. *Hydrology and Earth System Sciences Discussions*, 1–35.

Sellmeijer, J. (1988). *On the mechanism of piping under impervious structures* [Doctoral dissertation, Technische Universiteit Delft]. http://repository.tudelft.nl/assets/uuid:7f3c5919-1b37-4de9-a552-1f6e900eeaad/TR%20diss%201670(1).pdf

Service public de Wallonie. (2022). Voies Hydraulique Wallonie - Annuaires et statistiques [Date accessed: 26 June 2022]. http://voies-hydrauliques.wallonie.be/opencms/opencms/fr/hydro/Archive/annuaires/index.html

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Simard, R., & L'Ecuyer, P. (2011). Computing the two-sided Kolmogorov-Smirnov distribution. *Journal of Statistical Software*, *39*, 1–18.

Slijkhuis, K., Frijters, M., Cooke, R., & Vrouwenvelder, A. (1998). Probability of flooding: An uncertainty analysis. *Lydersen, S. Hansen, GK Sandtorv, HA, Safety and Reliability. Proceedings of the European Conference on Safety and Relibility-ESREL'98, Trondheim, Norway, 16-19 June, 1419-1425*.

Slomp, R., Knoeff, H., Bizzarri, A., Bottema, M., & De Vries, W. (2016). Probabilistic Flood Defence Assessment Tools. *E3S Web of Conferences*, *7*, 1–14. https://doi.org/10.1051/e3sconf/20160703015

Smirnoff, N. (1939). Sur les écarts de la courbe de distribution empirique. *Matematicheskii Sbornik*, *48*(1), 3–26.

't Hart, C. M. P., Leontaris, G., & Morales-Nápoles, O. (2019). Update (1.1) to ANDURIL — A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments: ANDURYL. *SoftwareX*, *10*, 100295. https://doi.org/10.1016/j.softx.2019.100295

't Hart, R., de Bruijn, H., & de Vries, G. (2016). *Fenomenologische beschrijving* (tech. rep.). Deltares.

TFFF. (2021). *Hoogwater 2021 - Feiten en duiding* (tech. rep.) (Last access: 21 June 2024). Task Force Fact-finding hoogwater 2021. Delft, Expertisenetwerk Waterveiligheid (ENW). https://www.enwinfo.nl/publish/pages/183541/211102_enw_hoogwater_2021-dv-def.pdf

Torres-Alves, G. A., & Morales-Nápoles, O. (2020). Reliability analysis of flood defenses: The case of the Nezahualcoyotl dike in the aztec city of Tenochtitlan. *Reliability Engineering & System Safety*, *203*, 107057. https://doi.org/10.1016/j.ress.2020.107057

Uemura, F., Rongen, G., Masuya, S., Yoshida, T., & Yamada, T. J. (2024). Calculating flood probability in Obihiro using a probabilistic method: incorporating the probability of dike failure with uncertainty. *Proceedings of IAHS*, *386*, 69–74.

van de Langemheen, W., & Berger, H. (2001). *Hydraulische randvoorwaarden 2001: Maatgevende afvoeren rijn en maas* (tech. rep.). RIZA. Ministerie van Verkeer en Waterstaat.

Van der Meer, J., Ter Horst, W., & Van Velzen, E. (2009). Calculation of fragility curves for flood defence assets. *Flood Risk Management: Research and Practice*, 567–573.

Van der Meij, R. (2019). *D-Stability - User Manual* (tech. rep.). Deltares.

van Hoven, A. (2019, November). *Schematiseringshandleiding grasbekleding* (tech. rep. No. november). Deltares.

Viglione, A., Merz, R., Salinas, J. L., & Blöschl, G. (2013). Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, *49*(2), 675–692.

Von Mises, R. (1928). Statistik und wahrheit. *Julius Springer*, *20*.

Vrijling, J. (2001). Probabilistic design of water defense systems in The Netherlands. *Reliability Engineering & System Safety*, *74*(3), 337–344. https://doi.org/10.1016/S0951-8320(01)00082-5

Waterschap Limburg. (2021). Discharge Measurements [Source: Waterschap Limburg (https://www.waterstandlimburg.nl/Home) Historical time series from personal communication. Date retrieved: 16 August 2021].

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., & Morales-Napoles, O. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. *European Journal of Operational Research*, *258*(3), 801–819.

Curriculum Vitæ

Guus Willem Franciscus Rongen

09-11-1990	Born in Geleen, the Netherlands.

Education

2003–2009	Pre-University Secondary Education VWO, Science and Engineering Bonnefanten College, Maastricht
2009–2013	BSc. Civil Engineering Delft University of Technology, Delft
2013–2016	MSc. Civil Engineering (Spec. Hydraulic Engineering) Delft University of Technology, Delft
2016-2022	Flood Risk advisor HKV consultants, Lelystad and Delft
2019-2023	PhD Researcher Section of Hydraulic Structures and Flood Risk Faculty of Civil Engineering and Geosciences Delft University of Technology, Delft
2022-present	Environmental Engineer Pattle Delamore Partners Ltd., Christchurch, New Zealand

Awards

2023	Stein Hugen Award ESREL 2023 (Best article from PhD entries)

List of Publications

Peer reviewed journal papers

8. **G Rongen**, GF Nane, O Morales-Nápoles, R Cooke, *Continuous distributions and measures of statistical accuracy for structured expert judgment* (Submitted)

7. **G Rongen**, O Morales-Nápoles, D Worm, M Kok, *Structured expert elicitation of dependence between river tributaries using non-parametric Bayesian networks* (Under review)

6. **G Rongen**, O Morales-Nápoles, M Kok, *"Using the Classical Model for structured expert judgment to estimate extremes: a case study of discharges in the Meuse River"*, Hydrology and Earth System Sciences 28(12), 2831-2848 (2024)

5. B Ramousse, MA Mendoze-Lugo, **G Rongen**, O Morales-Nápoles, *"Elicitation of Rank Correlations with Probabilities of Concordance: Method and Application to Building Management"*, Entropy 26 (2024)

4. **G Rongen**, O Morales-Nápoles, *"Matlatzinca: A PyBANSHEE-based graphical user interface for elicitation of non-parametric Bayesian networks from experts"*, SoftwareX 26 (2024)

3. KM de Bruijn, B van den Hurk, K Slager, **G Rongen**, M Hegnauer, KJ van Heeringen, *"Storylines of the impacts in the Netherlands of alternative realizations of the Western Europe July 2021 floods"*, Journal of Coastal and Riverine Flood Risk 2 (2023)

2. **G Rongen**, O Morales-Nápoles, M Kok, *"Expert judgment-based reliability analysis of the Dutch flood defense system"*, Reliability Engineering & System Safety 224 (2022)

1. **G Rongen**, M 't Hart, G Leontaris, O Morales-Nápoles, *"Update (1.2) to ANDURIL and ANDURYL: Performance improvements and a graphical user interface"*, SoftwareX 12 (2020)

Conference papers

3. F Uemura, **G Rongen**, S Masuya, T Yoshida, TJ Yamada, *Calculating flood probability in Obihiro using a probabilistic method: incorporating the probability of dike failure with uncertainty*, Proceedings of IAHS (2024)

2. **G Rongen**, B Throssell, *Schematizing rainfall events with multivariate depth-duration dependence*, Proceedings of the 33rd European Safety and Reliability Conference (2023)

1. **G Rongen**, O Morales-Nápoles, M Kok, *"Extreme Discharge Uncertainty Estimates for the River Meuse Using a Hierarchical Non-Parametric Bayesian Network"*, Proceedings of the 32rd European Safety and Reliability Conference (2022)