Structured Expert Elicitation of Dependence Between River Tributaries Using Nonparametric Bayesian Networks

Rongen, Guus; Morales-Nápoles, Oswaldo; Worm, Daniël; Kok, Matthijs

DOI
10.1111/risa.70111

Publication date
2025

Document Version
Final published version

Published in
Risk Analysis

Risk Analysis

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Structured Expert Elicitation of Dependence Between River Tributaries Using Nonparametric Bayesian Networks

Guus Rongen[1] | Oswaldo Morales-Nápoles[1] | Daniël Worm[2] | Matthijs Kok[1]

[1]Delft University of Technology, Civil Engineering and Geosciences, Delft, the Netherlands | [2]TNO, Applied Cryptography & Quantum Algorithms, Delft, the Netherlands

Correspondence: Guus Rongen (g.w.f.rongen@tudelft.nl)

Received: 18 April 2023 | **Revised:** 4 May 2025 | **Accepted:** 14 August 2025

Funding: This research was funded by the TKI project EMU-FD. This research project is funded by Rijkswaterstaat, Deltares and HKV consultants.

Keywords: dependence elicitation | structured expert judgment | flood risk | nonparametric Bayesian network | d-calibration score

ABSTRACT

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 nonparametric Bayesian network, describing the correlation between river tributary discharges. Specialized software was provided to assist the experts. 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 were investigated theoretically. Individual expert judgments were combined based on performance into a group opinion aka decision maker. All experts were able to create and quantify a correlation matrix between 10 variables that resembled the correlations between observed discharges 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. Adding a bad performing expert increased the positive effect of performance-based weighting, underscoring the importance of developing scoring rules for dependence elicitation. The overall results are promising: Aided by specialized graphical software, the experts in this study were able to quickly create and quantify dependence structures.

1 | Introduction

Scientific models can involve substantial uncertainty, especially when used to predict unprecedented events. In absence of data or resources to quantify these uncertainties, for example, because of the unfeasibility of large experiments or data collection, 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, just like univariate uncertainties.

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 (Brown 1968) and the Classical Model (CM), also known as Cooke's method (Cooke 1991). Most expert judgments studies in science and engineering focus 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

https://doi.org/10.1111/risa.70111

the assessment of the relationships between these distributions. Consequently, 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. With 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 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 nonparametric Bayesian networks (NPBNs) (i.e., a specific form of a BN) by experts may be found in (Delgado-Hernández et al. 2014 "and" 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). For a comprehensive overview of dependence models 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 (DMs) 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 NPBN. They first estimated the tributary discharges (marginals) and then their correlations. Combined, they were 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 and Worm 2013). These were then used as weights to create 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 was calculated to indicate whether an expert's estimate is significantly better than an uninformed guess. Finally, we showed theoretical properties of the dependence-calibration (or d-calibrations) score as a desirable metric of expert performance when eliciting dependence. The methods for estimating the marginals, as well as the Meuse River discharge statistics resulting from the elicited marginals and dependencies, are described in Rongen et al. (2024).

2 | Background on BNs and Copulas

This study quantifies dependence using NPBNs which are based on Gaussian (Normal) copulas. This section provides some background on these concepts.

A BN (Darwiche 2009; Pearl 2000) is a directed acyclic graph (DAG) that represents the dependence between random variables through nodes and arcs (see Figure 2 for some examples of DAGs). In a BN, the used probability distributions for the nodes are generally discrete and the conditional probability functions to be estimated are conditional probability tables. As an example, consider a hypothetical BN that describes the dependence between X_1 having COVID and X_2 testing positive to it. Assume both random variables have two states. Then the BN given by $X_1 \rightarrow X_2$ would render four conditional probabilities to be quantified (i.e., having COVID conditional on testing positive, having COVID conditional on testing negative, not having COVID conditional on testing positive, and not having COVID conditional on testing negative). These probabilities are conditional because having COVID changes the probability of testing positive and, reversely, testing positive changes the probability that someone has COVID (i.e., it makes it more likely if the test is any good).

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 with arcs 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 and Van Der Gaag 2000; Renooij 2001) and increases rapidly with the number of states of the variables in the network. Continuous variables can be discretized into several states. A finer discretization gives a better representation but simultaneously requires a larger number of conditional probabilities to be assessed.

The NPBN is a special form of a BN that uses Gaussian copulas to describe the relationships between variables. Each arc in a NPBN represents a (conditional) rank correlation. The structure of the graph defines which child node is dependent on which parent node, and through that the conditional (in)dependence between nodes. The (conditional) rank correlations, in combination with the graphical structure, give a positive semi-definite correlation matrix, that is, a unique and valid correlation matrix. Although NPBNs are based on Gaussian copulas, the marginal distributions of the random variables do not need to be normally distributed. The MVN can be transformed to its percentiles in the $[0, 1]$ range using the cumulative distribution function and subsequently be transformed to any desired distribution using its percentile function, facilitating calculations such as computing conditional distributions analytically. This absence of a need to parametrize 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 NPBN. For a more formal and detailed explanation of NPBNs, as well as a description of some applications, refer to

TABLE 1 | List of experts with their affiliation and professional interests.

Name	Affiliation	Field of expertise
Alexander Bakker	Rijkswaterstaat & Delft University of Technology	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

(A. Hanea et al. 2015). For this study, the most relevant feature about NPBNs is that it may be characterized by a rank correlation matrix, which is used to quantify the dependence in a multivariate normal copula.

A copula is a multivariate cumulative distribution for which the marginals are uniform in the interval $[0,1]$. Transforming the aforementioned MVN distribution to its percentiles gives a Gaussian copula, which can be used to model different correlation strengths between different variable pairs. The Gaussian copula is however limited in its ability to differentiate dependence strength between parts of the distributions (e.g., it does not model asymmetries in the joint distribution such as tail dependence). Archimedean copulas on the contrary can take many different forms and describe asymmetries in joint distributions such as tail dependence. However, their ability to describe dependence between more than two variables is limited as they fail, for example, to model different dependence patterns for different pairs of margins of the (three or more variable) joint distribution. In this study, we therefore use the Gaussian copula as dependence model to represent the elicited correlation matrices.

3 | Methods

In this study, experts estimated dependence between discharge peaks of tributaries within a catchment by quantifying a NPBN network using supporting software. The method description includes the experts and elicitation process (Section 3.1), the discharge data used (Section 3.2), and the method for scoring the experts' estimates (Section 3.3).

3.1 | 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 (Rongen et al. 2024). Seven experts participated in the elicitation that took place on July 4, 2022. An overview of their names, affiliations, field of expertise is shown in Table 1. All experts have a hydrology or flood risk background, and work in academia, governmental organizations, research institutes, or consultancy. Note that the experts are listed alphabetically by their first names, and that the listed order holds no reference to the letters (A–G) used when presenting the results.During this session, the experts estimated the discharge that is exceeded on average once per 10 and 1000 years. These estimates were then combined with data to form extreme value distributions (Rongen et al. 2024). For calculating extreme discharges along the downstream parts of the Meuse, the statistical dependencies between tributary discharges were also elicited. The dependence results are presented in this article.

Participants were tasked with estimating a correlation matrix representing the dependence between 10 tributaries. A NPBN was deemed an appropriate tool for this task, for three reasons: First, experts can intuitively consider a "causal" structure when specifying correlations. Second, 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 a value for each pair of nodes (see Section 2). Instead, 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. That is because, as stated previously, the elements of the correlation matrix are not algebraically independent. Assigning a number in $[-1,1]$ to every element of the matrix will not guarantee that the remaining matrix will be a correlation matrix since a correlation matrix has to be symmetric and positive semi-definite. The simplest example demonstrating 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. Using the NPBN prevents possible inconsistencies with this. Through the assignment of *conditional* rank correlations to the arcs of a NPBN (as described in A. Hanea et al. 2015), the constraints required for the resulting matrix to be a correlation matrix will always be met, as they describe the strength of the remaining correlation between two variables while holding other variables constant. Consequently, the conditional rank correlation can take any value in $[-1, 1]$. See (Morales et al. 2008) for more background on conditional correlations.

To assist the experts in creating the NPBN, we developed a GUI-based program called Matlatzinca. This program, based on (Paprotny et al. 2020; Koot et al. 2023), 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 above example 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, figs. 3 and 4), which are intended to clarify the relationship between correlation coefficients and conditional probabilities.

In structured expert judgment, seed questions are used to determine expert performance. Performance-based weights are derived by comparing the expert's estimates to the questions realizations (answers), which are then used to obtain the answers for the (unknown) target variables. In this study, experts estimated a "known" correlation matrix, in the sense that it was calculated from observations. This enables testing the experts' performance. We did not separately define tail-dependence for the correlations (i.e., different dependence for the extremes) since the used dependence model (Gaussian copula) does not facilitate the possibility to model these in detail.

3.2 | 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 the events' magnitudes) than the absolute values.

Figure 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).

The time ranges for which data are available differ between different stations. Creating a valid correlation matrix requires complete records, which is why we excluded 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 as well would further increase the number of events to 111 but we considered it to be a more significant tributary, that is, not worth excluding for 5 extra events. After excluding the Sambre, 9 of the 10 elicited tributaries remain in the correlation matrices.

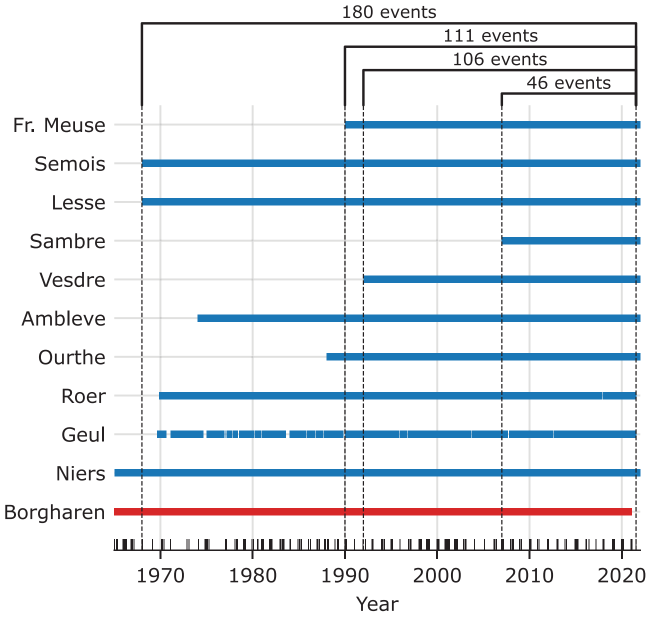


FIGURE 1 | Horizontal bars indicating the availability of measured discharge records for the different tributaries (blue) and the main river at Borgharen (red).

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 nonparametric 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.

3.3 | Scoring the Experts' Performance

We applied performance based weighting to combine the different experts' estimates into a DM. For the (univariate) tributary discharges, we combined the estimates using the CM (Cooke and Goossens 2008). The underlying idea is that a (performance based) weighting of expert estimates gives a better estimate than a single expert or an equally weighted combination. Continuing on this assumption necessitates a different score to assess expert performance, because the CM is not suitable for scoring dependence. We used the d-calibration score (Morales Nápoles and Worm 2013; Morales-Nápoles, Hanea, & Worm et al. 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:

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

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:

$$dCal(e) = 1 - d_H(R_q, R_e). \tag{2}$$

This score consequently varies on a scale from 0 to 1 and can be used as weights (after normalization) to calculate DMs similar to the CM. In Equation (1), 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 Appendix 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 (Appendix 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:

$$f(\mathbf{q}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{q}-\mu)^T \Sigma^{-1} (\mathbf{q}-\mu)\right). \tag{3}$$

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:

$$\ell(R|\mathbf{x}) = \log\left(\frac{1}{\sqrt{(2\pi)^k |R|}}\right) - \frac{1}{2}\mathbf{x}^T R^{-1} \mathbf{x}. \tag{4}$$

Note that this evaluates the joint probability distribution by its likeliness to the Gaussian copula (the MVN's cumulative distribution function). By transforming the observed discharges through their ranks, no assumption is made for their marginal distribution.

The log-likelihood is not a probability and does not range from 0 to 1. With a nine-variable MVN-distribution, the likelihoods are generally 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.

4 | Results

4.1 | BNs and Correlation Matrices

The BNs quantified by individual experts are shown in Figure 2. Recall 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 nonconditional correlations matches their estimate. Experts C and 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. A full overview of the (conditional) rank correlations specified by the experts is shown in Tables C1–C7.

The BNs in Figure 2, together with the experts' assessments of (conditional) rank correlations, give the correlation matrices shown in Figure 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 Experts 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.

4.2 | Experts' and DMs' Performance

4.2.1 | Scores

Table 2 shows the d-calibration scores (higher is better) and log-likelihoods (less negative is better) calculated from the expert correlation matrices. The experts' statistical accuracy for estimating the marginals (i.e., the tributary discharge extreme value distributions) are calculated using the CM, and are shown in the last column. These scores are calculated using a chi-square test that compares the expert estimates to observed discharges, checking both for over- and underconfidence and for location bias. The method and elicitation for this are described in (Rongen et al. 2024).

Two additional d-calibration scores are presented to provide context for the experts' performance. The first is the score for the observed correlation matrix. Estimating this would give the best possible d-calibration score and log-likelihood. The second is the 5% significance level. A score above this level indicates that it is unlikely (<5% probability) that the expert's matrix is uninformed, or, part of the population of randomly drawn NPBNs. Because there is no well-established method for deriving such

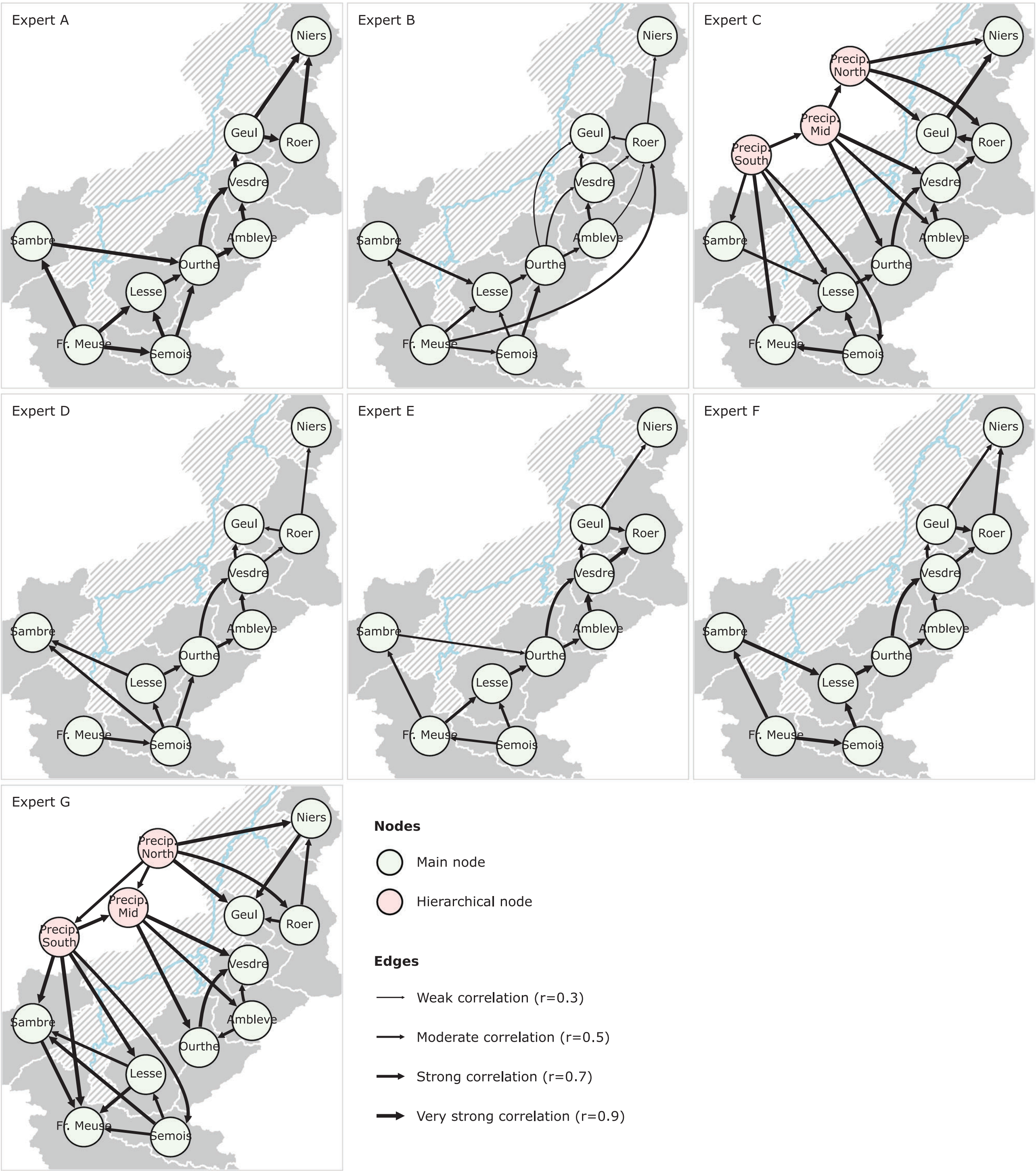


FIGURE 2 | Bayesian networks as drawn by the experts. The thickness of the arrows show the strength of the *nonconditional* correlation. The gray areas on the background represent a map of the catchments between which the dependence is elicited, with the blue line showing the main branches of the Meuse River.

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 95th 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 matrices. The method and results for this are explained in Section B.2.

Based on both the d-calibration score and the log-likelihood shown in Table 2, 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

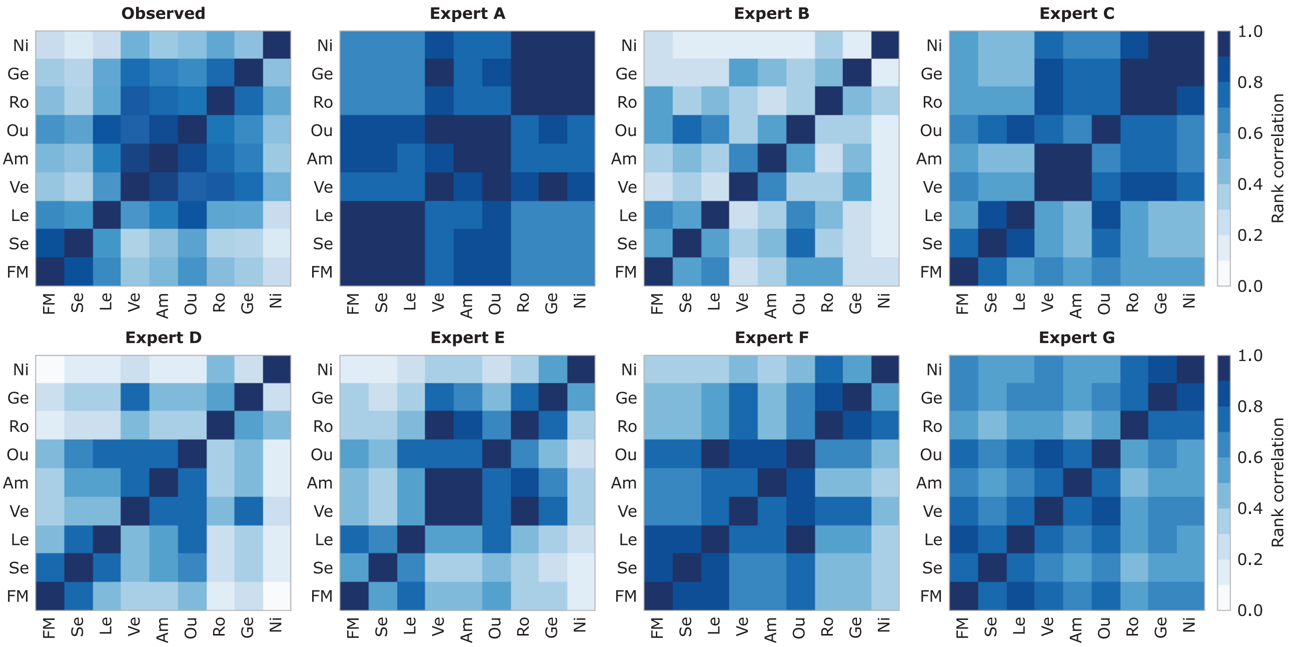


FIGURE 3 | Correlation matrix for observed discharges (top left panel) and correlation matrices representing the expert drawn BNs (other panels).

TABLE 2 | d-Calibration scores and likelihood for experts' and DMs' correlation matrices.

	d-Calibration score	Log-likelihood	Statistical accuracy (CM)
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% sign. level	0.150		

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. This means that even though a few experts score significantly worse than the rest, this hardly affects the result for the equal weight combination. This phenomenon is further investigated in Section 4.2.3. The global weight DM with optimization (GL opt.) is calculated by selecting experts based on a minimum required d-calibration score and calculating 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 nonzero 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.

The statistical accuracy for estimating the marginals shows much steeper varying values than the d-calibration scores. The difference between the best and worst expert is roughly a factor 3 for the d-calibration score, where it is a factor 10^8 for the CM statistical accuracy. Morales-Nápoles, Hanea, & Worm et al. (2014) found that statistical accuracy from CM 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. When comparing the two sets of scores for this study, we find them to be only weakly correlated; a Spearman rank coefficient of 0.21, with a p-value of 0.64 for being independent.

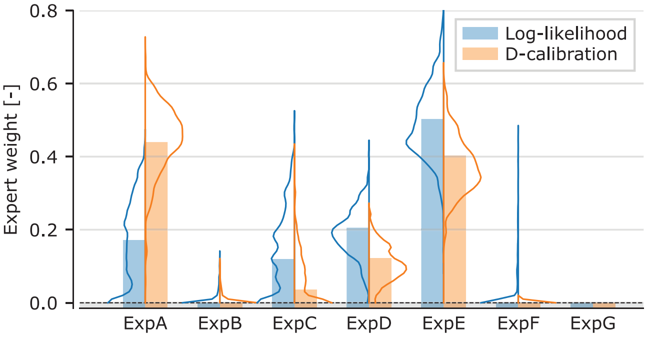


FIGURE 4 | Decision maker weights for each expert, optimized using log-likelihood (blue bars) and d-calibration score (orange bars). The lines show the uncertainty in the weights that were derived from bootstrapping.

This indicates that the two scores are not well correlated for this case study. This is best illustrated by Expert F, who scored worst for univariate estimates, scored second best (d-calibration) or best (log-likelihood) for dependence estimates.

4.2.2 | Finding the "Best" DM

While it is encouraging to see that all DMs score similar to the best expert, their 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 et al. 2014). This is partly due to the experts' d-calibration scores being close together, especially when compared to the scores from the CM. 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. Notice that this approach is different from the GL opt. DM, where the weights are constructed by including d-calibration scores above a cut-off level. Instead, we optimized while allowing all experts' weights to vary freely. To ensure stability of the optimum, we made sure the same optimum was found when using different starting points. Figure 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 weights under re-sampling.

Surprisingly, the results of the weights optimization are different from the d-calibration scores in Table 2. Experts B, F, and G are given almost zero weight, despite having well-approximated correlation matrices (e.g., F had the second-best d-calibration score). Simultaneously, Expert A, with the lowest score, is assigned a very large weight. This is due this experts' high estimated correlations (see Figure 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 case study, remains an open question.

4.2.3 | Robustness of the DMs

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 all unique combinations of experts. This is similar to checking experts' robustness in the CM. The results for using equal weights are visualized in Figure 5, with (a) showing the DM score for each combination including a specific expert and (b) by showing the score per combination of 1–7 experts. The results for the global weight DM are very similar and therefore presented in Section 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 Figure 5 shows). Surprisingly, it matters little which experts are combined, the amount of experts is more important for a good score, as Figure 5b 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 of random correlation matrices to another random correlation matrix. Appendix B shows the details of this.

While a combination with two experts gives the highest score (this is the GLopt DM in Table 2), including more experts is a more robust option as every combination of four 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 B.3 shows this, by including a hypothetical low-scoring expert. Doing this makes the results more sensitive to individual experts, demonstrating the importance of scoring, especially because the bad expert still has 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.

5 | Discussion and Final Remarks

In this study, we elicited the dependence of a the Meuse River tributaries' peak discharges from experts, by having them construct and quantify a NPBN. The experts were scored by the d-calibration score and likelihood of their matrix. Sensitivity analyses were done for the results, to see what combination

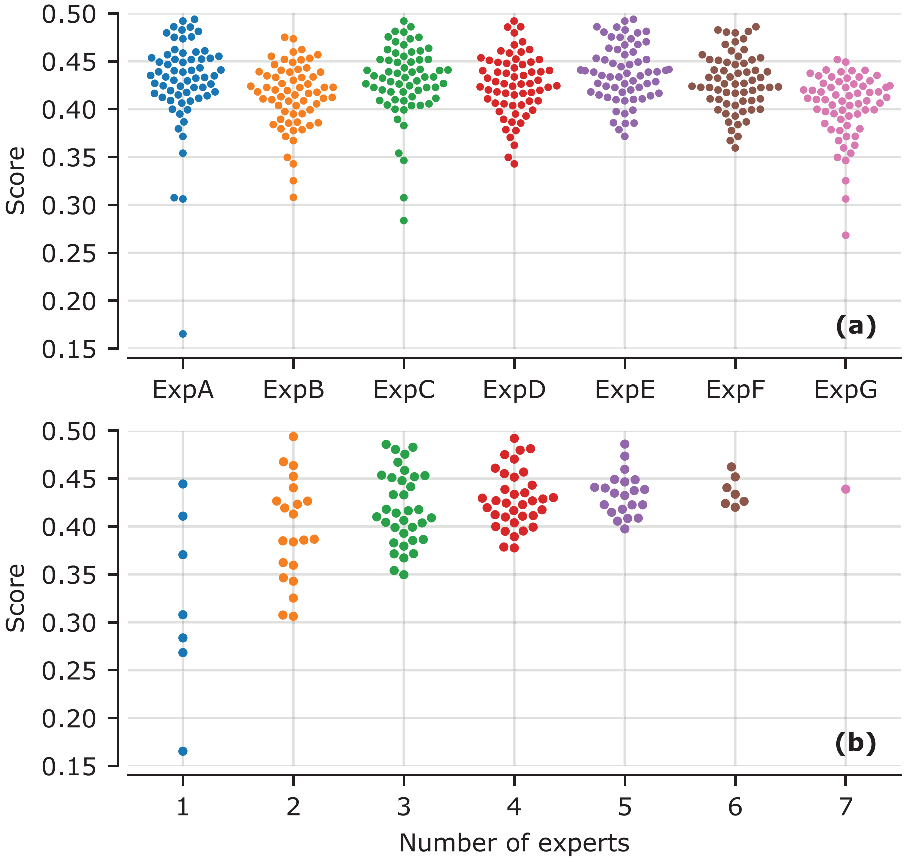


FIGURE 5 | Equal weight decision maker d-calibration scores for every unique combination of 1–7 experts. The top panel (a) groups the scores for all combinations including Expert A (blue), Expert B (orange), etc. The bottom panel (b) groups the scores per number of experts in the combination; 1 expert (blue), 2 experts (orange), etc.

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 5.1) and the more theoretical aspect of scoring (Section 5.2), we hope to contribute to the progression of the field of dependence elicitation.

5.1 | Practice of Expert Dependence Elicitation

NPBNs are not uncommon in scientific hydrological modeling studies (e.g., Paprotny and Morales-Nápoles 2017; Ragno et al. 2022) but unknown to most hydrologists. The concept of a NPBN is for a nonstatistician 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 with relative ease that represented the observed correlations well. 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 Figure C1) in which experts can directly see the

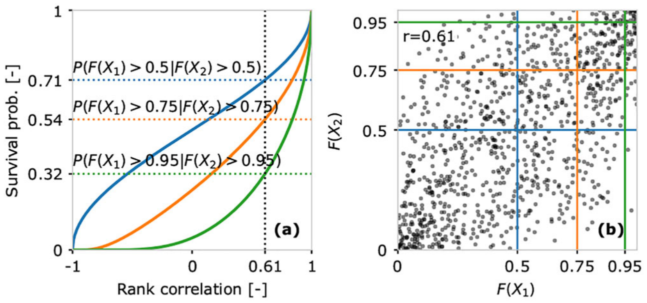


FIGURE 6 | Example of graphical information provided to the experts to aid them in estimating rank correlations. The left graph shows probability that X_2 exceeds a certain quantile provided that X_1 exceeds the quantile, for a given rank correlation. The right scatter illustrates this visually through the point density.

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, of which an example is given in Figure 6a. 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 Tributaries A and B). Figure 6b 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 Figure 6b was perceived easier than based on Figure 6a).

5.2 | Dependence Scoring

The d-calibration score was used to score the experts performance and provide weights for the DMs. 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 indicates 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 4.2.2). 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 best 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 5% significance level (a d-calibration score of 0.15), while only two of the seven experts scored above the significance level for the CM (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 DM 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 CM). 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, since all expert performed decently, and 5% significance level included all.

It is encouraging to see that all seven experts were able to provide good estimates for dependence, while only two 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.

5.3 | 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 DM. Experts estimated the dependence between peak discharges of tributaries within a river catchment, using a NPBN 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 DM, 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 advise 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 variation in correlations, including negative coefficients. 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.

Acknowledgments

We would like to thank all experts that participated in the study, Alexander, Eric, Ferdinand, Helena, Jerom, Nicole, and Siebolt, for

their time and effort in making this research possible. Second, we thank Dorien Lugt en Ties van der Heijden, who's hydrological and statistical expertise greatly helped in preparing the study through test rounds.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Armstrong, M. 1983. *Basic Topology*. Springer-Verlag.

Barons, M. J., S. Mascaro, and A. M. Hanea. 2022. "Balancing the Elicitation Burden and the Richness of Expert Input When Quantifying Discrete Bayesian Networks." *Risk Analysis* 42, no. 6: 1196–1234.

Brown, B. B. 1968. *Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts*. Technical Report. Rand Corp.

Cooke, R. M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press.

Cooke, R. M., and L. L. Goossens. 2008. "TU Delft Expert Judgment Data Base." *Reliability Engineering and System Safety* 93, no. 5: 657–674.

Daneshkhah, A., and J. Oakley. 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.

Delgado-Hernández, D.-J., O. Morales-Nápoles, D. De-León-Escobedo, and J.-C. Arteaga-Arcos. 2014. "A Continuous Bayesian Network for Earth Dams Risk Assessment: An Application." *Structure and Infrastructure Engineering* 10, no. 2: 225–238.

Druzdel, M. J., and L. C. Van Der Gaag. 2000. "Building Probabilistic Networks: 'Where do the numbers come from?'." *IEEE Transactions on Knowledge and Data Engineering* 12, no. 4: 481–486.

Hanea, A. M., Z. Hilton, B. Knight, and A. P. Robinson. 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, no. 6: 1235–1254.

Hanea, A., O. Morales Napoles, and D. Ababei. 2015. "Non-Parametric Bayesian Networks: Improving Theory and Reviewing Applications." *Reliability Engineering & System Safety* 144: 265–284.

Joe, H. 2006. "Generating Random Correlation Matrices Based on Partial Correlations." *Journal of Multivariate Analysis* 97, no. 10: 2177–2189.

Koot, P., M. A. Mendoza-Lugo, D. Paprotny, O. Morales-Nápoles, E. Ragno, and D. T. Worm. 2023. "PyBanshee Version (1.0): A Python Implementation of the MATLAB Toolbox BANSHEE for Non-Parametric Bayesian Networks With Updated Features." *SoftwareX* 21: 101279.

Land NRW. 2022. "ELWAS-WEB." https://www.elwasweb.nrw.de.

Marvin, M., and M. Henryk. 1992. *A Survey of Matrix Theory and Matrix Inequalities*, Volume 14. Courier Dover Publications.

Morales, O., D. Kurowicka, and A. Roelen. 2008. "Eliciting Conditional and Unconditional Rank Correlations From Conditional Probabilities." *Reliability Engineering & System Safety* 93, no. 5: 699–710. [Expert Judgement].

Morales-Nápoles, O., D. J. Delgado-Hernández, D. De-León-Escobedo, and J. C. Arteaga-Arcos. 2014. "A Continuous Bayesian Network for Earth Dams' Risk Assessment: Methodology and Quantification." *Structure and Infrastructure Engineering* 10, no. 5: 589–603.

Morales-Nápoles, O., A. M. Hanea, and D. T. H. Worm. 2014. "Experimental Results About the Assessments of Conditional Rank Correlations by Experts: Example With Air Pollution Estimates." In *Safety, reliability and risk analysis* (1st ed.), 1359–1366. CRC Press.

Morales Nápoles, O., and D. Worm. 2013. *Hypothesis Testing of Multidimensional Probability Distributions*. TNO report.

Moustafa, A., T. Karim, F. De La Torre, and F. Ferrie. 2010. *Designing a Metric for the Difference Between Gaussian Densities*, Volume 83. Springer.

Nelsen, R. B. 2007. *An Introduction to Copulas*. Springer Science & Business Media.

Nyberg, E. P., A. E. Nicholson, K. B. Korb, et al. 2022. "BARD: A Structured Technique for Group Elicitation of Bayesian Networks to Support Analytic Reasoning." *Risk Analysis* 42, no. 6: 1155–1178.

Paprotny, D., and O. Morales-Nápoles. 2017. "Estimating Extreme River Discharges in Europe Through a Bayesian Network." *Hydrology and Earth System Sciences* 21, no. 6: 2615–2636.

Paprotny, D., O. Morales-Nápoles, D. T. Worm, and E. Ragno. 2020. "BANSHEE—A Matlab Toolbox for Non-Parametric Bayesian Networks." *SoftwareX* 12: 100588.

Pearl, J. 2000. *Models, Inference and Reasoning*. Cambridge, UK: Cambridge University Press, 19(2), 3.

Ragno, E., M. Hrachowitz, and O. Morales-Nápoles. 2022. "Applying Non-Parametric Bayesian Networks to Estimate Maximum Daily River Discharge: Potential and Challenges." *Hydrology and Earth System Sciences* 26, no. 6: 1695–1711.

Renooij, S. 2001. "Probability Elicitation for Belief Networks: Issues to Consider." *Knowledge Engineering Review* 16, no. 3: 255–269.

Rijkswaterstaat. 2022. "Waterinfo." https://waterinfo.rws.nl/expert/Afvoer?parameters=Debiet___20Oppervlaktewater___20m3___2Fs.

Rongen, G., and O. Morales-Nápoles. 2024. "Matlatzinca: A PyBANSHEE-Based Graphical User Interface for Elicitation of Non-Parametric Bayesian Networks From Experts." *SoftwareX* 26: 101693.

Rongen, G., O. Morales-Nápoles, and M. Kok. 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, no. 13: 2831–2848.

Service public de Wallonie. 2022. "Voies Hydraulique Wallonie - Annuaires et statistiques." http://voies-hydrauliques.wallonie.be/opencms/opencms/fr/hydro/Archive/annuaires/index.html.

Waterschap Limburg. 2021. "Discharge Measurements." Waterschap Limburg. https://www.waterstandlimburg.nl/Home. [Historical time series from personal communication].

Werner, C., T. Bedford, R. M. Cooke, A. M. Hanea, and O. Morales-Napoles. 2017. "Expert Judgement for Dependence in Probabilistic Modelling: A Systematic Literature Review and Future Research Directions." *European Journal of Operational Research* 258, no. 3: 801–819.

Appendix A: Proofs for Properties of the d-Calibration Score

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 Equation (A1). 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.

$$\begin{aligned} \eta(N_1, N_2) = {} & \frac{|\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{\frac{1}{2}}} \times \\ & \exp\{-\frac{1}{8}(\mu_1 - \mu_2)^T \frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2(\mu_1 - \mu_2)\} \end{aligned} \tag{A1}$$

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 advantages 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 Equation (A1) 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:

$$d_H(\Sigma_1, \Sigma_2) = \sqrt{1 - \frac{|\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2|^{\frac{1}{2}}}} \tag{A2}$$

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 A.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 A.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 A.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 A.2. 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 Figure 3.

The following properties of the d-calibration score hold:

Theorem A.3. *Let the d-calibration score be defined as in Definition A.2. 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 (Moustafa et al. 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 A.1. □

Remark A.4. Each property from Theorem A.3 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 B: Behavior of the d-Calibration Score

In Section 4.2.3, the robustness of the DMs was tested by evaluating the d-calibration score for different combinations of DMs. 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 for that analysis (Section B.1) and investigates if these findings hold for randomly sampled correlation matrices as well (Section B.2). This random matrix sampling is used to define a significance level for the d-calibration score. Finally, Section B.3 shows the effect of adding a low-scoring expert to the pool.

B.1. | Robustness of Mean Matrices for Global DM

Where Figure 5 shows the robustness for equal weights DM to the individual experts (a) and number of experts (b), Figure B1 shows this for the global weights DM. The weights are calculated by normalizing the experts d-calibration scores. The difference between the EQ and GL robustness results is negligible. The global optimized DM (for every number of experts) is circled in Figure B1b. The actual global optimized DM is the combination 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 4.2.2, which showed that the "best" combination is not necessarily a combination of experts with the highest d-calibration scores.

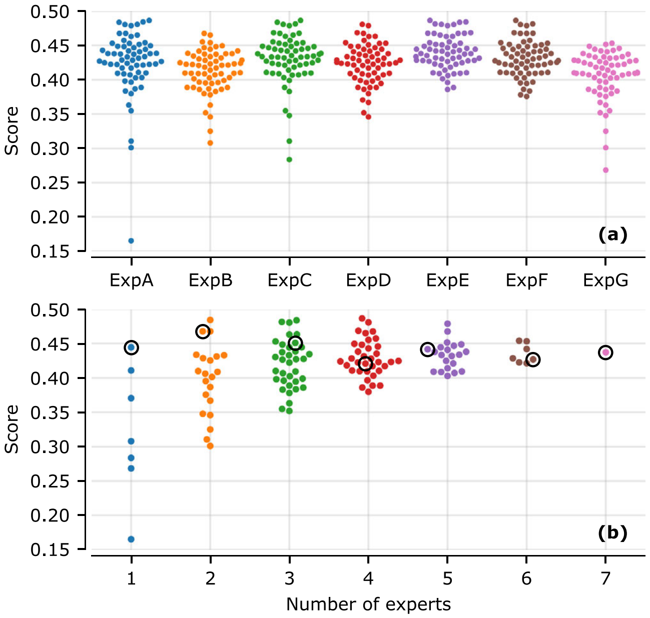


FIGURE B1 | Global weight decision maker d-calibration scores for every unique combination of 1–7 experts. The top panel (a) groups the scores for all combinations including Expert A (blue), Expert B (orange), etc. The bottom panel (b) groups the scores per number of experts in the combination; 1 expert (blue), 2 experts (orange), etc. The optimized combination is circled.

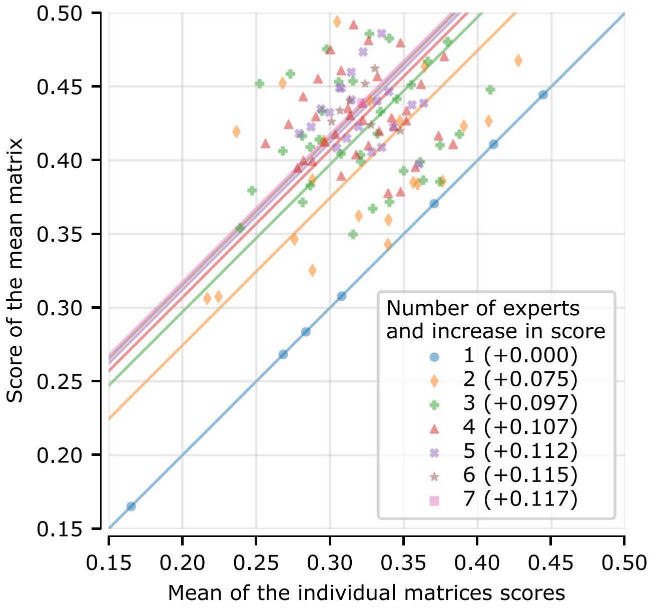


FIGURE B2 | Comparison between the d-calibration score of the mean matrix and the mean score of the individual contributing matrices. Each marker represent a unique combination of 1–7 experts.

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 CM for univariate uncertainty (see, for example, Cooke and 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 Figure B1b, as this is the number of experts that usually share the majority of the weight in the CM.

A different representation of the effect of averaging the matrices is shown in Figure B2. 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.

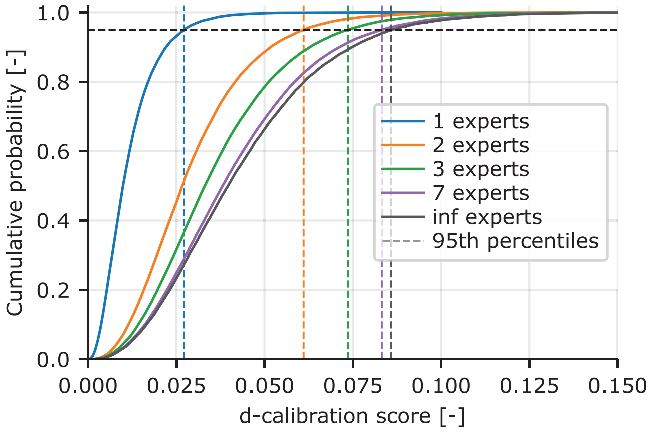


FIGURE B3 | CDFs of the d-calibration scores of averaged randomly sampled correlation matrices, compared to a random "true" matrix. Each line represents the distribution when averaging a number of experts' matrices (1, 2, 3, 7, and "infinite"). Matrices were generated by sampling conditional rank-correlations from $\mathcal{U}(-1, 1)$.

B.2. | d-Calibration Scores for Random Matrices

Last section's analysis showed that combining correlation matrices elicited from the experts consistently results in a better d-calibration score than using individual expert's matrices. To test this improvement of the scores in a more general setting, random correlation matrices were sampled for 1, 2, 3, 7, or "infinite" experts, averaged (using equal weights), and compared to a different 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 $\mathcal{U}(-1, 1)$. 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 randomly quantifying a BN (under the assumption that the graph is saturated and the coefficients are drawn from $\mathcal{U}(-1, 1)$) and (b) it is easy to create matrices with constraints on the distribution of the conditional rank correlations. The sampling procedure is as follows:

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 B3 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.

The consistent improvement in d-calibration score when averaging random matrices is similar to the pattern observed in Section 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 Figure B4, 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 deciding the rank correlations should be drawn from $\mathcal{U}(0, 1)$ instead of $\mathcal{U}(-1, 1)$ (no expert estimated a negative rank correlation coefficient during the elicitation). This would result in a significance level of 0.15, as shown in Figure B4. Note that this result is used as the 5% significance level for judging expert performance in Section 4.2. If an expert scores higher than this, the chance is less than 5% that the expert was making (almost) uninformed guesses.

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 4.2 showed that the DMs do not perform significantly better than the best experts, and Section 4.2.3 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 B1. 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

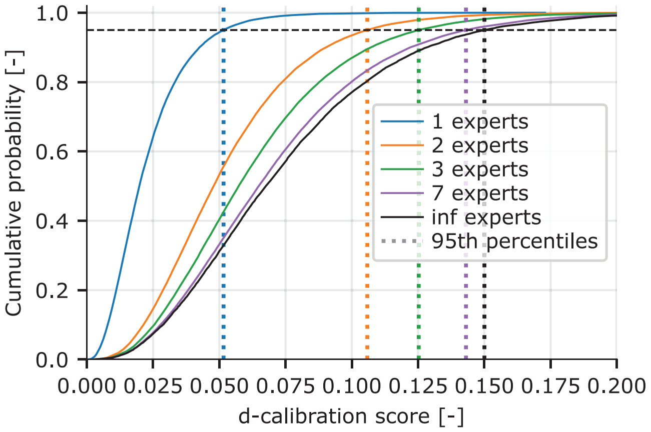


FIGURE B4 | CDFs of the d-calibration scores of averaged randomly sampled correlation matrices, compared to a random “true” matrix. Each line represents the distribution when averaging a number of experts’ matrices (1, 2, 3, 7, and “infinite”). Matrices were generated by sampling conditional rank-correlations from $\mathcal{U}(0, 1)$.

TABLE B1 | d-Calibration scores and likelihood for experts’ and DMs’ correlation matrices, when a low-scoring expert is added to the pool.

	d-Calibration score	**Log-likelihood**
Low-scoring expert	0.107	−1348
Other experts (A–G)	[0.165, 0.444]	[−2443, −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	0.150	

d-calibration score, an estimate with high correlations gives significantly worse log-likelihood than an estimate of independence (compare the scores of Experts A and C to the “low-scoring” expert).

In the robustness analysis where are all combination of experts are calculated, the effect of the low-scoring expert becomes clearer. Figure B5 shows scores for each combination of 1–8 (i.e., 7 + the hypothetical low-scoring expert) experts. For the EQ DM (panel a), two clusters are distinguished, 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.

The same clustering can be observed in Figure B6 where, similar to Figure B2, 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, including 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.

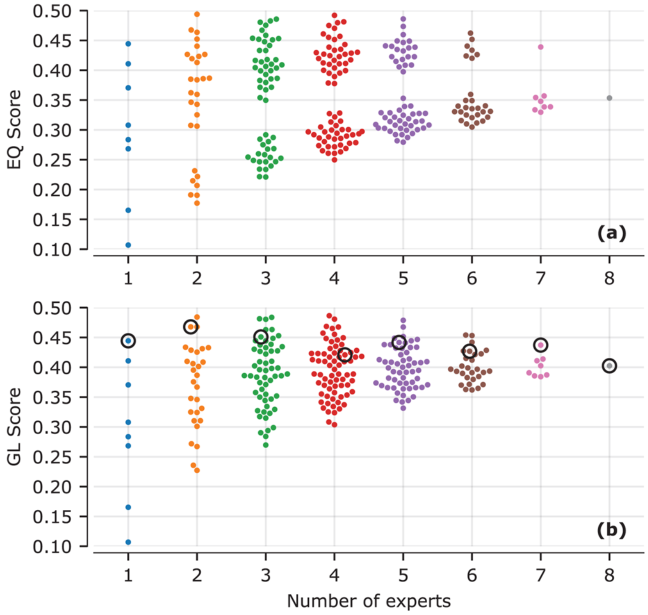


FIGURE B5 | Equal weights (top) and global weights (bottom) decision maker d-calibration scores for every unique combination of 1–8 experts (an eighth, low-scoring, expert was added to the pool). Both panels group the scores per number of experts in the combination; 1 expert (blue), 2 experts (orange), etc. The optimized combination is circled in the bottom panel.

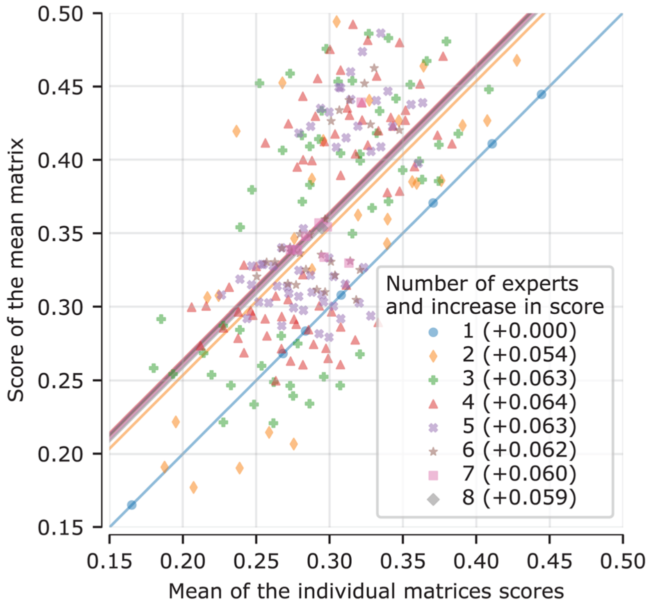
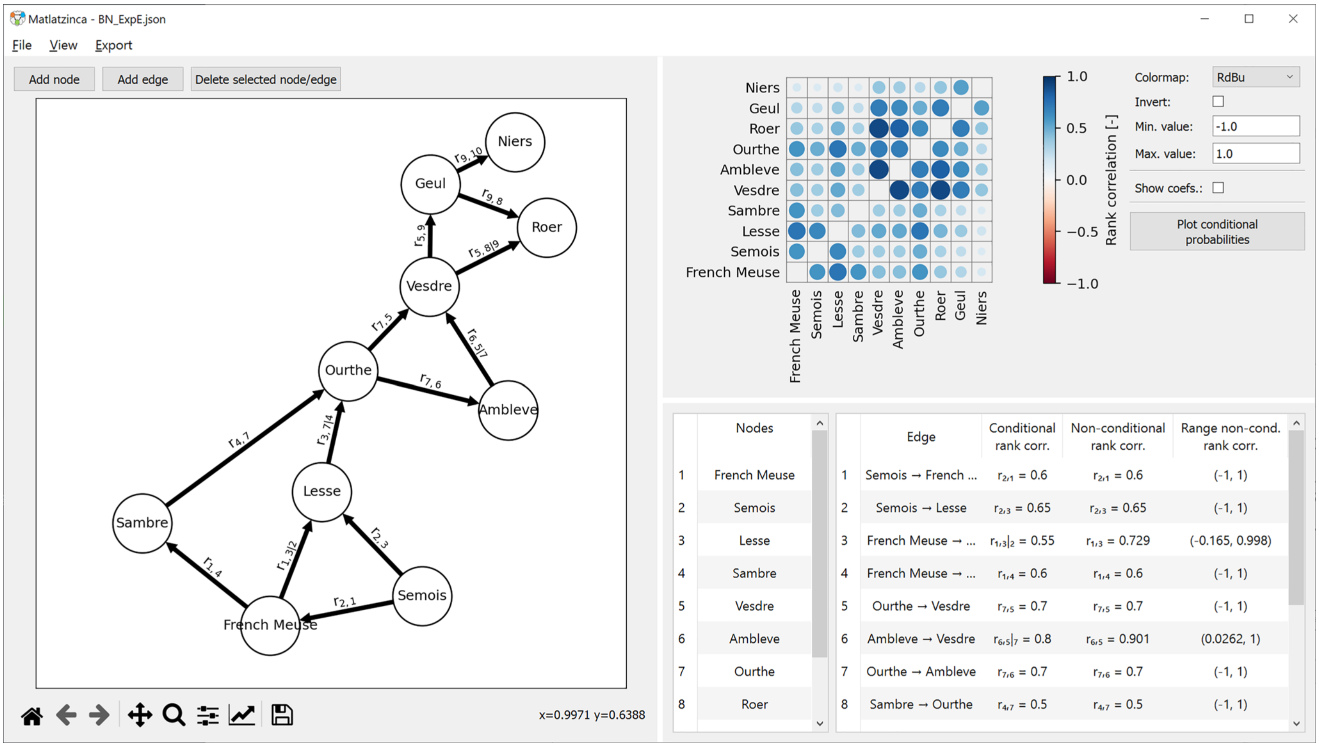


FIGURE B6 | Comparison between the d-calibration score of the mean matrix and the mean score of the individual contributing matrices. An eighth, low-scoring, expert was added to the pool. Each marker represent a unique combination of 1–8 experts.

Appendix C: Matlatzinca

The software used for eliciting correlations through the schematization and quantification of a BN is called Matlatzinca. Its details are described in (Rongen and Morales-Nápoles 2024). Figure C1 shows a screenshot of the program. The BN is shown on the left hand side. The table on the bottom right is used to quantify the conditional or nonconditional rank correlations for the edges. The resulting correlation matrix is shown on the top right.



Colormap: RdBu

Invert: □

Min. value: -1.0

Max. value: 1.0

Show coefs.: □

	Nodes
1	French Meuse
2	Semois
3	Lesse
4	Sambre
5	Vesdre
6	Ambleve
7	Ourthe
8	Roer

	Edge	Conditional rank corr.	Non-conditional rank corr.	Range non-cond. rank corr.
1	Semois → French ...	$r_{2,1} = 0.6$	$r_{2,1} = 0.6$	(-1, 1)
2	Semois → Lesse	$r_{2,3} = 0.65$	$r_{2,3} = 0.65$	(-1, 1)
3	French Meuse → ...	$r_{1,3\|2} = 0.55$	$r_{1,3} = 0.729$	(-0.165, 0.998)
4	French Meuse → ...	$r_{1,4} = 0.6$	$r_{1,4} = 0.6$	(-1, 1)
5	Ourthe → Vesdre	$r_{7,5} = 0.7$	$r_{7,5} = 0.7$	(-1, 1)
6	Ambleve → Vesdre	$r_{6,5\|7} = 0.8$	$r_{6,5} = 0.901$	(0.0262, 1)
7	Ourthe → Ambleve	$r_{7,6} = 0.7$	$r_{7,6} = 0.7$	(-1, 1)
8	Sambre → Ourthe	$r_{4,7} = 0.5$	$r_{4,7} = 0.5$	(-1, 1)

FIGURE C1 | Screenshot of the program "Matlatzinca," which the experts used to schematize and quantify their Bayesian networks.

Tables C1–C7 show the rank correlations estimated by the experts. Each table contains four columns. The first column (Edge) indicates the edge for which a correlation is estimated, including the node IDs. For example, "Semois → Lesse ($r_{2,3|1}$)" in Table C1 indicates the rank correlation between Semois (node 2) and Lesse (node 3), conditional to the French Meuse (node 1). The second column (Cond.) indicates the rank correlations conditional to any parent nodes (such as the French Meuse in the previous example). These conditional correlations range from –1 to 1, but the "would-be-observed" correlation might be different. These nonconditional correlations are shown in the third column (Non-c). Matlatzinca allows specifying these coefficients, which may range between the values in the fourth column (Range (non-c). The round values in the "Cond." column indicates that most expert chose to specify the conditional, rather than the unconditional correlations.

TABLE C1 | Correlation estimates for Expert A.

Edge	Cond.	Non-c	Range (non-c)
Fr. Meuse → Semois ($r_{1,2}$)	0.99	0.99	(−1, 1)
Fr. Meuse → Lesse ($r_{1,3}$)	0.95	0.95	(−1, 1)
Semois → Lesse ($r_{2,3\|1}$)	0	0.941	(0.897, 0.984)
Fr. Meuse → Sambre ($r_{1,4}$)	0.95	0.95	(−1, 1)
Ourthe → Vesdre ($r_{7,5}$)	0.9	0.9	(−1, 1)
Ambleve → Vesdre ($r_{6,5\|7}$)	0	0.856	(0.725, 0.991)
Ourthe → Ambleve ($r_{7,6}$)	0.95	0.95	(−1, 1)
Sambre → Ourthe ($r_{4,7}$)	0.9	0.9	(−1, 1)
Lesse → Ourthe ($r_{3,7\|4}$)	0	0.814	(0.636, 1)
Semois → Ourthe ($r_{2,7\|3,4}$)	0	0.847	(0.737, 0.961)
Geul → Roer ($r_{9,8}$)	0.95	0.95	(−1, 1)
Vesdre → Geul ($r_{5,9}$)	0.9	0.9	(−1, 1)
Geul → Niers ($r_{9,10}$)	0.95	0.95	(−1, 1)
Roer → Niers ($r_{8,10\|9}$)	0	0.903	(0.808, 1)

TABLE C2 | Correlation estimates for Expert B.

Edge	Cond.	Non-c	Range (non-c)
Fr. Meuse → Semois ($r_{1,2}$)	0.55	0.55	(−1, 1)
Semois → Lesse ($r_{2,3}$)	0.55	0.55	(−1, 1)
Fr. Meuse → Lesse ($r_{1,3\|2}$)	0.55	0.69	(−0.34, 1)
Sambre → Lesse ($r_{4,3\|1,2}$)	0.5	0.696	(−0.091, 0.967)
Fr. Meuse → Sambre ($r_{1,4}$)	0.6	0.6	(−1, 1)
Ourthe → Vesdre ($r_{7,5}$)	0.4	0.4	(−1, 1)
Ambleve → Vesdre ($r_{6,5\|7}$)	0.6	0.682	(−0.441, 0.969)
Ourthe → Ambleve ($r_{7,6}$)	0.6	0.6	(−1, 1)
Lesse → Ourthe ($r_{3,7}$)	0.65	0.65	(−1, 1)
Semois → Ourthe ($r_{2,7\|3}$)	0.6	0.742	(−0.223, 0.991)
Ambleve → Roer ($r_{6,8}$)	0.3	0.3	(−1, 1)
Fr. Meuse → Roer ($r_{1,8\|6}$)	0.5	0.552	(−0.766, 0.999)
Vesdre → Roer ($r_{5,8\|1,6}$)	0.3	0.38	(−0.353, 0.781)
Vesdre → Geul ($r_{5,9}$)	0.6	0.6	(−1, 1)
Ourthe → Geul ($r_{7,9\|5}$)	0.1	0.318	(−0.441, 0.969)
Roer → Geul ($r_{8,9\|5,7}$)	0.3	0.464	(−0.416, 0.955)
Roer → Niers ($r_{8,10}$)	0.4	0.4	(−1, 1)

TABLE C3 | Correlation estimates for Expert C.

Edge	Cond.	Non-c	Range (non-c)
Semois → Fr. Meuse ($r_{2,1}$)	0.8	0.8	(–1, 1)
Prec. S → Fr. Meuse ($r_{11,1\|2}$)	0.7	0.882	(0.238, 0.997)
Prec. S → Semois ($r_{11,2}$)	0.75	0.75	(–1, 1)
Fr. Meuse → Lesse ($r_{1,3}$)	0.6	0.6	(–1, 1)
Semois → Lesse ($r_{2,3\|1}$)	0.8	0.864	(0.0435, 0.956)
Prec. S → Lesse ($r_{11,3\|1,2}$)	0.8	0.747	(0.375, 0.786)
Sambre → Lesse ($r_{4,3\|1,2,11}$)	0.65	0.653	(0.347, 0.715)
Prec. S → Sambre ($r_{11,4}$)	0.7	0.7	(–1, 1)
Ourthe → Vesdre ($r_{7,5}$)	0.8	0.8	(–1, 1)
Ambleve → Vesdre ($r_{6,5\|7}$)	0.9	0.942	(0.148, 0.983)
Prec. C → Vesdre ($r_{12,5\|6,7}$)	0.8	0.933	(0.739, 0.953)
Prec. C → Ambleve ($r_{12,6}$)	0.8	0.8	(–1, 1)
Prec. C → Ourthe ($r_{12,7}$)	0.85	0.85	(–1, 1)
Lesse → Ourthe ($r_{3,7\|12}$)	0.75	0.805	(0.0853, 0.908)
Prec. N → Roer ($r_{13,8}$)	0.9	0.9	(–1, 1)
Vesdre → Roer ($r_{5,8\|13}$)	0.7	0.878	(0.406, 0.96)
Prec. N → Geul ($r_{13,9}$)	0.85	0.85	(–1, 1)
Roer → Geul ($r_{8,9\|13}$)	0.8	0.951	(0.551, 0.994)
Prec. N → Niers ($r_{13,10}$)	0.85	0.85	(–1, 1)
Geul → Niers ($r_{9,10\|13}$)	0.7	0.92	(0.466, 1)
Prec. S → Prec. C ($r_{11,12}$)	0.75	0.75	(–1, 1)
Prec. C → Prec. N ($r_{12,13}$)	0.8	0.8	(–1, 1)

TABLE C4 | Correlation estimates for Expert D

Edge	Cond.	Non-c	Range (non-c)
Fr. Meuse → Semois ($r_{1,2}$)	0.7	0.7	(–1, 1)
Semois → Lesse ($r_{2,3}$)	0.7	0.7	(–1, 1)
Lesse → Sambre ($r_{3,4}$)	0.7	0.7	(–1, 1)
Semois → Sambre ($r_{2,4\|3}$)	0.4	0.699	(0.0262, 1)
Ourthe → Vesdre ($r_{7,5}$)	0.7	0.7	(–1, 1)
Ambleve → Vesdre ($r_{6,5\|7}$)	0.36	0.7	(0.0953, 0.997)
Ourthe → Ambleve ($r_{7,6}$)	0.75	0.75	(–1, 1)
Lesse → Ourthe ($r_{3,7}$)	0.7	0.7	(–1, 1)
Semois → Ourthe ($r_{2,7\|3}$)	0.4	0.699	(0.0262, 1)
Vesdre → Roer ($r_{5,8}$)	0.5	0.5	(–1, 1)
Vesdre → Geul ($r_{5,9}$)	0.7	0.7	(–1, 1)
Roer → Geul ($r_{8,9\|5}$)	0.235	0.5	(–0.216, 0.965)
Roer → Niers ($r_{8,10}$)	0.5	0.5	(–1, 1)

TABLE C5 | Correlation estimates for Expert E.

Edge	Cond.	Non-c	Range (non-c)
Semois → Fr. Meuse ($r_{2,1}$)	0.6	0.6	(−1, 1)
Semois → Lesse ($r_{2,3}$)	0.65	0.65	(−1, 1)
Fr. Meuse → Lesse ($r_{1,3\|2}$)	0.55	0.729	(−0.165, 0.998)
Fr. Meuse → Sambre ($r_{1,4}$)	0.6	0.6	(−1, 1)
Ourthe → Vesdre ($r_{7,5}$)	0.7	0.7	(−1, 1)
Ambleve → Vesdre ($r_{6,5\|7}$)	0.8	0.901	(0.0262, 1)
Ourthe → Ambleve ($r_{7,6}$)	0.7	0.7	(−1, 1)
Sambre → Ourthe ($r_{4,7}$)	0.5	0.5	(−1, 1)
Lesse → Ourthe ($r_{3,7\|4}$)	0.65	0.729	(−0.503, 0.998)
Geul → Roer ($r_{9,8}$)	0.7	0.7	(−1, 1)
Vesdre → Roer ($r_{5,8\|9}$)	0.8	0.901	(0.0262, 1)
Vesdre → Geul ($r_{5,9}$)	0.7	0.7	(−1, 1)
Geul → Niers ($r_{9,10}$)	0.55	0.55	(−1, 1)

TABLE C6 | Correlation estimates for Expert F.

Edge	Cond.	Non-c	Range (non-c)
Fr. Meuse → Semois ($r_{1,2}$)	0.85	0.85	(−1, 1)
Semois → Lesse ($r_{2,3}$)	0.85	0.85	(−1, 1)
Sambre → Lesse ($r_{4,3\|2}$)	0.8	0.89	(0.23, 0.963)
Fr. Meuse → Sambre ($r_{1,4}$)	0.8	0.8	(−1, 1)
Ambleve → Vesdre ($r_{6,5}$)	0.7	0.7	(−1, 1)
Ourthe → Vesdre ($r_{7,5\|6}$)	0.7	0.863	(0.17, 0.987)
Ourthe → Ambleve ($r_{7,6}$)	0.8	0.8	(−1, 1)
Lesse → Ourthe ($r_{3,7}$)	0.9	0.9	(−1, 1)
Vesdre → Roer ($r_{5,8}$)	0.7	0.7	(−1, 1)
Geul → Roer ($r_{9,8\|5}$)	0.7	0.851	(0.0262, 1)
Vesdre → Geul ($r_{5,9}$)	0.7	0.7	(−1, 1)
Roer → Niers ($r_{8,10}$)	0.7	0.7	(−1, 1)
Geul → Niers ($r_{9,10\|8}$)	0	0.6	(0.254, 0.968)

TABLE C7 | Correlation estimates for Expert G.

Edge	Cond.	Non-c	Range (non-c)
Semois → Fr. Meuse ($r_{2,1}$)	0.7	0.7	(−1, 1)
Lesse → Fr. Meuse ($r_{3,1\|2}$)	0.7	0.851	(0.0262, 1)
Sambre → Fr. Meuse ($r_{4,1\|2,3}$)	0.7	0.847	(0.436, 0.919)
Prec. N → Fr. Meuse ($r_{13,1\|2,3,4}$)	0.8	0.967	(0.837, 0.98)
Prec. N → Semois ($r_{13,2}$)	0.8	0.8	(−1, 1)
Semois → Lesse ($r_{2,3}$)	0.7	0.7	(−1, 1)
Prec. N → Lesse ($r_{13,3\|2}$)	0.8	0.905	(0.17, 0.987)
Lesse → Sambre ($r_{3,4}$)	0.7	0.7	(−1, 1)
Semois → Sambre ($r_{2,4\|3}$)	0.7	0.851	(0.0262, 1)
Prec. N → Sambre ($r_{13,4\|2,3}$)	0.8	0.894	(0.586, 0.927)
Ambleve → Vesdre ($r_{6,5}$)	0.7	0.7	(−1, 1)
Ourthe → Vesdre ($r_{7,5\|6}$)	0.7	0.851	(0.0262, 1)
Prec. C → Vesdre ($r_{12,5\|6,7}$)	0.8	0.949	(0.637, 0.982)
Prec. C → Ambleve ($r_{12,6}$)	0.8	0.8	(−1, 1)
Ambleve → Ourthe ($r_{6,7}$)	0.7	0.7	(−1, 1)
Prec. C → Ourthe ($r_{12,7\|6}$)	0.8	0.905	(0.17, 0.987)
Prec. S → Roer ($r_{11,8}$)	0.8	0.8	(−1, 1)
Roer → Geul ($r_{8,9}$)	0.7	0.7	(−1, 1)
Niers → Geul ($r_{10,9\|8}$)	0.7	0.851	(0.0262, 1)
Prec. S → Geul ($r_{11,9\|8,10}$)	0.8	0.949	(0.637, 0.982)
Roer → Niers ($r_{8,10}$)	0.7	0.7	(−1, 1)
Prec. S → Niers ($r_{11,10\|8}$)	0.8	0.905	(0.17, 0.987)
Prec. S → Prec. C ($r_{11,12}$)	0.7	0.7	(−1, 1)
Prec. N → Prec. C ($r_{13,12\|11}$)	0.7	0.851	(0.0262, 1)
Prec. S → Prec. N ($r_{11,13}$)	0.7	0.7	(−1, 1)