Using structured expert judgment to ~~Estimate extreme river discharges~~estimate extremes: a case study of discharges in the Meuse River

Accurate estimation of extreme discharges in rivers, such as the Meuse, is crucial for effective flood risk assessment. However, ~~existing statistical and~~ hydrological models that estimate ~~these~~such discharges often lack transparency regarding the uncertainty of their predictions~~, as~~. This was evidenced by the devastating flood event that occurred in July 2021 which was not captured by the existing model for estimating design discharges. This article proposes an ~~alternative~~ approach to obtain uncertainty estimates for extremes with ~~a central role for~~structured expert judgment, using Cooke's method. A simple statistical model was developed for the river basin, consisting of correlated GEV~~-~~ distributions for discharges ~~in~~from upstream ~~sub-catchments.~~tributaries. The model was fitted to ~~expert judgments,~~seven experts' estimates and historical measurements~~, and the combination of both,~~ using ~~Markov chain Monte Carlo.~~Bayesian inference. Results ~~from the model~~ fitted to only ~~to~~the measurements were ~~accurate~~ solely informative for more frequent events, ~~but less certain for extreme events. Using~~while fitting to only the expert ~~judgment~~estimates reduced uncertainty solely for ~~these~~ extremes~~ but was less accurate for more frequent events. The combined approach~~. Combining both historical observations and estimates of extremes provided the most plausible results~~, with~~. Cooke's method ~~reducing~~reduced the uncertainty by appointing most weight to the two ~~of the seven~~ most accurate experts~~.~~, according to their estimates of less extreme discharges. The study demonstrates that ~~utilizing~~with the presented Bayesian approach that combines historical data and expert-informed priors, a group of hydrological experts ~~in this manner~~ can provide plausible ~~results~~estimates for discharges, and potentially also other (hydrological) extremes, with a relatively ~~limited~~manageable effort~~, even in situations where measurements are scarce or unavailable~~.

1 Introduction

~~Quantifying the uncertainty that comes with estimating~~Estimating the magnitude of extreme flood events ~~is a difficult matter.~~comes with considerable uncertainty. This became clear once more on the 18th of July 2021~~,~~ when ~~the~~a flood wave on the Meuse River~~, following~~ that followed from a few days of rain in the Eiffel and Ardennes, ~~reached its~~caused the highest peak discharge ever measured at Borgharen. Unprecedented rainfall volumes fell ~~within~~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~~(Task Force Fact-finding hoogwater 2021~~ (TFFF 2021; Mohr et al. 2022). 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 ~~irrelevant for extreme discharges on the Meuse.~~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 event was thus surprising in multiple ways. This might happen when we experience a new extreme, but given that Dutch flood risk has safety standards up to once per 100,000 years (Ministry of Infrastructure and Environment 2016) one would have hoped this to be less of a surprise. ~~This~~The event ~~underlines~~underscores the importance of understanding the uncertainty that comes with estimates of extreme flood events.

~~Quantifying~~Estimating the magnitude of events ~~that are more extreme~~greater than ~~ever measured (i.e., with return levels that are longer than the time period of~~ the largest from historical (representative ~~measurements),~~) records is a nontrivial task. It requires establishing a model that describes the occurrence of such events and subsequently extrapolating ~~from available data or knowledge.~~to specific exceedance probabilities from this model. For the Meuse, the traditional approach to this is ~~traditionally done by fitting~~to fit a probability distribution to extreme events and ~~extrapolating~~extrapolate from it (Langemheen and Berger 2001). ~~However, a~~A statistical fit ~~of extremes is often very~~to observations is, however, sensitive to the most extreme events in the ~~measured~~ time series available. Additionally, the hydrological and hydraulic response during extreme events might be ~~of a~~ different ~~type (statistical population) compared to regular events~~from that of events that occur more frequently, and therefore ~~badly~~incorrectly described by ~~a model that is~~ 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 river Meuse (and the Rhine) in the Netherlands. GRADE is a variant on ~~historical measurements~~ a conventional regional flood frequency analysis procedure. Instead of using only~~. Advances in computational power allow to narrow this (hydrological) uncertainty by, in the first place, generating~~

historical observations, it resamples these into long synthetic time series of rainfall ~~or~~ that contain the observed spatial and temporal variation. In then uses a hydrological model to calculate tributary flows and a hydraulic model to subsequently simulate river discharges (Leander et al. 2005; ~~Diederen et al. 2019). These can then be used to simulate a chain of models to approximate values of river discharges and potential flooding (e.g. Falter et al. 2015; Borgomeo, Farmer, and Hall 2015). For the Dutch rivers Meuse and Rhine, the GRADE instrument is used for this. It generates 50,000 years of rainfall and discharges (~~Hegnauer et al. 2014). ~~Advantages of such methods are~~Despite the fact that ~~it~~GRADE can create spatially coherent results~~, which~~ and can ~~correct for~~simulate changes in the catchment or climate~~. It~~, it is ~~however~~ still based on "~~re-sampling~~resampling" available measurements or knowledge. ~~Moreover, the computational resources required to simulate longer time series make it tedious or costly to quantify uncertainty in the method, let alone the uncertainty that comes with modelling choices.~~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.~~ The event could have been more extreme than a once in 50,000 year event, but a more likely cause is that the re-sampling approach uses (only) historic rainfall measurements to create new rain events. Underestimating uncertainty is however not~~

GRADE is only one example where the underestimation of uncertainty is observed. However, it is certainly not the only ~~an issue of GRADE. Meresa and Romanowicz (2017) show that hydrological model uncertainty can be larger than climate projection uncertainty and Winter et al. (2018) indicate that flood model uncertainties can lead up to a factor 3 difference in outcome (1.4 for flood frequency distribution).~~one. Boer-Euser et al. (2017; Bouaziz et al. 2020~~)~~), for example, compared different hydrological modelling 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.~~ flow. But regardless of the conceptual choices, all models would 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 occured. 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 and Thorndycraft 2005). Another approach utilizes 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, Demarée, and Willems (2017).

~~While most~~Another alternative to the data-based approach is the use of structured expert judgment (SEJ). Expert judgment (EJ), in terms of making estimates or verifying observations based on prior knowledge, is often unknowingly applied in everyday practice by researchers and practitioners. It is a way of assessing the truth or value of new information, and therefore indispensable in every scientific application. However, quantifying it is not straightforward. *Structured* expert judgment formalizes this process by eliciting expert judgments in such a way that they can be treated as scientific data. One structured method for this is Cooke's method, also called the *classical model* (Roger M. Cooke and Goossens 2008). Cooke's method assigns a weight to each expert within a group (usually 5 to 10 experts) based on their performance as uncertainty assessors 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. (Roger M. Cooke and Goossens 2008) shows an overview of the different fields in which Cooke's method 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, Mazzuchi, and Cooke (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 ~~obtaining better estimates of discharge extremes use hydrological or statistical modeling, some follow the approach of using expert judgment (EJ). (Sebok et al. 2021) recently applied expert elicitation to reduce~~informing the uncertainty in climate model predictions~~. (~~ (e.g., Oppenheimer, Little, and Cooke 2016; Bamber et al. 2019; Sebok et al. 2021). More closely related to this article, Kindermann et al. ((2020) reproduced historical water levels using structured expert judgment (SEJ), and ~~(~~G. Rongen, Morales-Nápoles, and Kok ((2022a) ~~recently~~ applied SEJ to estimate the probabilities of dike failure ~~probabilities~~ for the Dutch part of the ~~river~~ Rhine~~.~~ River.

While examples of using Cooke's method in hydrology are not abundantly available, ~~using~~they are for applications that use prior information to decrease ~~the~~ uncertainty and sensitivity ~~for extrapolation, is not new. Mostly this~~. This information often comes from ~~EJ~~expert judgments. Three examples in which a similar, Bayesian, approach was applied to limit the uncertainty in extreme discharge estimates are given by (Coles and Tawn 1996; Parent and Bernier 2003; Viglione et al. 2013). The mathematical approaches vary between the different studies, but the rationale for using EJ is the same: adding ~~"soft" or~~ uncertain prior information to available measurements ~~can~~to help ~~in achieving~~achieve more plausible extreme estimates. ~~In this~~

This study~~, we applied~~ applies structured expert judgment ~~as well,~~ 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 ~~aimed~~aim to get ~~credible~~uncertainty estimates ~~of extreme discharge events~~ for these discharges. Their credibility is assessed by comparing them to GRADE, the aforementioned model-based method for deriving the Meuse ~~that would otherwise require statistical extrapolation or complex modelling.~~River's design flood frequency statistics. A relatively simple model ~~was~~is quantified both with observed ~~data and with expert estimates, in which the latter serves to decrease uncertainty in~~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 combined opinion, while the 1000-year discharges are used to inform the extrapolated range. ~~By using Cooke's method for structured expert judgment (SEJ) (Cooke and Goossens 2008), participants can~~Participants use their own approach to ~~make their estimates, which are then combined based on the experts' performance in questions for which the analysts know the answer or will know the answer *post hoc*. Next, we investigate if this expert judgment-based method gives plausible results. For this, we compared the model results based on~~ come up with uncertainty estimates. To investigate how the method that combines data and expert judgments compares to the data-only or the expert estimates ~~only, on measured data only and on both.~~-only approaches, we quantify the model based on all three options. The differences show the added value of each component. This indicates ~~how good~~ the ~~method works,~~method's performance both when measurements are available and when they are not, for example in data scarce areas.

~~This study shows that the proposed method for estimating extreme river discharges with expert judgment leads to credible estimates for extremes. While the method that combines discharge measurements and expert estimates works best, the approach with equally weighted expert judgments only leads to plausible estimates as well, albeit with higher uncertainty in the extrapolated range.~~

2 Study area and ~~available~~ data used

Figure 1 shows an overview of the catchment of the Meuse River. The ~~catchment areas~~catchments that ~~discharge trough~~correspond to the ~~major~~main tributaries are outlined ~~with~~in red. ~~the~~The three locations for which we are interested in extreme discharge estimates, Borgharen, Roermond, and Gennep, are ~~shown in~~colored blue. We call these ~~the~~

‘downstream locations’ throughout this study. The river continues further downstream until it flows into the North Sea ~~at~~near Rotterdam. This part of the river becomes increasingly intertwined with the Rhine River and more affected by the downstream sea water level. ~~The~~Consequently, the water levels~~ consequently~~ can be ascribed decreasingly to the discharge from the upstream catchment. For this reason, we do not ~~go~~assess discharges further downstream than Gennep in this study.

The numbered dots indicate the locations along the tributaries where the ~~discharge of the upstream sub-catchments~~discharges are measured. ~~The gauge~~These locations’ names and ~~catchments’~~the tributaries’ names are shown on the lower left~~. The Semois catchment is part of the French Meuse catchment. The discharge estimates for this catchment are therefore only used for expert calibration, as the flow is part of the French Meuse flow~~.

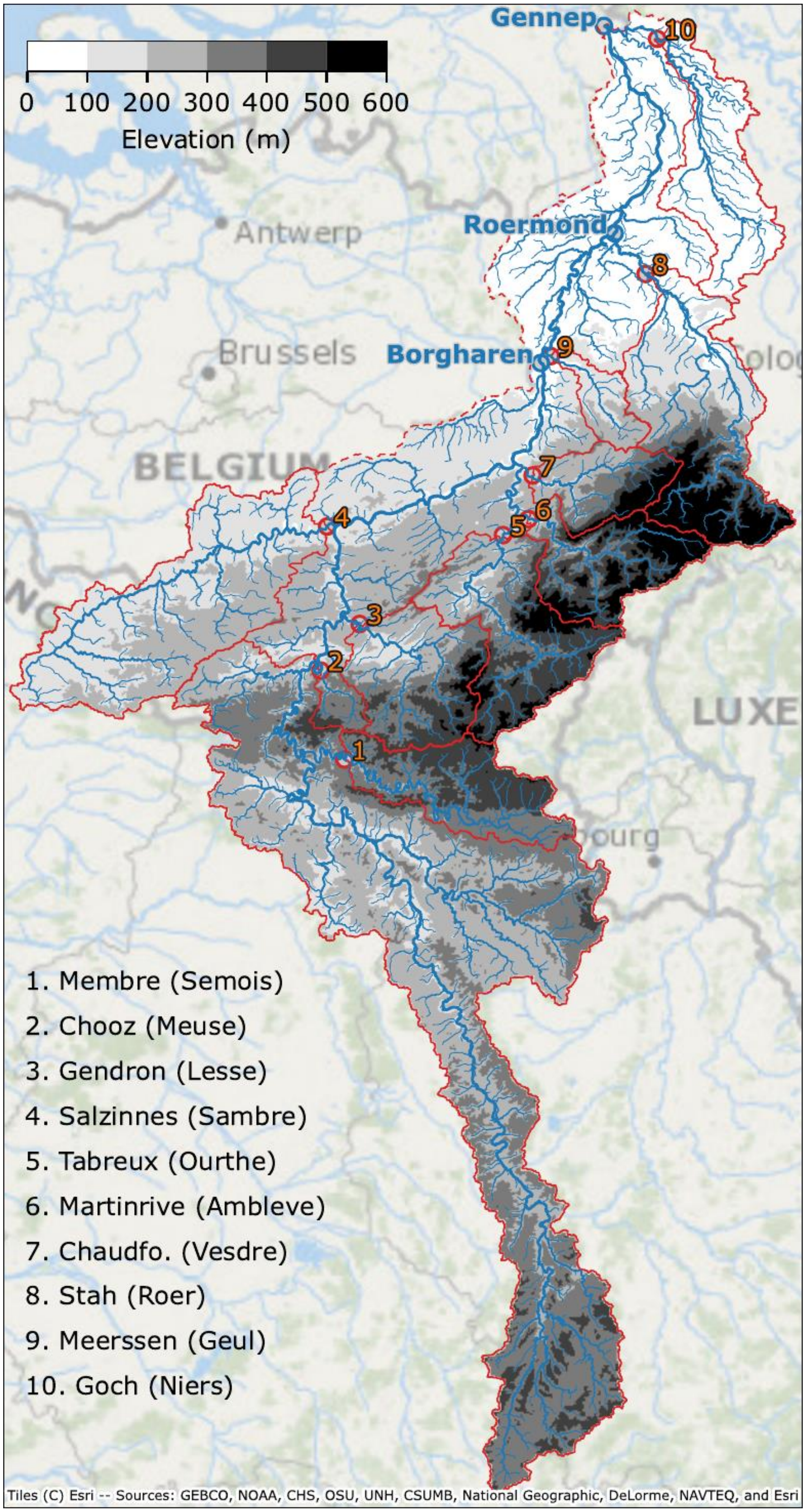


Figure 1: Map of the Meuse catchment considered in this study, with main river, tributaries, streams, and catchment bounds.

Elevation is shown with the grey-scale. ~~Data for this~~Elevation data were obtained from EU-DEM ~~((~~(Copernicus Land Monitoring Service 2017~~))~~) and used to derive catchment delineation and tributary steepness. ~~More~~These data were provided to the experts together with other hydrological characteristics, ~~such as~~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~~)), subsoil (~~)

- *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 annual recurrence intervals: 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 ~~statistics (Copernicus Land Monitoring Service 2020) and hydrograph shapes (from discharge data, see~~patterns and hydrographs for all catchments/tributaries during the 10 largest discharges measure at Borgharen (sources described below~~), that were provided to the experts, are shown~~).

This information, included in the supplementary information~~. This information~~, was provided to the experts to ~~help~~support them ~~make~~in making their estimates. The discharge data needed ~~for fitting~~to fit the model to the observations were obtained from (Service public de Wallonie 2022) for the Belgian gauges, (Waterschap Limburg 2021; ~~**Waterinfo?)**~~Rijkswaterstaat 2022) 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 ~~qualitative~~normalized form as hydrograph shapes.

3 Method for estimating extreme discharges with experts

3.1 Probabilistic model~~ for estimating extreme discharges~~

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

$$Q_d = \cancel{f_{\Delta t,u} \cdot \sum_{j} Q_u} f_{\Delta t} \cdot \sum_{u} Q_u,$$

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. ~~With the factor $f_{\Delta t,u}$ experts can compensate for inaccuracies in estimation. These are, for example, the time difference between discharge peaks and peak attenuation as the flood wave travels through the river (resulting in a lower factor), or rainfall on the Meuse catchment area that is not covered by one of the tributaries (leading to a higher factor). The factor can therefore be lower or higher than 1.0. The model is deliberately kept simple, so that the consequences of the experts' estimates remain traceable for them.~~ Location d can be any location along the river where the discharge is assumed to be dependent mainly on rainfall in the upstream catchment. ~~During an event, the tributary peak discharges Q_u are related; a rainfall event will likely span an area larger than the tributary, and therefore cause a lower or higher discharge in multiple tributaries. Additionally, hydrological characteristics of the catchments are similar for neighboring tributaries. These dependencies are modelled with a multivariate Gaussian distribution that is realized through Bayesian Networks estimated by the experts.~~The random variable Q_u is modelled 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 Frechet (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 Sect. 3.3). The factor or ratio $f_{\Delta t}$ in Eq. [eq:main_model] 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), 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 1.0. We elicited the discharges 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) from the experts, as well as the factor $f_{\Delta t}$, using structured expert judgment (SEJ), as described in Sect. 3.2. 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 and Tawn 1996, 467).

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 modelled with a multivariate Gaussian copula that is realized through Bayesian Networks estimated by the experts (Hanea, Morales Napoles, and Ababei 2015). The details of this concern the practical and theoretical aspects of eliciting dependence with experts and are beyond the scope of this article. They will ~~therefore~~ be presented in a separate article that is yet to be published. ~~The~~ We did use the resulting correlation matrices~~,~~ ~~presented in appendix 6.3.0.0.2, are nonetheless used~~ for calculating the discharge statistics in this study. They are presented in appendix 8.

~~Each random variable of this multivariate model is a univariate (marginal) distribution. We use~~The model from Eq. [eq:main_model] was deliberately kept simple to ensure that the ~~generalized extreme value distribution (GEV) as a statistical model~~effect of the experts' estimates on the result remains traceable for ~~extreme discharges (Jenkinson 1955). This family of distributions is used because it is suitable to estimate the probabilities of extreme events and gives flexibility by varying between Frechet, Gumbel and Weibull distributions (i.e., heavy tailed, exponentially tailed, and light tailed, respectively) to fit specific catchment properties.~~them. Section 3.4 explains how downstream discharges ~~are~~were generated from these model components~~,~~ (i.e., the different terms in Eq. [eq:main_model]), including uncertainty bounds.

The model ~~was~~is also described in more detail in (G. Rongen, Morales-Nápoles, and Kok 2022b) as well, where it was used in a ~~solely~~ data-driven context.

3.2 Assessing uncertainties with expert judgment

~~We use~~The experts' estimates are elicited using Cooke's method~~ for~~. This is a structured ~~expert judgment. Cooke's model is a method to elicit and combine~~approach to elicite uncertainty for unknown quantities. It combines expert judgments based on empirical control questions, with the aim to find a single~~,~~ combined~~,~~ estimate for the ~~variable~~variables of interest (rational consensus). ~~The approach~~ Cooke's method is typically employed when alternative approaches for quantifying uncertain variables are lacking or unsatisfying (e.g., due to costs or ethical limitations). It is extensively described in (Roger M. Cooke 1991) while applications are discussed in (Roger M. Cooke and Goossens 2008~~), here~~). Here, we discuss ~~some of~~ the basic elements of the method.

~~The expert makes an uncertainty estimate for each question by estimating a number of percentiles. In this study (and in most others) these are the 5th, 50th and 95th percentile, that are combined into a probability density function (PDF) f_{exp} for each expert for each variable of interest. With expert's answers to seed or calibration variables (unknown to the experts at the moment of the elicitation),~~ We applied Cooke's method ~~assigns a weight to each participating expert. The expert weight, $w_\alpha(e)$, is calculated by multiplying the *calibration* and *information* scores, if the experts calibration score is above the chosen significance level:~~because of its strong mathematical base, track record (Colson and Cooke 2017) and the authors' familiarity with this method.

In Cooke's method, a group of participants, often researchers or practitioners in the field of interest, provides uncertainty estimates for a set of $w_\alpha(e) = 1_\alpha \times$ ~~calibration score(e) × information score(e).~~

~~The calibration score is calculated from the~~ questions~~ for which the answer is known by the researcher~~. These can be divided into two categories; seed and target questions. Seed questions are used to assess the participants' ability to estimate uncertainty within the context of the study. The answers to these questions are known by the researchers but not by the participants at the moment of the elicitation. ~~These are referred to as seed or calibration variables. Calibration is a measure of~~Seed questions are often sourced from similar studies or cases unknown to the participants. They are as close as possible to the variables of interest and in any case related to the field of expertise of the participant pool. Target questions concern the variables of interest, for which the answer is unknown to both researchers as participants.

The uncertainty for each item is expressed by estimating percentiles (rather than a single value), from which two scores are calculated, the *statistical accuracy* ~~of the expert. The information score expresses the precision of an expert's answers.~~and *information* scores. Typically, the 5th, 50th, and 95th percentiles are elicited. This creates a probability vector with 4 inter-quantile intervals, $p = (0.05,0.45,0.45,0.05)$. The statistical accuracy is calculated by comparing the inter-quantile probability p_i to $s_i(e)$, the fraction of realizations within expert e’s inter-quantile interval. The score is based on the relative information $I(s(e)|p)$, which equals $\sum_{i=1,...,4} s_i \log(s_i/p_i)$. In case of 20 questions, the statistical accuracy is highest if the expert overestimates 1 seed question (i.e., the actual answer was below the 5th percentile), underestimates 1 question, and has 9 questions in both the [5%, 50%] and [50%, 95%] interval. This would result in s equaling p and ratios of 1. The farther away these interquantile ratios are from 1, the lower the statistical accuracy. Note that the maximum statistical accuracy is not achieved when all answers are close to the median, but it would give a high score nonetheless. The information score measures the degree of uncertainty of an expert’s answers compared to other experts. Percentile estimates that are close together (compared to the other participants) are more ~~precise, more~~ informative~~,~~ and get a higher information score. The product of the statistical accuracy and information score gives the expert’s weight $w_\alpha(e)$:

$w_\alpha(e) = 1_\alpha \times$ ~~In this study, the seed variables are the discharges that are exceeded on average once per 10 years, according to the measurements. 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?". Discharges with a 10 year recurrence interval are exceptional, but can in general still be well approximated from the available data. The information score is calculated from all questions, and is a measure of the degree of uncertainty of the experts answer. The decision maker (DM) is a~~$\text{statistical accuracy}(e) \times \text{information score}(e)$.

Experts with a statistical accuracy lower than α can be excluded from the pool by using a threshold, expressed by the 1_α in Eq. [eq:cookes]. This threshold is usually 5%. The (weighted) combination of the experts’ estimates~~.~~ is called the decision maker (DM). The ~~expert contributes~~experts contribute to the ith item’s DM by ~~the assigned weight, the product of the calibration score and information score~~their normalized weight:

$$\mathrm{DM}_\alpha(i) = \sum_e w_\alpha(e) f_{e,i} / \sum_e w_\alpha(e).$$

This is called the global weight (GL) DM. Alternatively, the experts can be given the same weight, which results in the equal weight (EQ) DM. This does not require eliciting seed variables, but neither does it distinguish experts based on their performance, a key aspect of Cooke’s ~~method. Other methods for expert elicitation could have been used as well. A well-known alternative to Cooke’s method is the Delphi method (Brown 1968), in which experts estimate and discuss in rounds, until consensus is reached on the estimates. Another option is a Bayesian approach, as described in Hartley~~

~~and French (2021). We chose Cooke's method because of its strong mathematical base and track record (Colson and Cooke 2017), and the authors' familiarity with this~~ method.

~~We~~ To construct the DM, probability density functions (PDFs) such as $f_{e,i}$ in Eq. [eq:DM], need to be created ~~PDFs~~ from the ~~experts' quantile~~ percentile estimates ~~by fitting a~~. We used the Metalog distribution ~~trough the percentiles~~for this (Keelin 2016). This distribution ~~allows to fit~~ is capable of exactly fitting any three~~-~~-percentile estimate ~~without changing the estimates.~~. For symmetric estimates, ~~the distribution~~ it is bell-shaped. For asymmetric ones, it becomes left- or right~~-~~-skewed. ~~Normally, in~~Typically, Cooke's method~~, the PDF is created by assuming~~ assumes a uniform distribution in between the percentiles (minimum information). This leads to a ~~piece-wise linear cumulative distribution,~~ stepped PDF where the Metalog gives a smooth ~~fit~~PDF. An example of using the Metalog distribution in an expert elicitation study ~~was~~is described by (Dion, Galbraith, and Sirag 2020). All calculations related to Cooke's method were performed using the open-source software ANDURYL (Leontaris and Morales-Nápoles 2018; Hart, Leontaris, and Morales-Nápoles 2019; Guus Rongen et al. 2020).

~~7~~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 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 ~~where~~were discussed before ~~making~~ the assessments to make sure that the ~~study~~ concepts and questions were clear. After this, an ~~training~~ exercise for the Weser catchment was done in which the experts ~~needed to answer~~ answered four questions that were subsequently discussed ~~afterwards. This~~, In this way, the experts could ~~see how~~compare their answers ~~compared~~ to the realizations~~,~~ and ~~subsequently what their~~view the resulting scores ~~in~~using Cooke's method ~~were.~~ .

Apart from the training exercise, the experts answered 26 questions~~,~~: 10 seed questions ~~for~~regarding the 10-year discharge ~~that is exceeded on average once per 10 years (1~~(one for each tributary), 10 target questions ~~for~~, regarding the 1,000-year discharge ~~exceeded on average once per 1000 years~~, and 6 target questions for the ~~factors~~ratios between upstream sum and downstream discharge (~~2 return periods × 3 locations). Note that we will use the shorthand~~ 10-year and 1,000-year ~~notation in the remainder of this article, when indicating the 'return period' of average recurrence interval of discharges.~~, for three locations). A list of the ~~7 participants and~~seven participants' names, their affiliations, and their field of expertise is shown in ~~table ~~Table [tab:experts]. ~~The alphabetic order in which the~~While the participants are pre-selected on their expertise, experts are ~~listed holds no relation to the number in which~~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 ~~are labelled~~in the table does not correspond to their labels in the ~~analysis carried out~~results. An overview of the data provided to the participants is given in ~~this article~~Sect. 2, while the data itself, as well as the questionnaire, are presented in the supplementary information.

Name	Affiliation	~~Specialism~~Field of expertise
Alexander Bakker	Rijkswaterstaat & Delft University of Technology	Risk analysis for storm surge barriers, extreme value analyses, climate change and climate scenario's.
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, focussed on hydrologic modelling 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

3.3 Determining model coefficients with Bayesian inference

~~We~~The model for downstream discharges (Eq. [eq:main_model]) is quantified using Bayesian inference. Firstly, because Bayesian methods explicitly incorporate uncertainty, which is a key aspect of this study. Secondly, because these methods provide a natural way to integrate expert judgment with data. Because the experts do not know the exact values of the discharges they are estimating, their estimates can be seen as prior information. This can subsequently be updated to a posterior distribution using available data. Note that the experts did not estimate the prior distributions of the GEV-parameters directly but they estimated the 10-year and 1000-year discharge from which the parameters were estimated.

To evaluate the performance of the combined data and EJ approach, we compared it to using each of these sources of information individually. Hence, we distinguished three approaches~~ for fitting the model from Eq. [eq:main_model],~~:

- ~~the~~The 'data-only' approach, ~~in which~~utilizing only measured discharges (the annual maxima per tributary that lead to a peak discharge at Borgharen~~) were used,~~.)
- ~~the~~The 'EJ-only' approach, ~~in which only~~solely relying on the expert's estimate for the 10-year and 1,000-year discharge~~ is used, and~~.
- ~~the~~The combined 'data and EJ' approach, ~~in which~~combining the measured discharges ~~are combined~~ with the expert estimate for the 1,000-year discharge (~~not~~excluding the 10-year discharge).

~~The fitting~~ A probability distribution for extreme discharges was fit for each one of the three ~~options is performed~~approaches using ~~the Bayesian inference technique~~ Markov-Chain Monte Carlo (MCMC~~). This results in an inference~~), a Bayesian inference technique commonly used to sample from a PDF. The MCMC algorithm generates a trace of ~~the~~distribution parameters ~~of the GEV distribution (i.e., the PDF for the tributary discharges). This~~ . After removing the spin-up and thinning it to remove autocorrelation, the trace ~~is~~becomes the empirical joint probability distribution of, in our case, the three GEV model parameters ~~(the GEV has three parameters) that is~~ for each tributary. These are subsequently used ~~for sampling~~to calculate the downstream discharges (see Sect. 3.4). ~~MCMC is the name commonly~~The Python module 'emcee' (Foreman-Mackey et al. 2013) was used for Bayesian inference. This module implements the affine invariant MCMC ensemble sampler as described by Goodman and Weare (2010).

The MCMC procedure relies on a ~~group~~likelihood-based criterion to assess the goodness of ~~algorithms used to sample from~~fit for a ~~PDF. One example~~proposed combination (θ) of ~~such algorithms~~the GEV distribution parameters, comprising the location (μ), scale (σ), and shape parameter (ξ). Consider $z = (x - \mu)/\sigma$. The probability density function (PDF) of the GEV is then,

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

This posterior likelihood function consists of three parts. The first component is the ~~Metropolis-Hastings algorithm (Hastings 1970).~~prior likelihood of the GEV-parameters. We prefer this to be weakly informative (i.e., uninformative, but within bounds that ensure a stable simulation), such that only the data and expert estimates inform the final result.

However, we did add an informative prior to the shape parameter (ξ) for two reasons. Firstly, when only using expert estimates and no data, two discharge estimates are not sufficient for fitting the three parameters of the GEV-distribution. Secondly, the standard deviation of the shape-parameter decreases with increasing number of years (or other block maxima) in a time series (Papalexiou and Koutsoyiannis 2013). Our 30 to 70 annual maxima per tributary are not sufficient to reach convergence. Therefore, we employ the geophysical prior as presented by Martins and Stedinger (2000); a beta distribution with $\alpha = 6$ and $\beta = 9$ for $x \in [-0.5, 0.5]$, for which the PDF is:

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

with $x = \xi + 0.5$, and Γ being the gamma-function. This ~~Bayesian technique combines *a-priori* distributions with observations to estimate *a-posteriori* distributions. We used a weakly informed prior for the GEV distribution, which is described in Appendix 6.0.0.0.1 in more detail. These were then updated with observations, with expert estimates, or~~ PDF is slightly skewed towards negative values of the shape parameter, preferring the heavy tailed Frechet 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.

We assigned equal probability to all values of μ and σ 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). Together with ~~both~~the beta-distribution for ξ, the prior likelihood function $\pi(\theta)$ equals $f_\beta(\xi + 0.5)$ for $-0.5 < \xi < 0.5$, $\sigma > 0$, and $\mu > 0$. $\pi(\theta) = 0$ for any other combination.

~~Just like most curve fitting algorithms, MCMC uses a (log-) likelihood-based criterion to find a good fit. Figure~~ After the prior, the second and third part of the posterior likelihood function are the likelihood of a GEV given the observed discharges and expert estimates. Figure 2 illustrates this. The top curve $f(Q|\theta)$ ~~shows how the likelihood of a specific GEV distribution is calculated, based on the observations and expert estimates. This calculation is implemented as a custom likelihood function in the used Python-package for Bayesian statistical modelling PyMC3 (Salvatier, Wiecki, and Fonnesbeck 2015).~~

represents a proposed GEV-distribution for the random variable Q (tributary peak discharge) with parameter vector θ.

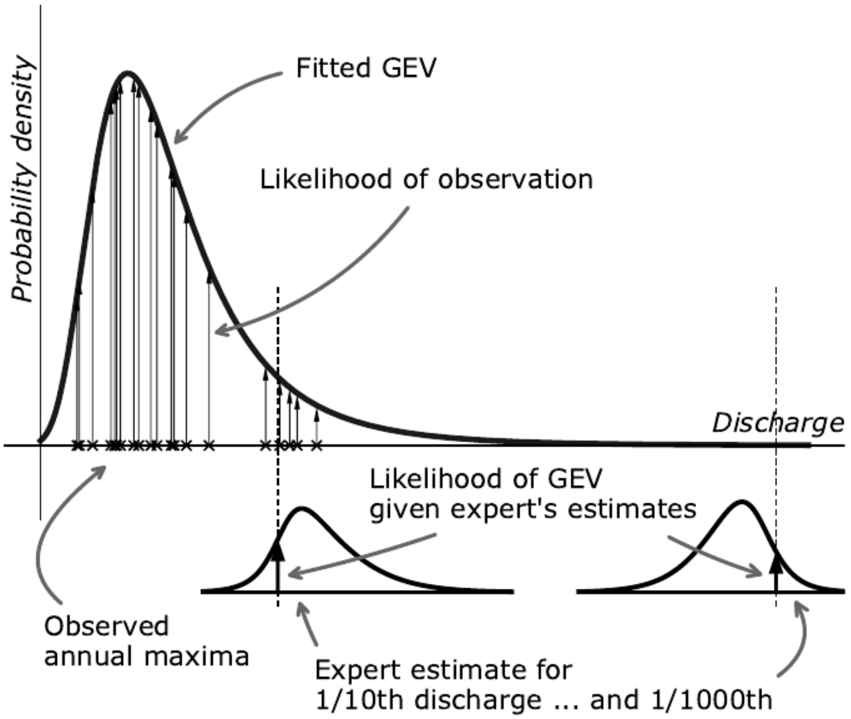


Figure 2: Conceptual visualization of elements in the likelihood-function of a tributary GEV-distribution.

~~The top curve $f(Q|\theta)$ represents a fitted GEV-distribution with parameter vector θ.~~ The log-likelihood of ~~the parameters that give this specific GEV can be~~ θ is calculated ~~with~~as the ~~product~~sum of the log-probability density ~~function~~ of ~~the observations (i.e., the product of~~each observation q given θ, or,

$$\ell(\theta|\mathrm{q}) = \sum_i \log\left(f(q_i|\theta)\right).$$

$f(q_i|\theta)$ corresponds to the length of the arrows in ~~the figure):~~

~~$\ell(\theta|q) = \sum_i \log\left(f_\theta(q_i)\right).$~~

~~The best fit of the curve is the set of parameters θ for which the log-likelihood ℓ, given the observations, q is maximal. The MCMC sampling algorithm gives a (joint) probability density function of~~ 2. The log-likelihood of θ~~, instead of a single value as a maximum likelihood estimate would.~~

~~The log-likelihood of~~ given the expert's ~~estimate~~estimates is calculated and added to the total likelihood, in a similar way as ~~follows:~~

described by Viglione et al. (2013): Given a GEV-distribution ~~$f_\theta(Q)$~~ $f(Q|\theta)$, the ~~discharges~~discharge q for ~~one or two~~a specific annual exceedance ~~probabilities are~~probability p is calculated ~~(1/~~from the inverse CDF,

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

~~1.~~ 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 ~~and 1/~~- or 1,000 ~~for EJ-only, 1/1,000 for data and EJ combined).~~-year discharge, $g\left(q_{p_j}\right)$. These ~~discharges correspond to an average recurrence interval of 10 and 1,000 year.~~

~~1. The expert is asked for an estimate of the discharge~~ estimates are illustrated with ~~this exceedance frequency resulting in the distribution f_{exp}. These estimates are displayed by~~ the curves on the bottom of Fig. 2.

The likelihood of the GEV~~-quantile~~ according to the expert or DM can then be calculated ~~in the expert estimated distribution:~~with,

~~$\ell(\theta|exp) = \sum_j \log\left(f_{exp,j}\left(q_{p_j}\right)\right)$~~

~~$= \sum_j \log\left(f_{exp,j}\left(F_\theta^{-1}(1 - p_j)\right)\right)$~~

~~where p_j is the exceedance probability for quantile j, and q_{p_j} the discharge corresponding to that exceedance probability based on the GEV-distribution f_θ. F_θ is the cumulative distribution function (CDF), so its inverse, F_θ^{-1}, is the quantile function.~~

~~By summing the likelihood in equations [eq:obs_likelihood] and [eq:exp_likelihood], the likelihood of the distribution given both the observations and expert opinions is calculated. This is an unbalanced sum because there are many more observations contributing to that part of the likelihood, than there are expert estimates (only one when combining data and EJ). We therefore add a factor $10/N_{obs}$ to the likelihood function that weights the observations as if only 10 were measured. Note that 10 is still much more than the single expert estimate. The estimate is however made for the 1,000-year discharge, which gives it more weight due to the cantilever effect (i.e., a small parameter variation results in a large difference for the q_{1000}, and therefore the $\ell(\theta|exp)$ as well). We found that 10 gave a good balance between observations and expert estimates but we realize that it is somewhat subjective. Appendix 6.3.0.0.1 shows a sensitivity analysis of the MCMC-fit to substantiate the choice for this factor. The complete likelihood function that is used to fit a distribution to both data and expert estimates, is:~~

~~$$\ell(\theta|q,exp) = \frac{10}{N_t} \cdot \sum_i \log\left(f_\theta(q_i)\right) + \log\left(f_{exp}\left(F_\theta^{-1}(0.999)\right)\right)$$~~

~~For the factor between the tributaries' sum and the downstream discharge~~

$$\begin{aligned} \ell(\theta|\mathrm{g}) &= \textstyle\sum_j \log\left(g\left(q_{p_j}\right)\right) \\ &= \textstyle\sum_j \log\left(g\left(F_\theta^{-1}(1-p_j)\right)\right). \end{aligned}$$

By summing the log-likelihood in equations [eq:prior_likelihood], [eq:obs_likelihood], and [eq:exp_likelihood], we get the total posterior likelihood function:

$$\ell(\theta|\mathrm{q},\mathrm{g}) = \log\left(\pi(\theta)\right) + \log\left(f(\mathrm{q}|\theta)\right) + \log\left(\mathrm{g}\left(F^{-1}(1-\mathrm{p}|\theta)\right)\right)$$

The posterior likelihood comprises the prior likelihood, the likelihood of the observations, and the likelihood of the expert judgment. If only data are used, the last term drops out. If only expert judgments are used, the second term drops out, and the last term contains two expert estimates. If both data and expert judgment are used, the last term contains only a single expert estimate. Equation [eq:combined_likelihood] is used to compare the likelihoods of specific proposed parameter-combinations in the MCMC-sampling. Notice that the expert judgment term does not take into account any information in the observed discharges q and can therefore be considered prior information.

With the procedure summarized in Eq. [eq:combined_likelihood], the probability distributions for the tributary discharges (Q_u in Eq. [eq:main_model]) 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 ~~responds~~corresponds to a practical choice for a distribution of positive values with sufficient shape flexibility. The experts estimated a distribution for the factor as well, which ~~is~~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. ~~In between, the factor was~~Values of $f_{\Delta t}$ for return periods T greater than 10 were interpolated~~, and for recurrence intervals larger than 1,000~~ (up to 1000-years~~, the factor was~~) or extrapolated~~.~~,

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

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

~~Finally, for~~For the correlation matrix describing the dependence between tributary extremes, ~~we have~~ the observed correlations~~,~~ were used for the data-only option~~,~~ and the expert-estimated correlations~~, used~~ for the expert-only option. For the combined option, we ~~simply take~~took the average of the observed correlation matrix and the expert-estimated correlation matrix. Other possibilities for combining correlation matrices are available (~~e.g.,~~see for example Al-Awadhi and Garthwaite 1998, for a Bayesian approach), however an in depth research of these ~~go~~options are beyond the ~~goal~~scope of this study~~. Notice also that in this study we investigate a "50-50" contribution from data and experts~~.

3.4 Calculating the downstream discharges

~~With the fitted model components described in Section 3.1, we calculated the discharge at a downstream location. To calculate a single exceedance frequency curve, we used a Monte Carlo approach in which the following samples for the $N_T = 9$ tributaries, $N_Q = 10{,}000$ discharge events, and $N_M = 8{,}000$ MCMC parameter combinations were combined:~~

~~3. N_T draws from the dependence model transformed to the $[1, N_M]$ interval. These are used as index to pick a GEV parameter combination from each tributary's inference trace after these sorting these based on the once per 1,000 year discharge. Doing this leads to relatively similar (1,000-year) discharges for tributaries with a strong dependence.~~

~~4. $N_T \times N_Q$ draws from the dependence model. These events (on a standard normal scale) are transformed to the discharge realizations for each tributaries GEV parameter combination.~~

~~N_Q draws from the factor between~~ The last section explained the three quantification alternatives of the components from Eq. [eq:main_model], needed to calculate the downstream discharges. These components are:

- Tributary (marginal) discharges, represented by the GEV-distributions from the Bayesian inference.

- The interdependence between tributaries, represented by a multivariate normal copula.

~~5.~~ The ratio between the upstream sum and downstream ~~discharge, used to multiply the sum of the upstream discharges to get the downstream discharge.~~

- ~~Assigning exceedance frequencies to the N_Q~~ discharges ~~using plot positions gives an exceedance frequency curve (plot positions from (Bernard and Bos-Levenbach 1955) were used). Repeating this procedure 2000 times, and calculating the percentiles per exceedance frequency, results in the uncertainty intervals of the exceedance frequency curves.~~ ($f_{\Delta t}$).

~~Note that in the sampling approach, the correlation model is not only used to model tributary dependence within individual events but also to model the dependence between different tributaries' GEV parameter combinations. With this we express our assumption that the correlation between tributaries is present as well in the uncertainty intervals. After all, the uncertainty in the fitted distributions is mainly the result from the largest observed events. Having a time series with (or without) high discharge(s) in multiple tributaries affects the fitted distributions in a similar way for these different tributaries.~~

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. Appendix 7 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. 1. However, the Semois catchment is 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 plotting positions, resulting in an exceedance frequency curve. This process is repeated 10,000 times with different GEV-realizations from the MCMC-trace. This results in 10,000 curves (each based on 10,000 discharges), from which the uncertainty bandwidth is determined. This is illustrated in Fig. 3. The grey lines depict 50 of the 10,000 curves (these can be both tributary GEV-curves, or downstream discharge curves). The (blue) histogram gives the distribution of the 1,000-year discharges. The colored 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.

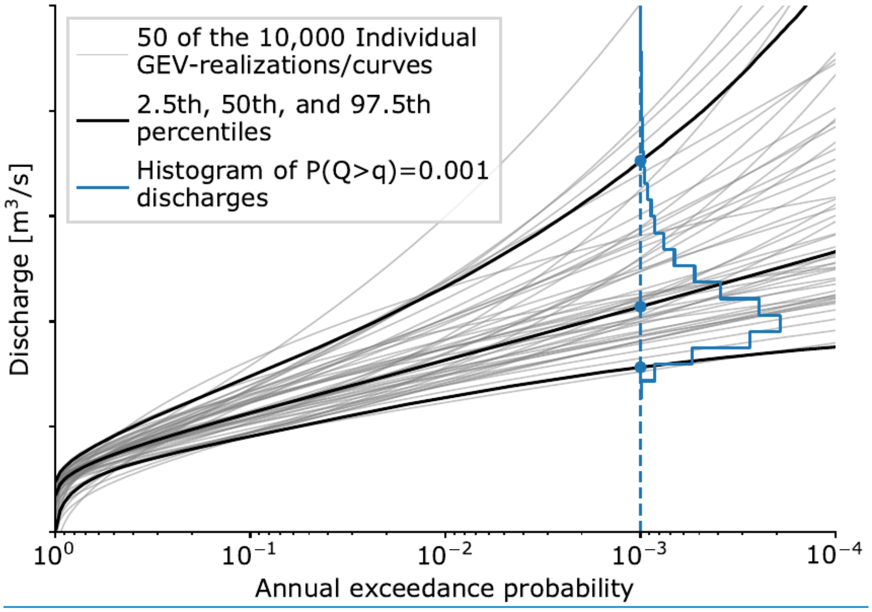


Figure 3: Individual exceedance frequency curves for each GEV-realization or downstream discharge, and the different percentiles derived from these.

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 x 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 inverse CDF). 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. 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 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 neighbouring tributary as well. Consequently, the neighbouring 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 blue 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.

4 Experts' performance and resulting discharge statistics

~~In this~~This result section~~, we~~ first ~~present~~presents the ~~expert~~experts' scores for Cooke's method~~,~~ (Sect. 4.1) and the experts' rationale for answering the questions~~.~~ (Sect. 4.2). After this, the extreme value results for the tributaries (Sect. 4.3) and downstream locations (Sect. 4.4) are presented.

4.1 Results for Cooke's method

The experts ~~estimate~~estimated three-percentiles (5th, ~~50eth~~50th and 95th) for the 10- and 1,000-year ~~discharges,~~discharge for all larger tributaries in the Meuse catchment. ~~The 10-year estimate is used for calibration.~~ An overview of the answers is given in the supplementary material. Based on these estimates, the scores for Cooke's method are calculated~~. The results~~ as described in Sect. 3.2. The resulting statistical accuracy, information score, and combined score (which, after normalizing, become weights) are shown in table 1.

Scores for Cooke's method, for the experts (top 7 rows) and decision makers (bottom 3 rows).

	~~Calibration score~~Statistical accuracy	Information score		Comb. score
		All	~~Calibr.~~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 ~~calibration score, a measure for the expert's~~ 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. Figure ~~3~~4 shows the position of each realization (answer) within the experts' ~~estimates. There were 10~~

~~calibration variables assessed~~three-percentile estimate for each ~~with three percentiles.~~of the 10-year discharges. A well calibrated score would capture the realizations to these seed variables accordingly to (or as close to) the mass in each inter-quantile bin~~. Thus,~~: 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 accuracy score. A concentration of dots on both ends indicates overconfidence (too ~~narrow~~close together estimates, resulting in realizations outside of the 90% bounds~~.)~~). We ~~can see~~observe that most experts tend to underestimate the measured discharges, since most realizations are higher than their estimated 95th percentile.

The information scores show less variation, as is usually the case. The expert with the highest calibration (or statistical accuracy) score (expert D) also has the lowest information score. Expert E, who has a high ~~calibration~~statistical accuracy as well, ~~provided~~estimated more concentrated ~~answers~~percentiles, resulting in a higher information score.

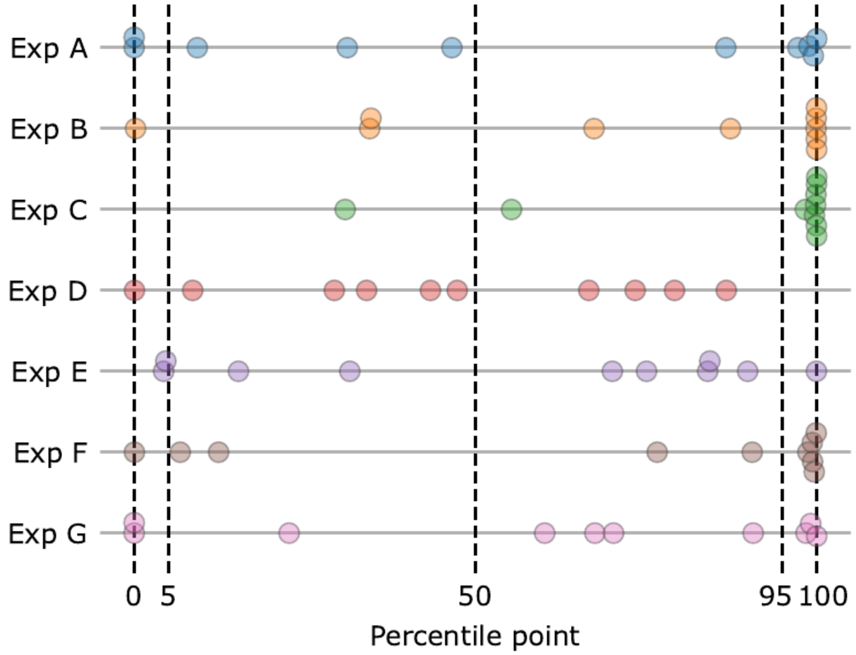


Figure ~~3~~4: Seed questions realizations' compared to each expert's estimates. The position of each realization is displayed as percentile point in the expert's distribution estimate.

The variation ~~in~~between the three decision makers (DMs) ~~shown~~in the table is limited. Optimizing the DM (i.e., excluding experts based on ~~calibration score~~statistical accuracy to improve the DM-score) has a limited effect~~: Only~~. In this case, only expert D and E ~~remain~~would have a non-zero weight, resulting in more or less the same results ~~as when~~compared to including all experts, even when some of them contribute with ~~"marginal"~~'marginal' weights. The equal weights DM ~~results~~ in ~~a good~~this case results in an outcome~~;~~ that is comparable to that of the performance based DM, i.e., a high ~~calibration score~~statistical accuracy with a slightly lower information score compared to the other two DMs.~~The item weights DM, which allocates more weight to precise answers, results in the same DM calibration score as the global weights DM. We chose not to use it as it assumes more precise (i.e., confident) answers are better, something we did not want to assume for this case study.~~

We present the model results ~~by fitting it to~~as discussed earlier through three cases i) only data, ii) only expert estimates, and iii) the two combined as described in Section 3.3. We used the global weights DM for the data and experts option (iii). 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 ~~calibration~~seed questions~~, a situation that could ultimately be used when~~ and the corresponding differential ~~weighting of expert judgments is not considered~~weights.

4.2 Rationale for estimating tributary discharges

We asked the experts to briefly describe the procedure they followed for making their estimates. From their responses we distinguished three approaches. The first was making, or thinking, of a simple conceptual hydrological model, in which the discharge follows from catchment characteristics like (a subset of) area, rainfall, evaporation and transpiration, ~~precipitation-hydrograph~~rainfall-runoff response, land-use, subsoil, slope, or the presence of reservoirs. Most of this information was provided to the experts, and if not so, they made estimates for it themselves. A second approach that experts followed was to compare the catchments to others that are 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 Sect. 2) was provided to the experts in spreadsheets as well, making it easier for them to do computations.

However, the time frame of one day (for the full elicitation) limited the possibilities for making detailed model simulations.

Figure ~~4~~5 gives an impression of how the different approaches ~~lead~~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. From the figure we can see 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.

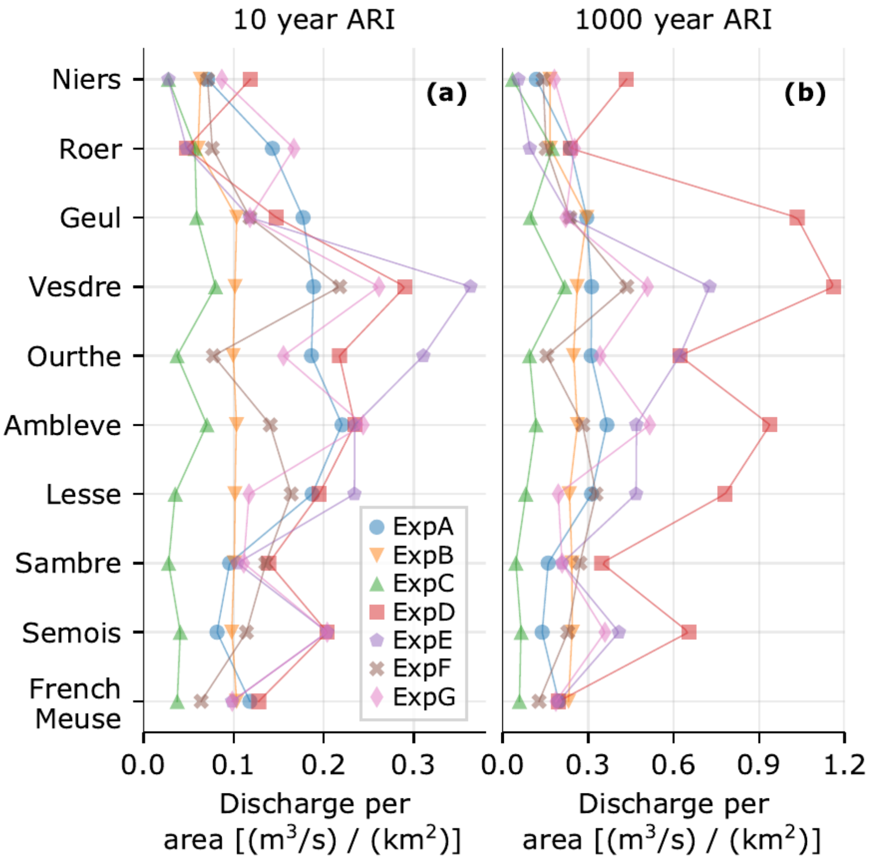


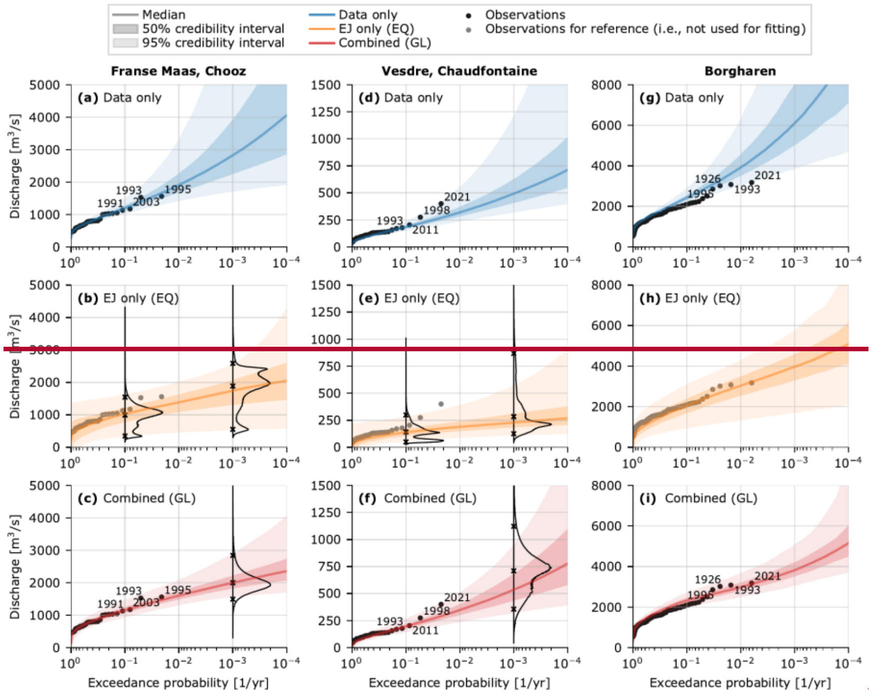
Figure ~~4~~5: Discharge per area for each tributary and experts, based on the estimate for the 50th percentile. ***(a)*** *for the 10-year, and* ***(b)*** *for the 1,000-year discharge. The lines are displayed to help distinguish overlapping markers.*

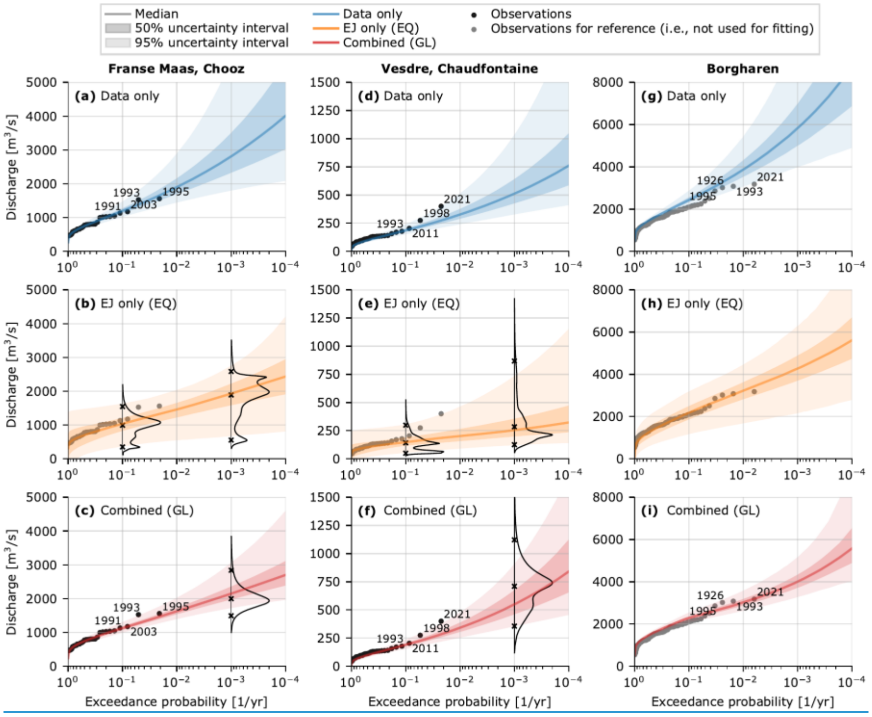
For estimating the factor between the tributaries' sum and the downstream discharge~~,~~ ($f_{\Delta t}$ in Eq. [eq:main_model]), experts mainly took into consideration that not 100% of the area is covered by ~~(i.e., it flows through) the locations~~the tributary catchments for which the discharge-estimates were made, and that there is a time difference between the downstream 'arrival' of the tributary peaks. Additional aspects noted by the experts were the effects of flood peak attenuation and spatial dependence between tributaries and rainfall.

~~The last part of the estimations required in order to calculate discharges at Borgharen is the dependence structure (correlations) between the tributaries. Experts estimated a correlation matrix for this by using a Non-parametric Bayesian Network. The resulting correlation matrices are shown in appendix 6.3.0.0.2.~~

4.3 Extreme discharges for tributaries

~~Based on the procedures described in this paper including experts' estimates we~~We calculated the extreme discharge statistics for each of the tributaries~~.~~ based on the procedures described in Sect. 3.3. Figure [fig:extreme_discharges_Borgharen] shows the results for Chooz and Chaudfontaine (left and middle column~~), a larger not to steep tributary, and a smaller steep tributary.~~). Chooz is a larger not too steep tributary, while Chaudfontaine is a smaller steep tributary (see figure 1). The right column shows the discharges for Borgharen, the location where we want to estimate the discharges through Eq. [eq:main_model], which is further discussed in Sect. 4.4. The results for the other tributaries are shown in the supplementary information for all experts and DMs.





The top row (a, d, g) in Fig. [fig:extreme_discharges_Borgharen] shows the uncertainty interval of these distributions when fitted only to the discharge measurements. The outer colored area is the 95% interval, the more opaque inner~~, darker,~~ 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 return period range where data are available, but the uncertainty interval grows for return periods further outside sample. ~~Opposite to this~~Contrarily, the experts-only option shows much more variation in the ~~“~~‘in ~~sample"~~sample’ range, while the out of sample return periods are more constrained. The combined option is accurate in the ~~“~~‘in ~~sample"~~sample’ range, while the influence of the DM estimates is visible in the 1,000 year return period range.

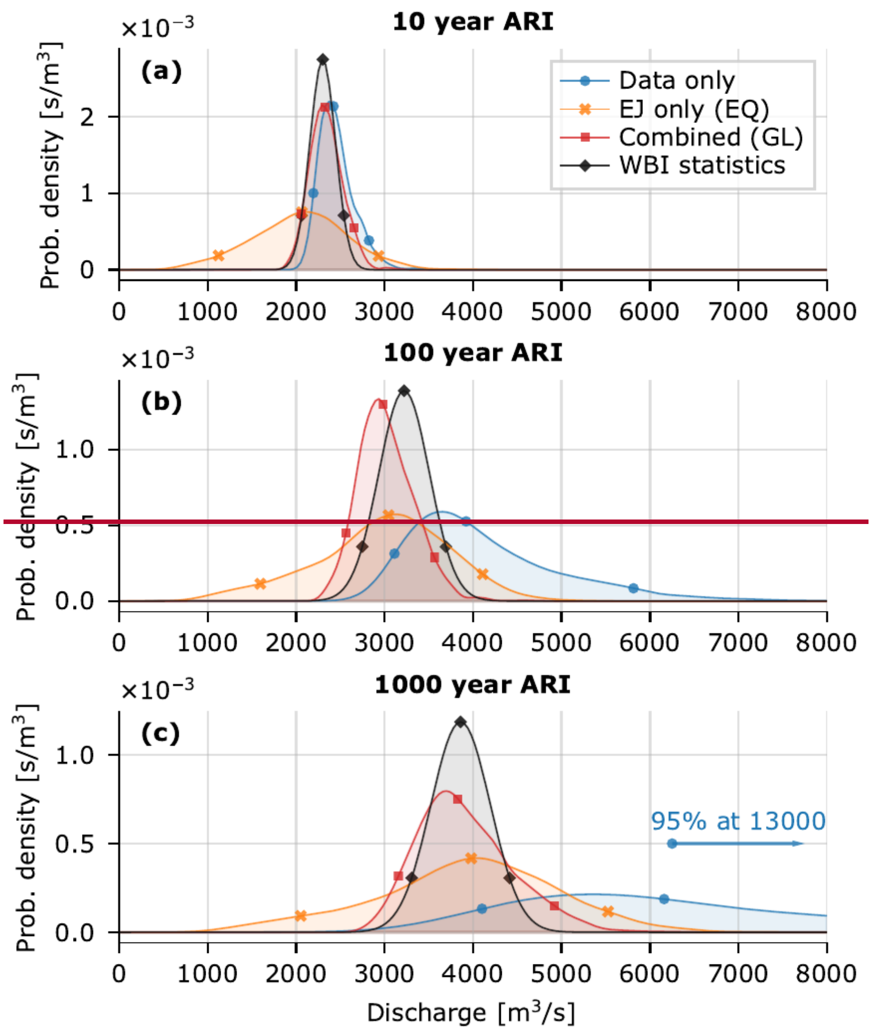
4.4 Extreme discharges for Borgharen

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. [fig:extreme_discharges_Borgharen] (g, h, i).

As with the statistics of the tributaries, we observe high accuracy for the data-only estimates in the ‘in ~~sample~~sample’ range, constrained uncertainty bounds for EJ-only in the range with higher return periods, and both when combined. ~~The data-only results deviate from the observations for frequent events as well. This is the result of the wide uncertainty in the marginal~~

~~distributions; a single large event at one of the larger tributaries (e.g. the French part of the Meuse) can by itself cause a large discharge event. Sampling from these wide uncertainty bounds will therefore (too) often result in a high discharge event. The combined results match the historical observations very well.~~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 equation [eq:main_model]. ~~The EJ-only estimate gives a much wider uncertainty estimate, but the median ('best estimate') matches the observations surprisingly well given that the model was not directly fitted to the data~~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 modelling. 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 ~~5~~6 shows the (conditional) probability distributions (smoothed with a kernel density estimate) for ~~the~~these discharges at ~~this~~the location of interest.


10 year ARI
×10⁻³
(a)
Prob. density [s/m³]
0
1
2
Data only
EJ only (EQ)
Combined (GL)
WBI statistics
0
1000
2000
3000
4000
5000
6000
7000
8000
100 year ARI
×10⁻³
(b)
Prob. density [s/m³]
0.0
0.5
1.0
1000 year ARI
×10⁻³
(c)
Prob. density [s/m³]
0.0
0.5
1.0
95% at 13000
Discharge [m³/s]

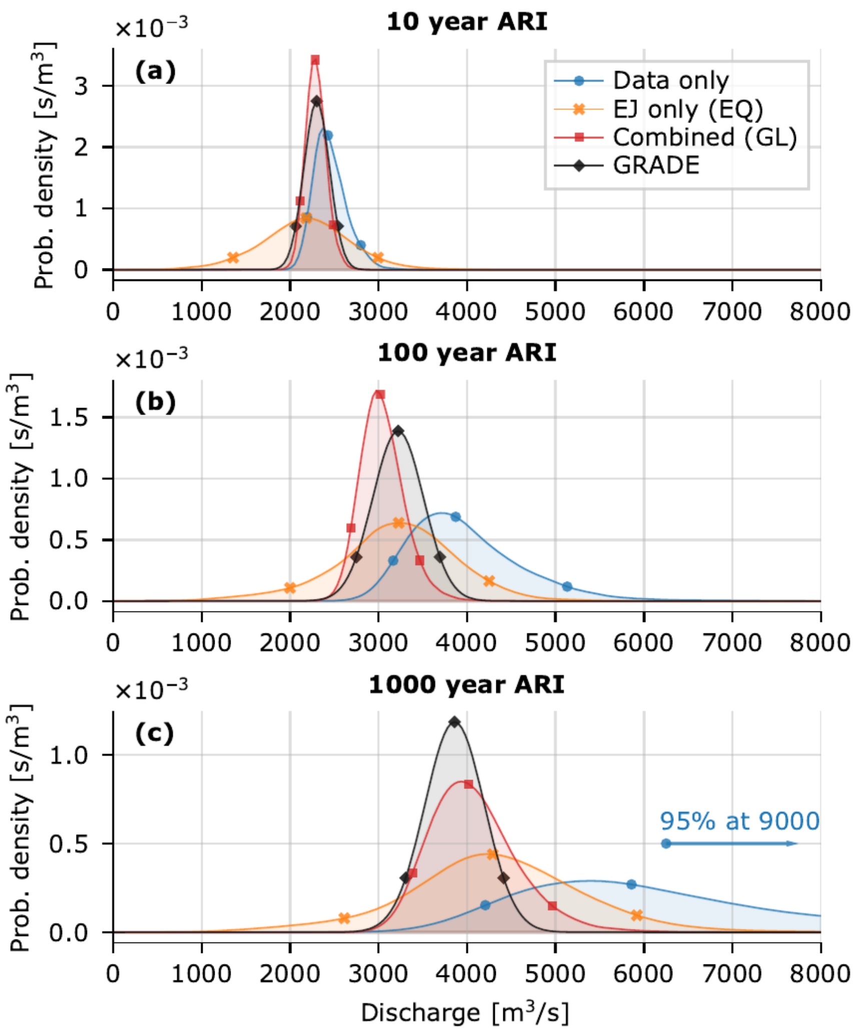


Figure ~~5~~6: Kernel density estimates for the 10-year ***(a)****, 100-year* ***(b)****, and 1,000-year* ***(c)*** *discharge for Borgharen. The dots indicate the ~~5~~5th, 50th and 95th percentile.*

Comparing the ~~same~~ three ~~variants~~modelling options discussed thus far, we see that the data-only option is ~~much too~~very uncertain, with a 95% ~~credibility~~uncertainty interval of ~~6~~4,000 to around ~~11~~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 (GWF Rongen 2016). For discharges higher than 6,000 m^3/s the ~~simple sum-~~applied model (Eq. [eq:main_model]) should be reconsidered, as the hydrodynamic properties of the system change due to upstream flooding.

The combined results are surprisingly close to the currently used GRADE-statistics for dike assessment; the uncertainty is slightly larger but the median is ~~close~~very similar. The EJ-only results are less precise, but the median values ~~(‘best estimate’)~~ are similar to the combined results and

GRADE-statistics. The large uncertainty is ~~largely~~mainly the results of equally weighting all experts, instead of assigning most weight to ~~expert~~experts D and E (the global weight DM). ~~The experts'~~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. [fig:extreme_discharges_Borgharen] a-f). Figure 6 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. [eq:main_model]), and therefore higher resulting downstream discharges. Overall, the combined effect of data and EJ is more difficult to identify in the downstream discharges (Fig. [fig:extreme_discharges_Borgharen] g-i) than it is in the tributary discharge GEVs (Fig. [fig:extreme_discharges_Borgharen] 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. [fig:extreme_discharges_Borgharen] that illustrate this are presented in the supplementary information. There, the results for the other two downstream locations, Roermond and Gennep~~ are ~~, are presented as well. These results behave similar to those for Borgharen~~,~~ and are therefore not presented here~~. They are however displayed in the supplementary material~~.

5 Discussion

~~The discharge estimates that result from this study show the value of fitting a relatively simple statistical model to both data and expert estimates. The predictive power of such models usually diminishes in the extrapolated range, but this is greatly improved by combining it with expert estimates. The data themselves help to increase the precision in the frequent range and it can point out the statistically accurate experts to improve the extrapolated range.~~

~~To test the model without data, we used an equal weight decision maker and left out the data in the fitting procedure. Note that the equal weight DM is a conservative choice, as the experts' statistical accuracy could still be determined based on another catchment where data are available. While the marginal distributions present wide bandwidths, the final results for Borgharen gave an accurate result, albeit with a large bandwidth (low precision).~~

~~The discharge statistics at Borgharen currently used for dike assessment give a lower and less uncertain range for the 1,000-year discharge (Hegnauer and Van den Boogaard 2016). The estimates given~~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.

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 tests 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 uncertain tail shape parameter is informed through this, as the location and scale parameter can with relative certainty be estimated from data. Alternatively, we could have estimated the tail shape parameter directly (however this is not an observed quantity like discharges are), or estimated a related parameter such as the ratios between, for example, the 10-year and 100-year discharge. Because we are ultimately interested in the 1000-year discharges, we preferred eliciting absolute discharges directly. This is in line with guidelines for structured expert judgment, where eliciting observable quantities is recommended over the elicitation of model parameters which are not necessarily observed. We weight expert judgments based on their performance in estimating 10-year discharges and use this information to combine the experts' 1,000-year estimates with data. This should increase the plausibility of a correct estimate of the shape parameter of the GEV or a ratio of extreme discharges with particular return periods. However, it is almost sure that if experts would have been assessed by their ability to estimate ratios of extreme discharges, 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. 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. This kind of study is however out of the scope of our research. Our research however 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 expert estimated 1,000-year discharges much higher of lower than would be expected from observations. We might considered this an indication that the GEV is not an adequate model to fit to this data. 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 observed in 30 to 70 years. 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's 'identically distributed' assumption. The GEV-distribution does however have sufficient shape flexibility to facilitate substantially higher 1,000-year, 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 Cooke's method scores, as described in Sect. 5.2.

A final remark regarding the model is the 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 in the model estimates would have been a valuable addition, and it would likely be the first addition we would consider. However, 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 effects from our research does not alter our main conclusion which is the possibility of enhancing estimation of extreme discharges through structured expert judgments.

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 Cooke's method and has extensively been discussed (Roger M. 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 Cooke's method. In Sect. 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 and 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~~. This was a choice made in order to still be able to draw conclusions regarding the method without data. These measurements~~, such that we could ~~have helped~~clearly distinguish the effect of observations and 'prior' expert judgments. Moreover, GRADE (at the ~~experts in making "less uncertain" estimates. On the other hand, if~~time) did not include the July 2021 event. If the ~~currently used~~ GRADE statistics ~~would be~~had been derived ~~again including~~with the inclusion of the July 2021 event, it ~~is not unlikely that they~~ would ~~appoint~~likely assign more probability to ~~slightly~~ 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.

~~Using expert judgment to provide answers for a model (like we did) can still give~~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. [fig:extreme_discharges_Borgharen] 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.

5.3 Final remarks on model choices

Finally, we note that using expert judgment to estimates 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 ~~choosing which~~providing specific information we ~~provide,~~steer the participants ~~are steered~~ towards a ~~certain~~specific way of reasoning. Every step in the method~~;~~, such as the choice for a GEV-distribution, the dependence model, or the choice for Cooke's method', affects the end result. By presenting the method and providing background information explicitly, we hope to ~~make~~have made this transparent and show the ~~use~~usefulness of the method for similar applications.

6 Conclusions

This study sets out to ~~estimate~~ establish a method for estimation of statistical extremes through structured expert judgment and Bayesian inference, in a case-study for extreme river discharges ~~with expert judgment in a case study of~~ on the River Meuse. Experts' estimates of tributary discharges ~~for large return periods were combined with measured high river discharges in ranges~~ that are ~~commonly "in sample". We combined the different tributary discharges~~ 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 ~~compared~~compare the results for three approaches, i) data only, ii) expert judgment only, and iii) the combination. ~~We used~~ The expert elicitation is formalized with Cooke's method for structured expert judgment~~, in which the experts estimated the discharge that is exceeded on average once per 10 and 1,000 years. The once per 10 year estimate is in the observed range and was therefore used as calibration question for Cooke's method~~.

The results ~~showed that~~ of applying our method show credible extreme river discharges resulting from the combined expert-and-data approach~~ are most plausible. Using only data gives relatively small uncertainty bounds in the range of lower return periods, while using only experts does constrains the uncertainty in the range of higher return periods. Note that results with smaller uncertainty bands does not mean they are also correct (in order to assess correctness we would need to observe thousands of years of~~

~~measurements in an unchanging environment), but they seem credible when compared to the most extreme discharges we have observed.~~

~~In conclusion, we found that with the method presented in this study we were successfully able to derive credible extreme discharges for the River Meuse. The combined data-EJ approach performed best.~~ A comparison to GRADE, the prevailing method for estimating ~~extreme river discharges, while the experts-only approach performed~~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 Cooke's method had discharge estimates that correspond well with those discharges that might be expected from the observations. This indicates that using Cooke's method 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~~. This indicates that~~ than the EJ-data option. The method can thus also be applied ~~as well~~ to river systems where measurement data are scarce or absent~~. A case study for a different river could verify these findings. The credible results, together with the relatively limited effort needed, makes the presented method an attractive alternative for a more complex hydrological model-study.~~

~~Appendix A. Prior distribution for GEV inference~~

~~1. A weakly informed prior~~

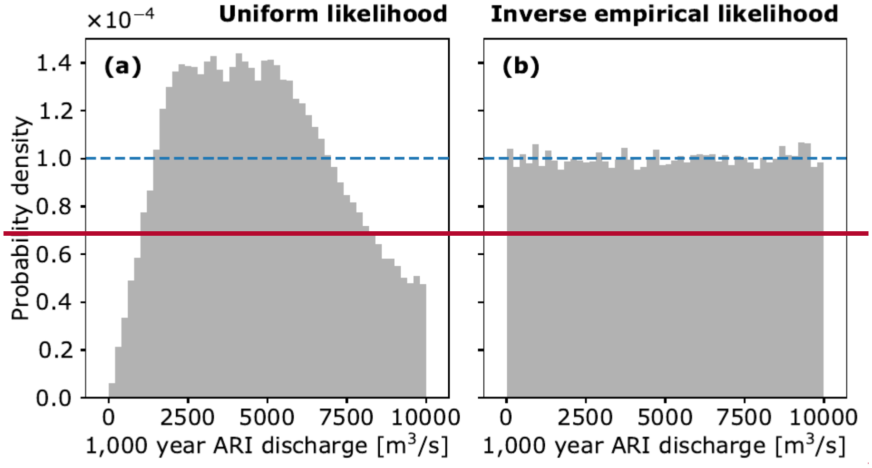
~~Section 3.3 described how Markov-Chain Monte Carlo (MCMC) was used to derive credibility bounds for the GEV fit. MCMC is an algorithm for Bayesian inference, meaning it updates a-priori distribution with observations to an a-posteriori distribution. The central theme of this paper is using structured expert judgment to quantify a discharge model, so we wanted the prior to be unbiased regarding the expert estimates. We chose that this meant using a prior for the GEV that gives a uniform distribution between 0 and 10,000 m^3/s for the 1,000-year discharge. This range is wide enough to cover the plausible range for any expert or data fit (remember that we are only fitting tributary discharges). As this is not truly uninformative, we call it a weakly informative prior.~~

~~2. Deriving the prior joint distribution~~

~~The GEV-distribution needs three parameters, a location parameter μ, scale-parameter σ, and shape parameter ξ. Ambivalence about the three parameters (i.e, a uniform distribution $\mathcal{U}$ with range $(-\infty,\infty)$ for μ and ξ, and with range $(0,\infty)$ for σ) does not lead to a uniform 1,000-year discharge, so we needed to derive the joint distribution of the three parameters that does give the required discharge distribution. We did this by using MCMC just like we fitted the expert~~, but adding information on less extreme events is recommended to increase the precision of the estimates~~ (i.e., with the likelihood function of Eq. [eq:exp_likelihood]), but then with a uniform probability density between 0 and 10,000 m³/s:~~.

$$\begin{aligned}\ell(\theta|exp) &= \sum_j \log\left(f_{\mathcal{U}_{[0,10000]}}\left(q_{p_j}\right)\right) \\ &= \sum_j \log\left(f_{\mathcal{U}_{[0,10000]}}\left(F_\theta^{-1}(1-p_j)\right)\right)\end{aligned}$$

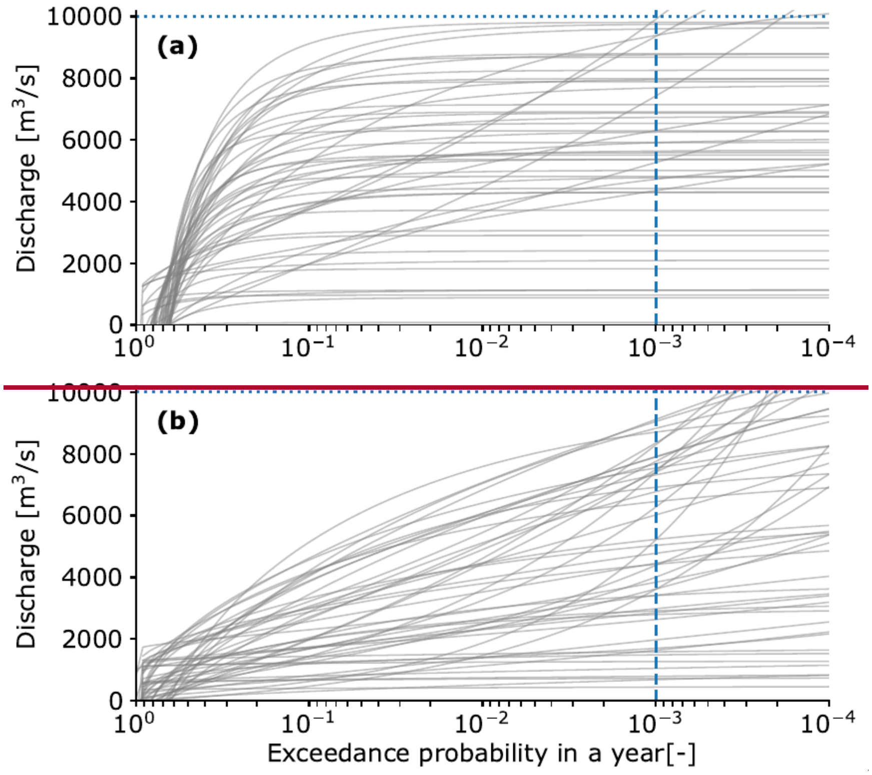
~~By using the weakly informed priors $\mu = \mathcal{U}_{[0,2000]}$, $\sigma = \mathcal{U}_{[0,10000]}$, and $\xi = \mathcal{U}_{[-2,2]}$ and doing the inference, we get the distribution for the 1,000-year discharge shown in Fig. 6 (a).~~



~~*Figure 6: Histograms of 1,000-year discharges resulting from the MCMC prior traces.* ***(a)*** *With a uniform probability density as likelihood function,* ***(b)*** *with the inverse of the left densities as likelihood function.*~~

~~The probability density is limited in between 0 and 10,000 m³/s, but the probability density is not uniform. Therefore, we repeat the above procedure, but now with an empirical probability density, of 1 divided by the densities shown in Fig. 6 **(a)**. This results in a sufficiently uniform pattern, Fig. 6 **(b)** shows.~~

~~The inference-trace can now be used as empirical prior. However, it still has a problem: The degrees a freedom within the (μ, σ, ξ)-combinations lead to too much prevalence of light-tailed distributions in the prior (i.e., a horizontal curve). Figure 7 **(a)** shows this.~~



~~*Figure 7: Example curves drawn from the priors.* ***(a)*** *Drawn from the prior with uninformed shape parameter ξ.* ***(b)*** *Using an informed shape parameter ξ.*~~

~~When using this prior to fit the expert estimates without data, the results become unfeasible, because the two expert distributions leave to much freedom for fitting the GEV shape. To solve this, we first did the inference for the data-only fits (Sect. 3.3, Eq. [eq:obs_likelihood]), and fitted a beta-distribution to the shape parameters. This distribution was then used as prior for ξ, with which we repeated the just described procedure. The wide tail shape that follows from the data (see for example Fig. [fig:extreme_discharges_Borgharen] **a.** or **d.**) leaves enough freedom for fitting differently shaped GEV-distributions to data and expert estimates. A sample of GEV-curves that results from this final prior, is shown in Fig. 7 **b.**.~~

~~To summarize, we followed these steps for deriving the GEV-prior:~~

~~1. Fit the GEV-distribution with the likelihood function from Eq. [eq:obs_likelihood] to the observed tributary discharges, using the weakly informed priors $\mu = \mathcal{U}_{[0,2000]}$, $\sigma = \mathcal{U}_{[0,10000]}$, and $\xi = \mathcal{U}_{[-2,2]}$.~~

~~1. Fit a beta-distribution through the resulting shape parameters ξ.~~

~~2. Sample from $\mu = \mathcal{U}_{[0,2000]}$, $\sigma = \mathcal{U}_{[0,10000]}$, and the fitted $\xi =$ Beta(5.35,3.72), ranging from -0.85 to 0.75, with a log-likelihood $\ell = \log\left(f_{\mathcal{U}_{[0,10000]}}(Q_{1000y})\right)$ for the 1,000-year discharge.~~

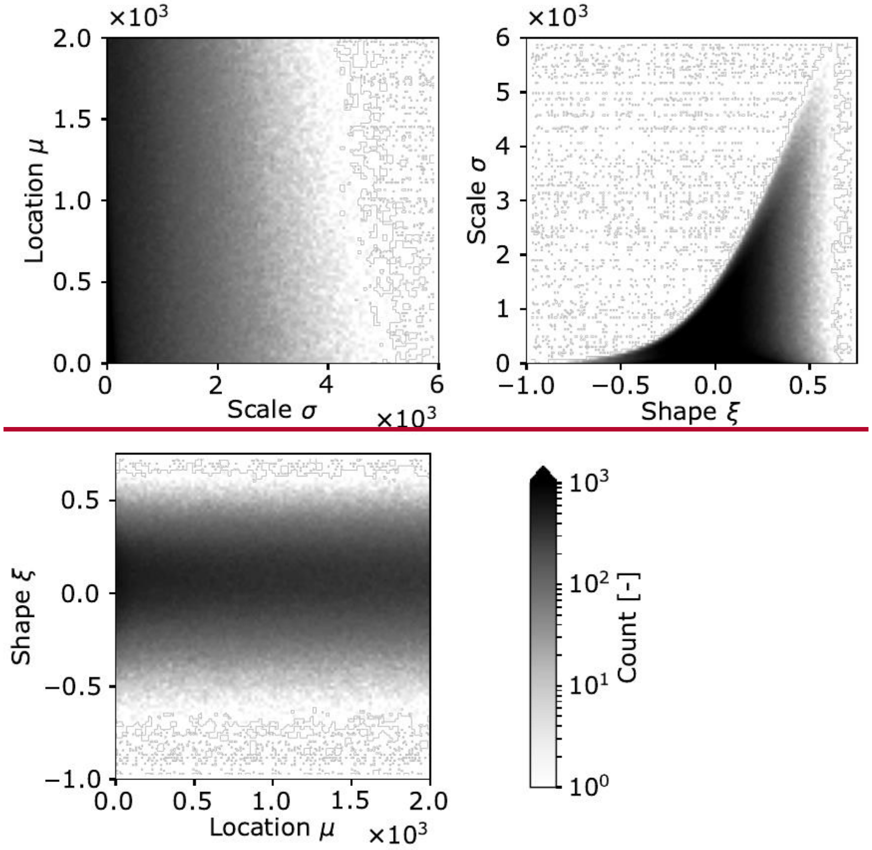
~~3. Sample again from the same priors, with an adjusted log-likelihood $\ell = \log(1/f_{obs}(q))$~~

~~4. Sample again from these priors, with an adjusted log-likelihood function, to obtain a uniform probability density for the 1,000-year discharge.~~

~~5. Use this sample as prior for all GEV-fits.~~

~~3. Sampling from the prior~~

~~The MCMC-inference gives a trace that is used as prior. A heatmap of the parameter-combinations is shown in Fig. 8.~~



~~*Figure 8: Joint prior distribution of the (μ, σ, ξ) combinations, plotted for each pair.*~~

~~To use this empirical distribution with MCMC, we sample from a uniform distribution for each of the parameters μ, σ, and ξ and transform these to the parameter-space, taking into account the dependencies:~~

- ~~$x_\mu = F^{-1}_{X_\mu}(u_\mu)$~~
- ~~$x_\sigma = F^{-1}_{X_\sigma}(u_\sigma|X_\mu = x_\mu)$~~
- ~~$x_\xi = F^{-1}_{X_\xi}(u_\xi|X_\mu = x_\mu, X_\sigma = x_\sigma)$~~

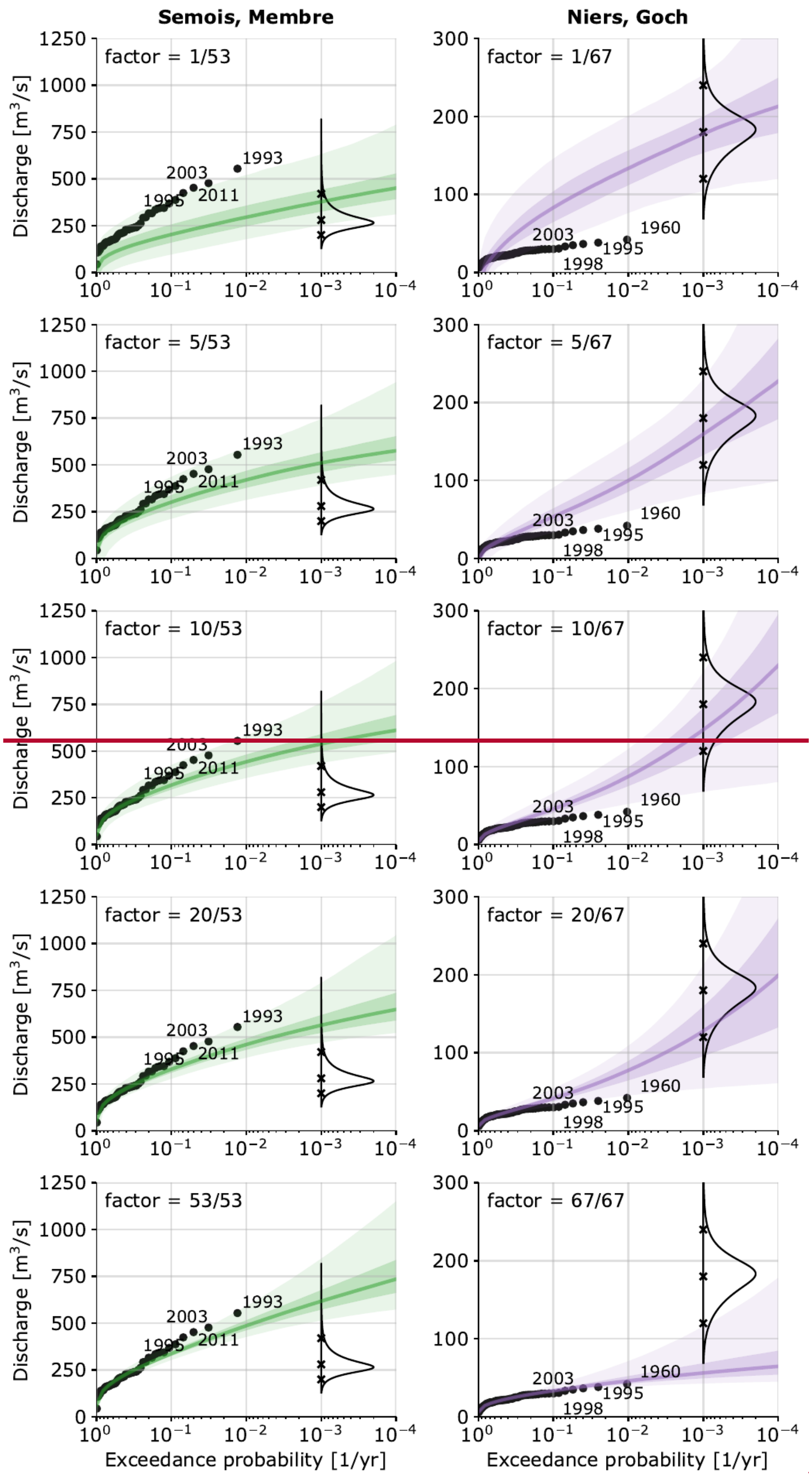
~~$F^{-1}_{X_\mu}$ is the inverse empirical CDF (i.e., the percentile point function) for parameter the location parameter X_μ. x_μ is the realization of X_μ in parameter space, and u_μ its realization in $\mathcal{U}_{[0,1]}$ space.~~

~~Numerically, x_μ, x_σ, and x_ξ are determined by discretizing the sampled values into 100 equally spaced bins per variable. The cumulative sum of the count per bin, divided by the total number of variables (i.e., normalized), gives the empirical CDF. x_μ is determined by interpolating u_μ within the normalized, cumulative bin count, and returning the bin x-values. Subsequently, x_σ is determined in a similar way, but now the empirical CDF is created from the values where $X_\mu \in bin(x_\mu)$ ($bin(x_\mu)$ is the bin that contains x_μ). Finally, x_ξ is determined by interpolating u_ξ in the empirical CDF created from the values where $X_\mu \in bin(x_\mu)$ and $X_\sigma \in bin(x_\sigma)$.~~

~~Appendix B. Sensitivity of observation to expert judgment factor~~

~~When fitting a GEV-distribution to both observations and expert judgments, like we do in this study, the contribution of each to the likelihood function affects the resulting fit. Normally, the more evidence (i.e., observations) are available, the closer the solution should follow these observations. We however strive for a balance between the two, irrespective of the number of observations, because the two sources are used for fitting the distribution to different ranges of return periods.~~

~~Equation [eq:combined_likelihood] shows the combined likelihood function, in which the factor $10/N_i$ gives the weight of the observations relative to the expert judgment. The 10 in this fraction means that the observations have a weight like (only) ten events were considered. Without the factor, the fraction would be 1.0, or N_i/N_i, while an equal weight between observation and expert judgment would give a factor $1/N_i$. Fig. 9 shows the resulting solution for the five options $1/N_i$, $5/N_i$, $10/N_i$, $20/N_i$, and N_i/N_i, for expert F's estimates for the tributaries Semois and Niers. Most experts underestimated and overestimated these two tributaries' discharges respectively. The comparison shows that $10/N_i$ gives a middle ground between observations and expert estimates in these two illustrative cases.~~



~~*Figure 9: Sensitivity of the fitted GEV for different observations-to-EJ weights. A underestimated tributary, Semois (left), as well as an overestimated tributary, Niers (right), are shown.*~~

~~The sensitivity of the outcome to the factor becomes less when expert judgment and observations are more in line with a single GEV distribution. This is expected to be the case for the high scoring experts but not necessarily, as the expert weights were determined from the on average once per 10-year discharge estimate rather than the 1000-year estimate shown in Fig. 9.~~

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 is needed, which cannot satisfactorily be derived (exclusively) from a (limited) data-record, the presented approach provides a means 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, a case studies for different rivers should verify these findings. Considering the credible results and the relatively manageable effort required, the approach presents an attractive alternative for complex hydrological studies where the uncertainty in extremes needs to be constrained.

Appendix A. Calculation of downstream discharges

Section 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:

$F_{GEV}^{-1}(1-0.001|\theta)$, in which F_{GEV}^{-1} 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 (x) are transformed to the $[1, N_M]$ interval, and are then used as index j to select a GEV-parameter combination for each of the N_T tributaries:

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

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 Y from the dependence model. These events (on a standard normal scale) are transformed to the discharge realizations Q for each tributaries' GEV parameter combination:

$$\mathrm{Q} = F_{GEV,\mathrm{j}}^{-1}\left(F_{norm}(\mathrm{Y})\right)$$

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

$$\mathrm{q} = \mathrm{f} \cdot \sum(\mathrm{Q}).$$

Note that this notation corresponds to Eq. [eq:main_model]. The N_Q discharges q are subsequently sorted and assigned a plot positions:

$$\mathrm{p} = \frac{\mathrm{k} - a}{N_Q + b},$$

with a and b being the plot positions, 0.3 and 0.4, respectively (from Bernard and Bos-Levenbach 1955). k indicates the order of the events in the set (1 being the largest, N_Q the smallest). The plot positions (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 a 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 C.~~Appendix B. Expert and DM correlation matrices

Figure ~~10~~7 shows the correlation matrices estimated by the experts. The DM correlation matrices are weighted combinations of the expert matrices, based on the weights from Table 1. See subsection 3.2 and equation [eq:DM].

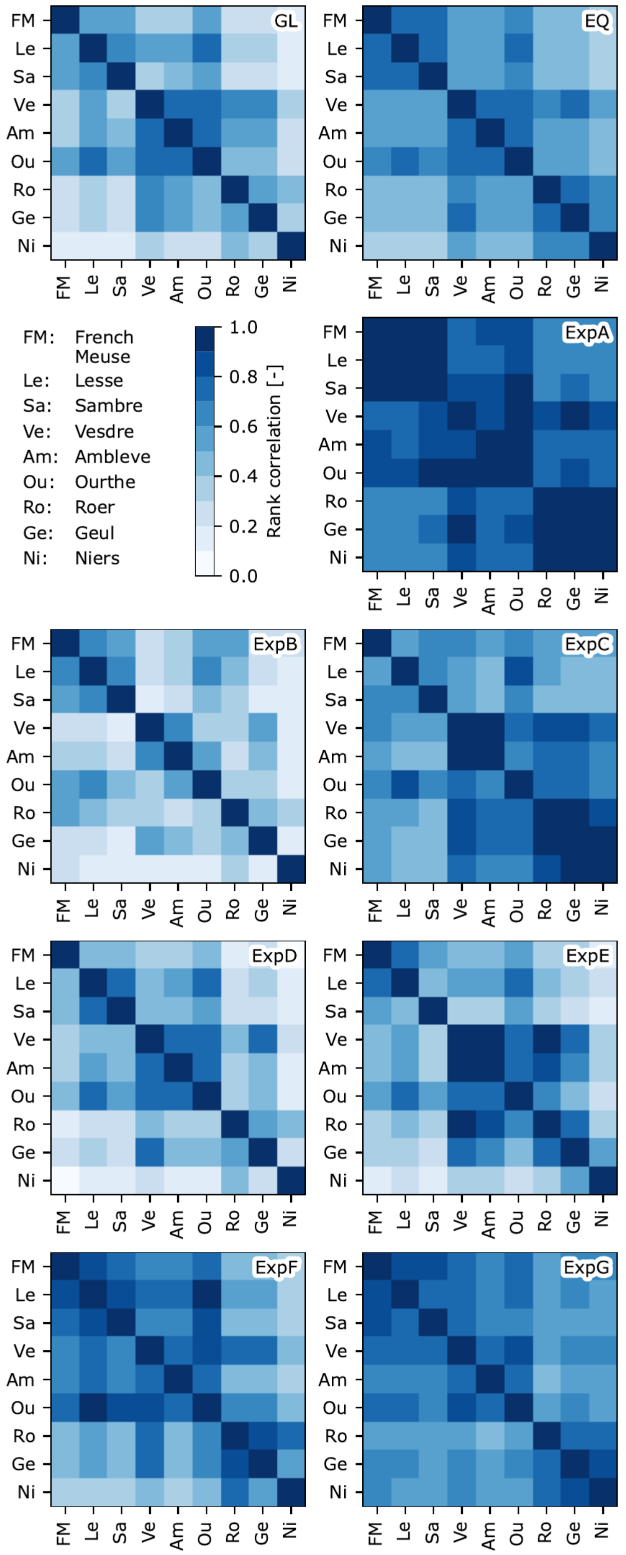


Figure ~~10~~7: Correlation matrices estimated by the expert

~~Acknowledgements~~

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. Secondly, we thank Dorien Lugt en Ties van der Heijden, who's hydrological and statistical expertise greatly helped in preparing the study through test rounds.

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

References

Al-Awadhi, Shafeeqah A, and Paul H Garthwaite. 1998. "An Elicitation Method for Multivariate Normal Distributions." *Communications in Statistics-Theory and Methods* 27 (5): 1123–42.

Bamber, Jonathan L, Michael Oppenheimer, Robert E Kopp, Willy P Aspinall, and Roger M Cooke. 2019. "Ice Sheet Contributions to Future Sea-Level Rise from Structured Expert Judgment." *Proceedings of the National Academy of Sciences* 116 (23): 11195–200.

Benito, Gerardo, and VR Thorndycraft. 2005. "Palaeoflood Hydrology and Its Role in Applied Hydrological Sciences." *Journal of Hydrology* 313 (1-2): 3–15.

Bernard, A, and EJ Bos-Levenbach. 1955. "The Plotting of Observations on Probability-Paper." *Stichting Mathematisch Centrum. Statistische Afdeling*, no. SP 30a/55.

Boer-Euser, Tanja de, Laurène Bouaziz, Jan De Niel, Claudia Brauer, Benjamin Dewals, Gilles Drogue, Fabrizio Fenicia, et al. 2017. "Looking Beyond General Metrics for Model Comparison–Lessons from an International Model Intercomparison Study." *Hydrology and Earth System Sciences* 21 (1): 423–40.

~~Borgomeo, Edoardo, Christopher L Farmer, and Jim W Hall. 2015. "Numerical rivers: A synthetic streamflow generator for water resources vulnerability assessments." *Water Resources Research* 51 (7): 5382–5405.~~

Bouaziz, Laurène JE, Guillaume Thirel, Tanja de Boer-Euser, Lieke A Melsen, Joost Buitink, Claudia C Brauer, Jan De Niel, et al. 2020. "Behind the Scenes of Streamflow Model Performance." *Hydrology and Earth System Sciences Discussions* 2020: 1–38.

~~Brown, Bernice B. 1968. "Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts." Rand Corp Santa Monica CA.~~

Brázdil, Rudolf, Zbigniew W Kundzewicz, Gerardo Benito, Gaston Demarée, Neil Macdonald, Lars A Roald, et al. 2012. "Historical Floods in Europe in the Past Millennium." *Changes in Flood Risk in Europe, Edited by: Kundzewicz, ZW, IAHS Press, Wallingford*, 121–66.

Coles, Stuart G, and Jonathan A Tawn. 1996. "A Bayesian Analysis of Extreme Rainfall Data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 45 (4): 463–78.

Colson, Abigail R, and Roger M Cooke. 2017. "Cross Validation for the Classical Model of Structured Expert Judgment." *Reliability Engineering & System Safety* 163: 109–20.

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

Cooke, Roger M., and Louis L. H. J. Goossens. 2008. "TU Delft expert judgment data base." *Reliability Engineering and System Safety* 93 (5): 657–74. https://doi.org/10.1016/j.ress.2007.03.005.

Copernicus Land Monitoring Service. 2017. "EU-DEM." https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1/view.

———. 2018. "CORINE Land Cover." https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download.

———. 2020. "E-OBS." https://cds.climate.copernicus.eu/cdsapp#!/dataset/insitu-gridded-observations-europe?tab=overview.

De Niel, J., G. Demarée, and P. Willems. 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): 8459–74. https://doi.org/https://doi.org/10.1002/2017WR020803.

Dewals, Benjamin, Sébastien Erpicum, Michel Pirotton, and Pierre Archambeau. 2021. "Extreme floods in Belgium. The July 2021 extreme floods in the Belgian part of the Meuse basin."

~~Diederen, Dirk, Ye Liu, Ben Gouldby, Ferdinand Diermanse, and Sergiy Vorogushyn. 2019. "Stochastic Generation of Spatially Coherent River Discharge Peaks for Continental Event-Based Flood Risk Assessment." *Natural Hazards and Earth System Sciences* 19 (5): 1041–53.~~

Dion, Patrice, Nora Galbraith, and Elham Sirag. 2020. "Using Expert Elicitation to Build Long-Term Projection Assumptions." In *Developments in Demographic Forecasting*, 43–62. Springer, Cham.

~~Falter, Daniela, Kai Schröter, Nguyen Viet Dung, Sergiy Vorogushyn, Heidi Kreibich, Yeshewatesfa Hundecha, Heiko Apel, and Bruno Merz. 2015. "Spatially Coherent Flood Risk Assessment Based on Long-Term Continuous Simulation with a Coupled Model Chain." *Journal of Hydrology* 524: 182–93.~~

Food and Agriculture Organization of the United Nations. 2003. "Digital Soil Map of the World." https://data.apps.fao.org/map/catalog/srv/eng/catalog.search?id=14116#/metadata/446ed430-8383-11db-b9b2-000d939bc5d8.

~~Hartley, David, and Simon French. 2021. "A Bayesian Method for Calibration and Aggregation of Expert Judgement." *International Journal of Approximate Reasoning* 130: 192–225.~~

~~Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57 (1): 97–109. https://doi.org/10.1093/biomet/57.1.97.~~

Foreman-Mackey, Daniel, David W. Hogg, Dustin Lang, and Jonathan Goodman. 2013. "emcee: The MCMC Hammer." *Publications of the Astronomical Society of the Pacific* 125 (925): 306. https://doi.org/10.1086/670067.

Goodman, Jonathan, and Jonathan Weare. 2010. "Ensemble Samplers with Affine Invariance." *Communications in Applied Mathematics and Computational Science* 5 (1): 65–80.

Hanea, Anca, Oswaldo Morales Napoles, and Dan Ababei. 2015. "Non-Parametric Bayesian Networks: Improving Theory and Reviewing Applications." *Reliability Engineering & System Safety* 144: 265–84. https://doi.org/https://doi.org/10.1016/j.ress.2015.07.027.

Hart, Cornelis Marcel Pieter 't, Georgios Leontaris, and Oswaldo Morales-Nápoles. 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/https://doi.org/10.1016/j.softx.2019.100295.

Hegnauer, M, JJ Beersma, HFP Van den Boogaard, TA Buishand, and RH Passchier. 2014. "Generator of Rainfall and Discharge Extremes (GRADE) for the Rhine and Meuse basins. Final report of GRADE 2.0." Delft: Deltares.

Hegnauer, M, and HFP Van den Boogaard. 2016. "GPD verdeling in de GRADE onzekerheidsanalyse voor de Maas." Delft: Deltares.

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): 158–71. https://doi.org/https://doi.org/10.1002/qj.49708134804.

Keelin, Thomas W. 2016. "The Metalog Distributions." *Decision Analysis* 13 (4): 243–77.

Kindermann, Paulina E, Wietske S Brouwer, Amber van Hamel, Mick van Haren, Rik P Verboeket, Gabriela F Nane, Hanik Lakhe, Rajaram Prajapati, and Jeffrey C Davids. 2020. "Return Level Analysis of the Hanumante River Using Structured Expert Judgment: A Reconstruction of Historical Water Levels." *Water* 12 (11): 3229.

Land NRW. 2022. "ELWAS-WEB." https://www.elwasweb.nrw.de/elwas-web/index.xhtml#.

Langemheen, W. van de, and H. E. J. Berger. 2001. "Hydraulische Randvoorwaarden 2001: Maatgevende Afvoeren Rijn En Maas." RIZA; Ministerie van Verkeer en Waterstaat.

Leander, Robert, Adri Buishand, Paul Aalders, and Marcel De Wit. 2005. "Estimation of extreme floods of the River Meuse using a stochastic weather generator and a rainfall." *Hydrological Sciences Journal* 50 (6).

Leontaris, Georgios, and Oswaldo Morales-Nápoles. 2018. "~~Meresa, Hadush K, and Renata J Romanowicz~~ANDURIL — a MATLAB Toolbox for ANalysis and Decisions with UnceRtaInty: Learning from Expert Judgments." *SoftwareX* 7: 313–17. https://doi.org/https://doi.org/10.1016/j.softx.2018.07.001.

Marti, Deniz, Thomas A. Mazzuchi, and Roger M. Cooke. 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, Eduardo S, and Jery R Stedinger. 2000. "Generalized Maximum-Likelihood Generalized Extreme-Value Quantile Estimators for Hydrologic Data." *Water Resources Research* 36 (3): 737–44.

~~. 2017. "The Critical Role of Uncertainty in Projections of Hydrological Extremes." *Hydrology and Earth System Sciences* 21 (8): 4245–58.~~

Ministry of Infrastructure and Environment. 2016. "Regeling Veiligheid Primaire Waterkeringen 2017 No IENM/BSK-2016/283517."

Mohr, Susanna, Uwe Ehret, Michael Kunz, Patrick Ludwig, Alberto Caldas-Alvarez, James E Daniell, Florian Ehmele, 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.

Oppenheimer, Michael, Christopher M Little, and Roger M Cooke. 2016. “Expert Judgement and Uncertainty Quantification for Climate Change.” *Nature Climate Change* 6 (5): 445–51.

Papalexiou, Simon Michael, and Demetris Koutsoyiannis. 2013. “Battle of Extreme Value Distributions: A Global Survey on Extreme Daily Rainfall.” *Water Resources Research* 49 (1): 187–201.

Parent, Eric, and Jacques Bernier. 2003. “Encoding Prior Experts Judgments to Improve Risk Analysis of Extreme Hydrological Events via POT Modeling.” *Journal of Hydrology* 283 (1-4): 1–18.

Rijkswaterstaat. 2022. “Waterinfo.” https://waterinfo.rws.nl/#!/kaart/Afvoer/Debiet 20Oppervlaktewater 20m3 2Fs/.

Rongen, G, O Morales-Nápoles, and M Kok. 2022a. “Expert Judgment-Based Reliability Analysis of the Dutch Flood Defense System.” *Reliability Engineering & System Safety* 224: 108535.

———. 2022b. “Extreme Discharge Uncertainty Estimates for the River Meuse Using a Hierarchical Non-Parametric Bayesian Network.” In *Proceedings of the 32th European Safety and Reliability Conference (ESREL 2022)*, edited by Maria Chiara Leva, Edoardo Patelli, Luca Podofillini, and Simon Wilson, 2670–77. Research Publishing. https://doi.org/10.3850/978-981-18-5183-4_S17-04-622-cd.

Rongen, Guus, Cornelis Marcel Pieter ’t Hart, Georgios Leontaris, and Oswaldo Morales-Nápoles. 2020. “Update (1.2) to ANDURIL and ANDURYL: Performance Improvements and a Graphical User Interface.” *SoftwareX* 12: 100497. https://doi.org/https://doi.org/10.1016/j.softx.2020.100497.

Rongen, GWF. 2016. “The effect of flooding along the Belgian Meuse on the discharge and hydrograph shape at Eijsden.” Master’s thesis, Delft University of Technology; Delft University of Technology.

~~Salvatier, John, Thomas Wiecki, and Christopher Fonnesbeck. 2015. “Probabilistic Programming in Python using PyMC.” *PeerJ Computer Science* 2 (July). https://doi.org/10.7717/peerj-cs.55.~~

Sebok, Eva, Hans Jørgen Henriksen, Ernesto Pastén-Zapata, Peter Berg, Guillume Thirel, Anthony Lemoine, Andrea Lira-Loarca, 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.

Service public de Wallonie. 2022. “Annuaires Et Statistiques.” http://voies-hydrauliques.wallonie.be/opencms/opencms/fr/hydro/Archive/annuaires/index.html.

~~Task Force Fact-finding hoogwater 2021.~~TFFF. 2021. “Hoogwater 2021 - Feiten en duiding.” Delft: Task Force Fact-finding hoogwater 2021; Expertisenetwerk Waterveiligheid (ENW).

Viglione, Alberto, Ralf Merz, José Luis Salinas, and Günter Blöschl. 2013. “Flood Frequency Hydrology: 3. A Bayesian Analysis.” *Water Resources Research* 49 (2): 675–92.

Waterschap Limburg. 2021. “Discharge Measurements.”

~~Winter, B, K Schneeberger, M Huttenlau, and J Stötter. 2018. “Sources of Uncertainty in a Probabilistic Flood Risk Model.” *Natural Hazards* 91 (2): 431–46.~~