Impact of model combination methods on extreme precipitation projections

Jessup, Sébastien[1] Mailhot, Mélina[1] Pigeon, Mathieu[2]

[1] Department of Mathematics, Concordia University, Montreal, Canada
[2] Département de mathématiques, UQAM, Montreal, Canada

Abstract

Climate change is expected to increase the frequency and intensity of extreme weather events. To properly assess the increased economical risk of these events, actuaries can gain in relying on expert models/opinions from multiple different sources, which requires the use of model combination techniques. From non-parametric to Bayesian approaches, different methods rely on varying assumptions potentially leading to very different results. In this paper, we apply multiple model combination methods to an ensemble of 24 experts in a pooling approach and use the differences in outputs from the different combinations to illustrate how one can gain additional insight from using multiple methods. The densities obtained from pooling in Montreal and Quebec City highlight the significant changes in higher quantiles obtained through different combination approaches. Areal reduction factor (ARF) and quantile projected changes are used to show that consistency, or lack thereof, across approaches reflects the uncertainty of combination methods. This shows how an actuary using multiple expert models should consider more than one combination method to properly assess the impact of climate change on loss distributions, seeing as a single method can lead to overconfidence in projections.

Keywords Extreme precipitation; Model combination; Non-parametric methods; Bayesian model averaging

1 Introduction

Climate change and global warming are expected to lead to increases in catastrophic weather events such as wildfires, droughts, and extreme precipitation. These changes can have many effects such as crop damage, soil erosion, and increased risk of flooding. Quantifying severe weather events is of particular interest to actuaries, since events such as flooding account for a large part of global economic losses (Boudreault et al., 2020). An increase in extreme rainfall can lead to a possibly greater increase in river

discharge (Breinl et al., 2021). Therefore, one would gain from obtaining reliable rainfall projections to assess flood risks.

Modelling precipitation behaviour, and weather events in general, requires complex models. For example, seasonality needs to be taken into account (e.g. Kodra et al. (2020)), as well as wind patterns, which also use advanced models (see for example Gracianti et al. (2021)). One further needs to model spatial interpolation (Wagner et al. (2012), Hu et al. (2019), etc.). As such, projecting changes in extreme precipitation would mean combining these elements with extreme value theory in a limited data context. Given that different models may capture different elements of a system's behaviour, when interested in extreme precipitation, one will often receive diverging information from multiple sources and may wish to combine these sources of information. These sources can often be considered as expert opinions, which are used in actuarial science, for example, particularly in mortality studies, where deterministic projections are incorporated into mortality forecasting via Continuous Mortality Investigation (Huang and Browne, 2017) and P-Splines (Djeundje, 2022). Combining expert opinions and models is especially important for actuaries to set credible hypotheses when modelling losses from weather events.

Extreme weather events caused $2.1 billion in insured damage in Canada alone in 2021 (Insurance Bureau of Canada, 2022), and losses from natural catastrophes have been increasing over the last 20 years. In this context, the last few years have seen increased demand for catastrophe insurance, particularly flood insurance, and private insurers have been developing new products to respond to this demand. The challenge with modelling flood losses, or severe weather events in general, is that the covered events do not occur frequently, and the changing nature of climate implies that only relatively short spans of time can be considered to have similar risks. This compounds the lack of data necessary for developing actuarial models with traditional techniques requiring a high volume of frequency and severity data. Given that expert climate models specialise in the complex dynamics of weather events, combining these models offers an appealing solution for insurers by allowing for an alternate way of obtaining reliable models for catastrophic events.

To efficiently combine models, one needs to determine how much weight to give to each expert's opinion. Clemen (1989) reviewed forecast combination literature, concluding that combining individual forecasts substantially improves accuracy, and that simple methods work reasonably well relative to more complex methods. By reviewing statistical techniques for combining multiple probability distributions, Jacobs (1995) showed that independent experts yield more information than dependent experts, where dependent experts might for example have models relying on one another. Cooke et al. (1991) also reviewed expert combination and offered a non-parametric approach for attributing weights to experts based on specific quantiles. From a different perspective allowing for the potential use of a prior opinion about each of the experts, Mendel and Sheridan (1989) and Raftery et al. (1997) used Bayesian approaches to combine expert distributions.

Such methods have been further developed, in particular with Bayesian Model Aver-

aging (BMA) gaining popularity in recent years. For example, Broom et al. (2012) considered BMA in a limited data context, and Fragoso et al. (2018) provided a review of its applications in 587 articles from 1990 to 2014, covering biology, social sciences, environmental studies, and financial applications. In the last few years, the concept of BMA has been generalised into Bayesian Predictive Synthesis (BPS) in a financial time series context (e.g. Johnson (2017), McAlinn and West (2019), McAlinn et al. (2020)). Model combination can be useful in areas such as climate modelling, where significant uncertainty is present, especially in the context of climate change, and different models rely on different hypotheses. BMA is currently used to this end, for example Massoud et al. (2020) used BMA to study mean precipitation changes in the US by region.

In the context of extreme rainfall leading to flooding, spatial distribution becomes important as it can significantly change risk exposure, where a local rainfall does not lead to the same risks as widespread rainfall. To analyse this spatial distribution, areal reduction factors (ARF) are often used to convert point rainfall into areal rainfall (see for example Svensson and Jones (2010)). The impact of climate change on ARFs was studied by Li et al. (2015) for the region of Sydney, Australia. A limitation of this study is that the authors used a single expert model to obtain precipitation projections. One would seek to improve this type of analysis by combining multiple expert projections. A challenge with this idea is that combination methods often require larger datasets than are available in an extreme precipitation context. This is especially true given that precipitation patterns are changing, where considering an extended span of time means differences in precipitation distribution within the dataset. To circumvent this issue, Innocenti et al. (2019) used a model pooling approach with a 50-member ensemble when studying extreme precipitation in Northeastern North-America, allowing the authors to use 3-year periods of data. Supposing that all expert projections are equally likely, the authors could then apply frequency analysis to study 99^{th} quantiles. An advantage of this method, beyond its simplicity and effectiveness, is that it allows for observing how variability between expert models can be used to improve the estimation of annual maxima statistics. A question that naturally arises is whether attributing weights to each expert based on combination methods instead of supposing all projections are equally likely would yield significantly different projections. This question is of particular interest to actuaries, since changing the underlying precipitation hypotheses would have an effect on event probabilities, and thus affect both pricing and reserving.

We thus focus on the impact of model combination methods on quantile and ARF projections when applied to the pooling approach of Innocenti et al. (2019) in Montreal and Quebec, Canada. The paper is divided as follows: Section 2 provides details regarding parametric and non-parametric model combination methods, Section 3 applies these methods to pooling to obtain extreme precipitation quantile and ARF projections, and briefly explains how such projections can be used for flood damage modelling. Finally, Section 4 provides concluding comments. Additional material can be found in Appendices I-IV.

2 Model combination methods

Expert climate research groups often provide diverging information based on varying methods and underlying hypotheses regarding greenhouse gas emissions, changes in global convection patterns, the impact of topography, etc. One may seek to combine this information by using an array of tools such as non-parametric approaches or Bayesian approaches. This section presents approaches from various combination methods relying on different hypotheses. To easily analyse the differences between approaches, we choose well known approaches allowing for establishing weights to attribute to each expert, as compared to less transparent machine learning methods such as neural networks, for example. Such methods are however increasing in popularity, as highlighted in a review of recent AI applications in actuarial science by Richman (2021). As will be shown in Section 3, the choice of method can lead to very different probabilities attributed to each expert's projections, suggesting that one can benefit from investigating the differences between expert models with higher probability.

Before going into each method's details, the following notation will be used throughout the remainder of this paper. Consider a vector of years $\vec{\tau} = \{s,s+1,\ldots,t\}$, where $s \in \{0,\ldots,t\}$, $t \leq T$, with $T \in \mathbb{N}$ the latest available year. Let $\vec{Y}_{\vec{\tau},x}$ be a vector of random variables representing the precipitation annual maxima of $G(x,\vec{\tau},d)$, the daily precipitation at site x for day d, for years in $\vec{\tau}$. Further let the vector of random variables $\vec{Y}_{\vec{\tau},A}$ be the annual maxima of $H(A,\vec{\tau},d)$ for the same period from s to t, where $H(A,\vec{\tau},d)$ is the average areal rainfall for day d, such that $H(A,\vec{\tau},d) = \frac{1}{\text{card}(X)} \sum_{x \in X} G(x,\vec{\tau},d)$ for a collection of sites $x \in X$ within the area A. The respective realisations of $G(x,\vec{\tau},d)$ and $H(A,\vec{\tau},d)$ are then $\vec{y}_{\vec{\tau},x}$ and $\vec{y}_{\vec{\tau},A}$, with length $t-s+1$.

Consider n experts providing a model $\mathcal{M}_e$ allowing for projections of annual maxima for site x and area A, $\vec{y}^{(e)}_{\vec{\tau},x}$ and $\vec{y}^{(e)}_{\vec{\tau},A}$ respectively, where $e \in \{1,\ldots,n\}$, over a period $\vec{\tau}$ as described above. With a certain weight w_e attributed to each expert, the objective is then to obtain a precipitation projection with a weighted sum of the experts' projections, that is,

$$\tilde{\vec{y}}_{\vec{\tau},x} = \sum_{e=1}^{n} w_e \vec{y}^{(e)}_{\vec{\tau},x}.$$

The goal of each method is then to obtain these w_e from calibration variables. These are variables for which we know the true values, while the experts providing their opinion do not. This information then allows us to calibrate how much weight we give to each expert. Consider K such calibration variables $V_1,\ldots,V_K$. We specify M percentages for which each one of n experts provides corresponding quantiles $v^{(e)}_{k,m}$, $k = 1,\ldots,K$; $m = 1,\ldots,M$; and $e = 1,\ldots,n$. In the context of extreme precipitation projection, we would have $\text{card}(X)$ calibration variables corresponding to $\vec{Y}_{\vec{\tau},x}$ for a calibration period $\vec{\tau}$.

2.1 Inverse Distance Weighting

A first possible approach to model combination is to intuitively build weights based on the distance between an expert's projection about a variable of interest, or vector of variables, and the true value of this variable. This idea can be achieved through Inverse Distance Weighting (IDW). The advantage of this approach is its intuitiveness and ease of use.
Classically, IDW was used with Euclidean distance. In a geometric context, Shepard (1968) used IDW to consider distance while taking angles into account. In a probabilistic setting, the challenge with this method is then to determine an appropriate distance measure. One such measure is the Wasserstein distance, which Kantorovitch and Rubinštein (1958) first realised was applicable to probability distributions. This idea was expanded on by Givens and Shortt (1984), and used recently by Pesenti et al. (2021) for sensitivity analysis. In the univariate case, the distance for expert e over time period $\vec{\tau}$ at location x is defined as

$$D^{(e)} = \int |F_{Y^{(e)}_{\vec{\tau},x}}(y) - F_{Y_{\vec{\tau},x}}(y)|dy,$$

with $F_{Y_{\vec{\tau},x}}$ the real cumulative distribution function and $F_{Y^{(e)}_{\vec{\tau},x}}$ the expert's CDF.
With this distance, the weight attributed to each expert's projection is then

$$w_e = \frac{1/D^{(e)}}{\sum_{l=1}^{n} 1/D^{(l)}}.$$

2.2 Non-parametric calibration

From a literature-based approach, model combination can be approached from many angles. Cooke et al. (1991) offered a review of early expert combination methods. Clemen and Winkler (1999) further elaborated on this review, suggesting issues that need to be considered when combining expert opinions such as expert selection and the role of interaction between experts. Since then, Cooke and Goossens (2008) and Hammitt and Zhang (2012) compared the performance of multiple combination methods, among which a classical approach which was first presented by Cooke et al. (1991).
This combination method uses desirable properties of scoring rules, namely that they should be coherent, strictly proper, and relevant (see Cooke et al. (1991) for details). A three-part method was established attributing weights to each expert distribution based on a relative information component, a calibration component, and an entropy component. This method has the advantage of being non-parametric, suggesting that an expert does not need to have a complete statistical model. Such a method can be appropriate for example in actuarial science, where an expert might reasonably provide an estimate for a small, medium, and large loss, but not a full loss distribution.
From the calibration variables V_1 to V_K defined previously, we set $v_{k,0}$ and $v_{k,M+1}$ such that

$$v_{k,0} < v^{(e)}_{k,m} < v_{k,M+1} \ \forall \ m,e.$$

We compare these selections and expert-provided values with the true observed values to find the proportion of calibration variables in each interquantile space. This forms an empirical distribution $\boldsymbol{q} = \{q_1, \ldots, q_{M+1}\}$ that we can compare to the theoretical proportion $\boldsymbol{p} = \{p_1, \ldots, p_{M+1}\}$. As shown by Cooke et al. (1991), we can obtain the calibration and entropy components, $C(e)$ and $O(e)$ respectively, as

$$C(e) = 1 - \chi^2_{K-1}((2K)I(q,p)),$$

where

$$I(q,p) = \sum_{k=1}^{M+1} q_k \ln\left(\frac{q_k}{p_k}\right)$$

is the relative information component, and

$$O(e) = \frac{1}{K}\sum_{k=1}^{K}\left(\ln(v_{k,M+1} - v_{k,0}) + \sum_{m=1}^{M+1} p_m \ln\left(\frac{p_m}{v^{(e)}_{k,m} - v^{(e)}_{k,m-1}}\right)\right).$$

It can readily be shown that the relative information component $I(q,p)$ multiplied by $2K$ (i.e. twice the number of calibration variables) follows a Chi-squared distribution. The calibration component uses this fact to measure the goodness of fit of each expert forecast, while the entropy component measures the distance of expert forecasts from a uniform distribution. The intuition for this component is that a uniform model provides very little useful information. From these, we finally obtain

$$w'_e = C(e)O(e)I_{\{C(e)>\alpha\}}$$

for a specified threshold α chosen by optimising the score of the combined distributions, where $0 < \alpha < 1$. This α can be seen as a hyperparameter representing the minimal calibration level that each model needs to satisfy to receive weight. As such, a higher α means we give probability to less models. This also implies that the maximal value for α is the highest value of $C(e)$. We can then recalibrate the weights to make their sum equal to 1 by dividing w'_e by the sum over all experts:

$$w_e = \frac{w'_e}{\sum_{l=1}^{n} w'_l}.$$

These w_e do not require the analyst to have a prior opinion of each expert's projections. We will refer to this method as Cooke's method for the sake of brevity. In the context of daily precipitation annual maxima, the corresponding calibration variable is then $\vec{Y}_{\vec{\tau},x}$, where we consider K different sites x.

2.3 Bayesian Model Averaging

As an alternative to the previous approaches, one may seek to exploit their prior knowledge using Bayesian methods, updating a prior belief with observed data to obtain a posterior distribution more representative of recent data.

Bayesian Model Averaging (BMA) is a widely used tool for model combination. Recently, in the United States, BMA was used to study extreme rainfall density as well as daily mean rainfall by Zhu et al. (2013) and Massoud et al. (2020) respectively. First made popular by Raftery et al. (1997) in linear models, BMA uses observed data to update weights to different models based on their likeliness. This relies on the premise that any of the models could be right, but selecting only one model would fail to capture the uncertainty around this choice. This in turn leads to reducing overconfidence from ignoring a model's uncertainty. BMA however implicitly relies on the assumption that one of the models must be right (Hoeting et al., 1999). Note that the method presented in Cooke et al. (1991) relies on a similar assumption, given that the optimal α requires at least one model to be chosen.

Let $\mathcal{M}$ be a discrete variable representing this best model, with possible values $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$. An analyst has some prior belief about the probability that each expert's model is right, $\Pr(\mathcal{M} = \mathcal{M}_e)$, which we will denote $\Pr(\mathcal{M}_e)$, normalised such that $\sum_{e=1}^{n} \Pr(\mathcal{M}_e) = 1$. In the absence of prior information, then $\Pr(\mathcal{M}_e) = 1/n$, $\forall\, e$. Given data $\vec{y}_{\vec{\tau},x}$, the analyst can update these probabilities through Bayesian updating, that is

$$\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},x}) = \frac{\Pr(\vec{y}_{\vec{\tau},x}|\mathcal{M}_e)\Pr(\mathcal{M}_e)}{\sum_{l=1}^{n}\Pr(\vec{y}_{\vec{\tau},x}|\mathcal{M}_l)\Pr(\mathcal{M}_l)},$$

where $\Pr(\vec{y}_{\vec{\tau},x}|\mathcal{M}_e)$ is the probability of observing $\vec{y}_{\vec{\tau},x}$ under model $\mathcal{M}_e$. Since we divide by $\sum_{l=1}^{n}\Pr(\vec{y}_{\vec{\tau},x}|\mathcal{M}_l)\Pr(\mathcal{M}_l)$, it follows that $\sum_{e=1}^{n}\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},x}) = 1$, and posterior probabilities $\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},x})$ can therefore be considered as updated weights attributed to each expert. This supposes that all models are independent since we ignore possible interactions between models. This assumption is appropriate in this case since all experts rely on different approaches, but this will be discussed in Section 4. There are different ways of calculating the expert-associated probabilities.

A first possibility is to use an Expectation-Maximisation (EM) algorithm, as shown by Darbandsari and Coulibaly (2019), where the residuals between the model projections $\vec{y}^{(e)}_{\vec{\tau},x}$, representing an expert's projection generated from model $\mathcal{M}_e$ about the variable $\vec{Y}_{\vec{\tau},x}$, and actual data are assumed to follow a Gaussian distribution. This assumption allows for iterating through these residuals' Gaussian likelihood while updating the weights attributed to each expert model until the difference between iterations is less than some threshold β. The algorithm is outlined in Appendix I. The algorithm allows for projecting a posterior distribution for a period $\vec{\psi} = \{s',s'+1,\ldots,t'\}$, with $s' \in \{t,t+1,\ldots,t'\}$, $t < t' \leq T$. This approach must be used carefully as it can lead to overfitting. With a low threshold, expectation-maximisation will be optimised for training data, but will also learn the noise surrounding the signal. Because of this, the algorithm can then perform poorly on testing data. This limitation of the EM algorithm will be further explored in section 3.

The same hypothesis that residuals follow a normal distribution was used by Zhu et al. (2013), but with a different approach due to limited datasets, where the authors used bootstrapping under Generalised Likelihood Uncertainty Estimation (GLUE, see Beven

and Freer (2001)) to obtain the posterior likelihoods. The algorithm is presented in Algorithm 1, where $y_{\vec{\tau},x,m}$ is the m^{th} quantile of the vector $\vec{y}_{\vec{\tau},x}$, $y_{\vec{\tau},x,m,b}$ is the b^{th} bootstrap resampling of this quantile with B resamplings, and $\Pr(Y_{\vec{\psi},x} = y|\mathcal{M}_e)$ is the probability distribution of extreme precipitation under model $\mathcal{M}_e$ for a future period $\vec{\psi}$.

Algorithm 1: Generalised Likelihood Uncertainty Estimation

1: Resample $y_{\vec{\tau},x,m}$ to obtain B bootstrap iterations $y_{\vec{\tau},x,m,b}$.

2: Calculate the variance for quantile m as $\sigma_m^2 = \frac{1}{B}\sum_{b=1}^{B}\left(y_{\vec{\tau},x,m,b} - \frac{1}{B}\sum_{i=1}^{B} y_{\vec{\tau},x,m,i}\right)^2$.

3: Calculate the likelihood assuming residuals follow a normal distribution:

$$L(\vec{y}^{(e)}_{\vec{\tau},x},m) = \frac{1}{\sqrt{2\pi}\sigma_m}\exp\left(-\frac{\frac{1}{B}\sum_{b=1}^{B}\left(y_{\vec{\tau},x,m,b} - y^{(e)}_{\vec{\tau},x,m,b}\right)^2}{2\sigma_m^2}\right)$$

$$L(\vec{y}^{(e)}_{\vec{\tau},x}) = \frac{1}{M}\sum_{m=1}^{M} L(\vec{y}^{(e)}_{\vec{\tau},x},m).$$

4: Update the probability of each expert as

$$\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},x}) = \frac{L(\vec{y}^{(e)}_{\vec{\tau},x})\Pr(\mathcal{M}_e)}{\sum_{l=1}^{n} L(\vec{y}^{(l)}_{\vec{\tau},x})\Pr(\mathcal{M}_l)}.$$

5: Calculate posterior distribution as

$\Pr(y|\vec{y}_{\vec{\tau},x}) = \sum_{e=1}^{n}\Pr(y|\mathcal{M}_e)\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},x})$.

3 Application to Areal Reduction Factors

In the context of extreme precipitation, where projections from multiple models are available, model combination can become a particularly useful tool. The issue with combining models with annual maxima data is that datasets are limited. To find projected precipitation trends in annual maxima at a 1 in 100 return level, Innocenti et al. (2019) pooled $\vec{y}^{(e)}_{\vec{\psi},x}$ across all experts for projected time period $\vec{\psi}$, thus significantly increasing available data for small spans of time. Let $\vec{Y}_{\vec{\psi},x}$ be the vector of random variables describing annual maxima for period $\vec{\psi}$. The pooled "observations" for this variable are then

$$\vec{y}_{\vec{\psi},x} = (\vec{y}^{(1)}_{\vec{\psi},x}, \vec{y}^{(2)}_{\vec{\psi},x}, \ldots, \vec{y}^{(n)}_{\vec{\psi},x}),$$

where all elements of $\vec{y}_{\vec{\psi},x}$ are considered equiprobable, such that

$$\Pr(Y_{\vec{\psi},x} = y) = \frac{1}{(t' - s' + 1)n},$$

with $y \in \vec{y}_{\vec{\psi},x}$, n experts, and $\vec{\psi}$ having length $t' - s' + 1$.

Applying frequency analysis to this pooled set, we define the quantile corresponding to a certain frequency R as $Y \in \vec{Y}_{\vec{\psi},x}$ such that

$$Y_{\vec{\psi},x,R} = \min\{Y_{\vec{\psi},x} : \Pr(Y \leq Y_{\vec{\psi},x}) \geq 1 - 1/R\},$$

where for example for a 1 in 20 year return level, we would have $1 - 1/20 = 0.95$.

3.1 Non-equiprobable pooling

In the previous section, we saw different methods to attribute weights to expert opinions depending on the probability of each expert projection being accurate. We can incorporate these ideas into the pooling idea of Innocenti et al. (2019). We use their pooling method as a baseline, where one may consider all expert-provided models as equally likely, which we will refer to as the equiprobable scenario. Instead of supposing that all model projections are equally likely ($\Pr(\mathcal{M}_e) = 1/n$), we can update our belief about the probability of each model with observed data. By defining

$$\Pr(Y_{\vec{\psi},x} = y) = \frac{\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},x})}{t' - s' + 1},$$

with $y \in \vec{y}_{\vec{\psi},x}$, we obtain a shifted distribution reflecting this updated belief, where $t' - s' + 1$ is the number of years in the future projection period $\vec{\psi}$, and $\vec{\tau}$ is the historical observed period.

3.2 Calculating areal reduction factors

We can now incorporate the model combination methods and pooling presented previously into ARFs to investigate their impact on extreme precipitation quantile and ARF projections.

Although there are slightly varying definitions of ARFs, we will focus on the one used by Le et al. (2018), which can be thought of as a quantile of average areal precipitation over an average of point precipitation quantiles. This particular definition has the advantage of being applicable to any station within a region and not only one station. Starting from the notation introduced in Section 2, let $Y_{\vec{\tau},A,R}$ and $Y_{\vec{\tau},x,R}$ respectively represent the areal and point rainfall for area A, point x, and frequency R over period $\vec{\tau}$. The ARF based on daily precipitation is then

$$\text{ARF}_{(A,R,\vec{\tau})} = \frac{Y_{\vec{\tau},A,R}}{\frac{1}{\text{card}(X)} \sum_{x \in X} Y_{\vec{\tau},x,R}},$$

where there are a collection of sites $x \in X$ within area A.

An issue that arises when calculating ARFs with climate models is that expert projections are often not available at each point x, but rather at a grid scale. This issue can however be solved by assuming that scaling from point precipitation to grid average precipitation is time invariant. Li et al. (2015) demonstrated the validity of this

hypothesis, enabling the use of grid cells for ARF calculation, where we would have grid-to-area instead of point-to-area.

With this notion of time-constant scaling, we can thus consider the points x as grid cell coordinates instead of stations. This enables us to calculate ARFs using grid data, as made available by climate agencies such as Climate Data Canada and Copernicus Climate Change Service. Grid cells are available at a resolution of approximately 0.1 degrees of latitude and longitude, and represent average precipitation over the grid cell. We consider zones of 6×4 grid cells in the regions of Montreal and Quebec. We have access to 24 different climate models using historical data from 1951 to 2005 to project precipitation from 2006 to 2100. These models rely on three different Representative Concentration Pathways (RCP) emission scenarios: a low emissions scenario (RCP 2.6), a moderate emissions scenario (RCP 4.5) and a high emissions scenario (RCP 8.5). In keeping with Innocenti et al. (2019), we will focus on the 8.5 scenario, corresponding to a 4.5 degree increase by 2100. We calibrate weights using data from 2001 to 2020, for which we have both real and projected precipitation. This allows us to compare quantiles for Bayesian Model Averaging, or interquantile space for Cooke's method and inverse Distance Weighting, and so calibrate combination weights using each method. With the obtained weights, all future time periods are then forecasted. It is worth noting that this relies on the hypothesis that weights remain the same whether forecasting near or far future.

To use pooling, we need to have sufficient data for frequency analysis. Due to having 24 models instead of the 50 in Innocenti et al. (2019), we consider 6-year periods, such as precipitation from 2016 to 2021, rather than 3-year periods to obtain a similar number of data points. Applying weights calculated using the different methods presented in Section 2, we calculate shifted densities reflecting these adjusted weights, as can be observed in Figures 4 and 5. However, before using the BMA-EM algorithm, a threshold or number of iterations must be chosen to prevent overfitting. This is because too many iterations of the expectation-maximisation algorithm will lead to learning the signal as well as the noise in the training data. Figure 1 illustrates the average MSE resulting from splitting data from 2001 to 2020 into ten-year training and testing periods. Overfitting occurs passed 4 iterations of the expectation-maximisation algorithm, where we see that the testing sample MSE starts increasing significantly while the training sample MSE stabilises and even slightly increases. To prevent this overfitting, we choose to stop the algorithm after 4 iterations. This is a known issue of BMA (see for example Domingos (2000)), added to BMA tending to select only one model asymptotically, as BMA implicitly relies on the assumption that one of the models is true (Hoeting et al., 1999). An α of 0.65 is also selected for Cooke's method by optimising the error as shown in Figure 1.

We first note that different combination methods can yield very different weights attributed to each model. Figure 2 illustrates the difference in weights for the cities of Montreal and Quebec for a period from 2001 to 2020. Note that for the rest of the article, when we refer to Quebec, this will imply Quebec City and not the province. We see that for Montreal, the two BMA methods generally agree, whereas they do not for

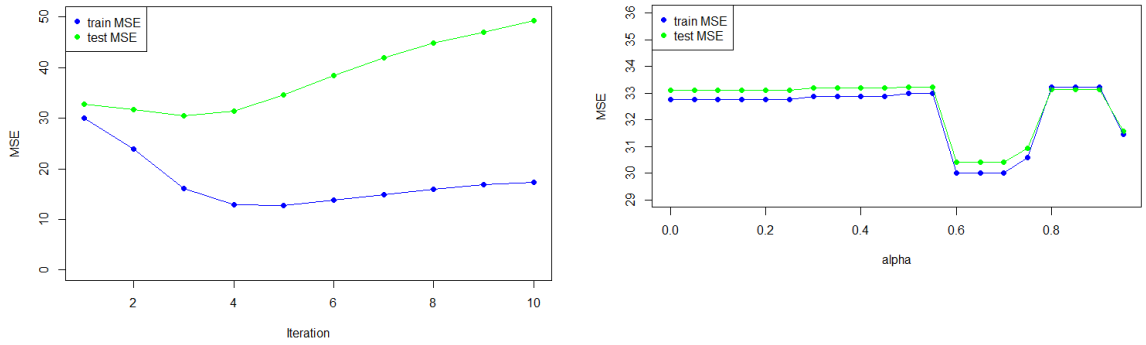

Figure 1 – Grid cell MSE of the expectation-maximisation algorithm (left) and Cooke's method (right) in the Montreal region from 2001 to 2020

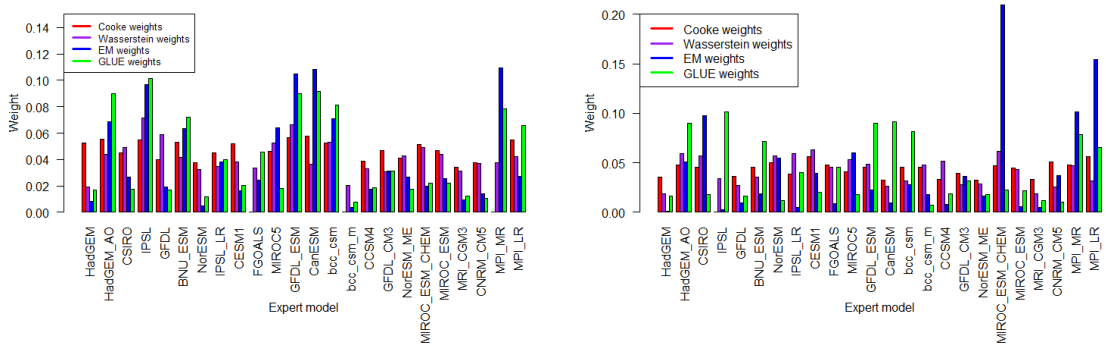

Figure 2 – Model weight by method for Montreal (left) and Quebec (right) for precipitation from 2001 to 2020

Quebec. On the other hand, both Cooke and IDW lead to relatively similar weights in both locations, but they differ from BMA results. These different weight attributions can lead to different projected quantiles.

One may seek to investigate the expert models with larger probability to ensure they agree with those models' hypotheses. For example, in Montreal, the next to last model (MPI_MR) receives a large weight from the EM algorithm, but gets truncated by the calibration approach. This happens because the model has a jump in precipitation level around the 50^{th} quantile, as illustrated in Figure 3. 7 observations out of 20 fall in the 45-50% interquantile space for model MPI_MR. This causes a poor fit in calibration in terms of Cooke's method, but the quantile-to-quantile residuals are quite small, meaning that we still have a good fit in terms of low residuals when compared to real data, making the BMA methods give this model high weight. In similar fashion, one can gain additional insight by comparing the outputs of different combination approaches.

Figures 4 and 5 illustrate the upper tail of the resulting empirical cumulative distribu-

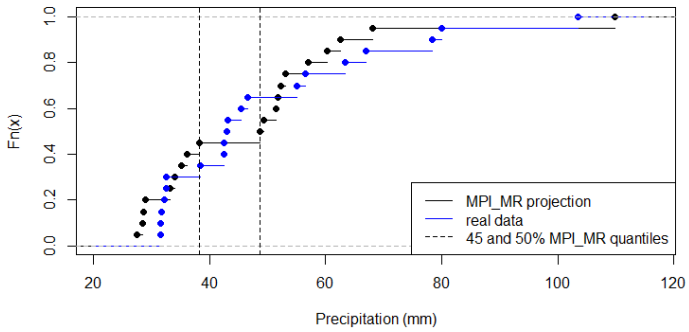

Figure 3 – Cumulative distribution for model MPI_MR and real data in Montreal for a grid cell between 2001 and 2020

tion functions under different possible combination methods for Montreal and Quebec respectively. We see that the quantiles obtained from varying combination methods are substantially different depending on the weights attributed to each model. From a risk management perspective, such differences can alter conclusions reached by an analyst concerning risk level. As such, one would benefit from considering multiple combination methods, given that this would allow for better understanding of projection uncertainty. Since different combination methods yield different results, one may be interested in the variability induced by attributing weights to each expert. Let $F^{(e)}_{\vec{\tau},x}$ be the cumulative distribution function corresponding to model $\mathcal{M}_e$. We define the CDF of $Y_{\vec{\tau},A}$ as

$$F_{\vec{\tau},A}(y) = w_1 F^{(1)}_{\vec{\tau},A}(y) + \ldots + w_n F^{(n)}_{\vec{\tau},A}(y),$$

where $w_1, \ldots, w_n$ are the weights attributed to each expert (which correspond to probabilities $\Pr(\mathcal{M}_e|\vec{y}_{\vec{\tau},A})$). It is easy to show that for a given return level, the boundaries for $Y_{\vec{\tau},A,R}$ will be the minimum and maximum of $\{Y^{(1)}_{\vec{\tau},A,R}, \ldots, Y^{(n)}_{\vec{\tau},A,R}\}$. Indeed, we have

$$\begin{aligned} Y_{\vec{\tau},A,R} &= \min\left(Y_{\vec{\tau},A} : \Pr(Y \leq Y_{\vec{\tau},A}) \geq 1 - 1/R\right) \\ &= \min\left(Y_{\vec{\tau},A} : F_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq 1 - 1/R\right) \\ &= \min\left(Y_{\vec{\tau},A} : w_1 F^{(1)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) + \ldots + w_n F^{(n)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq 1 - 1/R\right). \end{aligned}$$

Now suppose $F^{(i)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq F^{(j)}_{\vec{\tau},A}(Y_{\vec{\tau},A})$ for some $i \in \{1, \ldots, n\}$ and $\forall\, j \in \{1, \ldots, n\}$. Then it follows that $F^{(i)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq w_1 F^{(1)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) + \ldots + w_n F^{(n)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq 1 - 1/R$, provided that $w_1, \ldots, w_n \in [0,1]$ with $\sum w_i = 1$, and so $F^{-1}_{\vec{\tau},A}(Y_{\vec{\tau},A})$ must be the minimum for $Y_{\vec{\tau},A,R}$ for any combination of weights. Similarly, the reverse logic allows for stating that the lowest CDF must yield the maximum quantile.

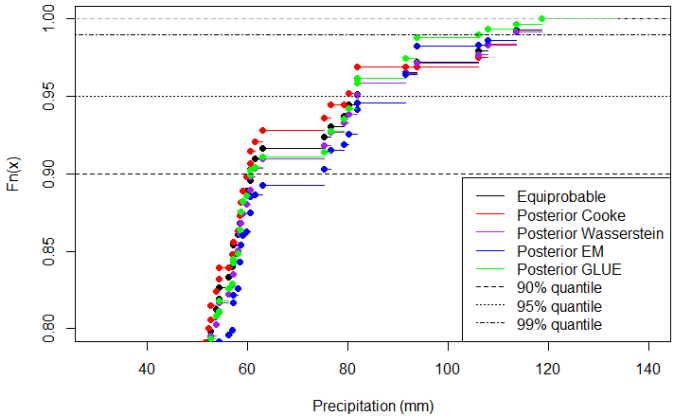

Figure 4 – Upper tail of empirical cumulative distribution functions of pooled annual maximum daily rainfall (mm) for Montreal from 2016 to 2021 with different weighting methods

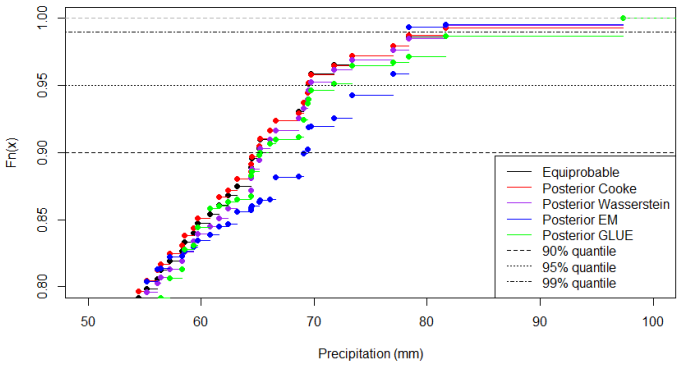

Figure 5 – Upper tail of empirical cumulative distribution functions of pooled annual maximum daily rainfall (mm) for Quebec from 2016 to 2021 with different weighting methods

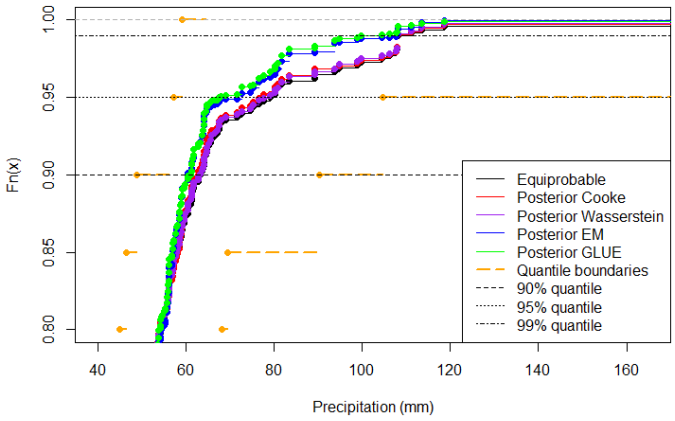

Figure 6 – Upper tail of empirical cumulative distribution functions of pooled annual maximum daily rainfall (mm) for Montreal from 2001 to 2020 with different weighting methods, and minimum and maximum boundaries

From this reasoning, Figure 6 illustrates the CDF obtained with each combination method in Montreal between 2001 and 2020 compared to the minimum and maximum boundaries of quantiles, where the period is expanded to 20 years to allow for empirical quantiles from each expert in a short enough period that precipitation is not expected to change significantly. We notice that the combination methods are grouped within a much narrower range than the theoretical boundaries from the minimum and maximum projections.

We can suppose that the weights provided by the different combination methods will improve the variance around a quantile estimate compared to having no information about each expert. While we cannot obtain this variance mathematically, we can use bootstrap resampling to compare the quantile distribution under each scenario. Figure 7 illustrates the resulting density distributions for the 95^{th} quantile in Montreal between 2001 and 2020. In keeping with intervals presented in Climate Data Canada, the 10% and 90% quantiles of the distribution supposing no information about experts are shown (corresponding to the equiprobable scenario), which can be thought of as the lower and upper bounds that a user with no evaluation of the expert models might consider as plausible. We notice that the two BMA methods differ largely from the other two methods, with modes lying outside the 10%-90% boundaries, while the other methods are more similar to not evaluating experts, particularly for the 95^{th} quantile.

This difference is driven by the same phenomenon as the difference in weight attribution. BMA methods rely on the assumption that residuals between projections and real data follow a normal distribution, whereas Cooke's method and IDW using Wasserstein distance use the distance between (cumulative) densities of the projections and real data. If expert distributions have jumps in their CDFs, this will cause aggregation for both Cooke and Wasserstein, leading to these models receiving little weight.

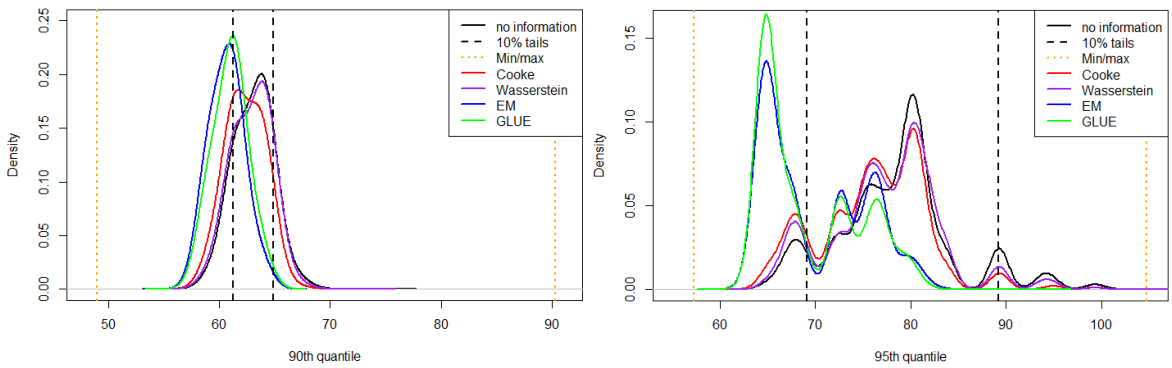

Figure 7 – Comparison of bootstrap densities under different combination methods for the 90th quantile (left) and 95th quantile (right) in Montreal between 2001 and 2020 for 10000 iterations

Table 1 – Comparison of mean and variance of uniform weight attribution and model combination weights for Montreal and Quebec from 2001 to 2020 at the 95^{th} quantile

	Montreal		Quebec	
	Mean	Variance	Mean	Variance
No information	78.4	40.2	72.3	3.4
Cooke	76.4	33.1	72.2	4.3
Wasserstein	77.4	34.8	71.3	3.9
EM	70.0	28.4	70.1	3.7
GLUE	69.3	26.9	70.0	4.0

Nonetheless, the residuals between these experts' projections and real data might still be small, such that BMA methods will attribute larger weight to these models. These different weights cause the gap between quantile values of BMA methods compared to the other methods, as observed in Figure 6. Given the similarity in results between the non-parametric methods using densities, and the BMA methods using residuals as in Figure 7, it is natural to suppose that keeping only one method using densities and another using residuals provides sufficient information for analysis purposes.

Moreover, these combination methods allow for alternate confidence bounds based on an evaluation of expert models as opposed to supposing all expert projections are equally likely. Table 1 also highlights the reduction in variance for the 95^{th} quantile in Montreal, while the much lower variance is similar for all methods in Quebec.

Applying the same exercise to multiple grid cells within the Montreal region, we can calculate the resulting ARF for each method. Given that we observe a 10% difference in 95^{th} quantiles between methods, we can expect different weights to yield significantly different ARF curves.

It is worth noting that directly using quantiles found with model combination methods can yield nonsensical results when computing ARFs. This is because the spatiotemporal

relation between the full region $Y_{\vec{\tau},A,R}$ and the underlying grid cells $Y_{\vec{\tau},x,R}$ for each expert's projection is broken when comparing a weighted average of $y^{(e)}_{\vec{\tau},x,R}$ and $y^{(e)}_{\vec{\tau},A,R}$, leading to ARFs possibly exceeding 1. From a point-to-area point of view, this would not make sense, seeing as a whole area cannot have more intense precipitation than its maximal component, limiting the applicability of such a method. This effect is lessened by using the same weights for all grid cells within an area.

From the significant variability in higher quantiles observed in the previous figures depending on the weights attributed to model projections, we choose to study percentage changes in ARF and quantiles because they yield more comparable information between the different combination methods than actual quantile and ARF values. Mathematically, the modelled quantile change for area A corresponds to $\Delta_{quant} = Y_{\vec{\psi},A,R}/Y_{\vec{\tau},A,R}$, and the ARF change to $\Delta_{ARF} = ARF_{A,R,\vec{\psi}}/ARF_{A,R,\vec{\tau}}$ for future period $\vec{\psi}$ and current period $\vec{\tau}$.

Using the quantile boundaries found previously, we can establish boundaries for possible quantile change by comparing the future maximum to the current minimum, and vice-versa for the minimum possible change. This exercise is not well-defined for ARFs, since the area value depends on the underlying grid cells, and so we cannot for example use the highest area quantile with the lowest grid quantiles, as this would not make sense from a rainfall perspective. Keeping the same 20-year period, we compare it to a near-future period of 2011-2030 and a far future of 2071-2090. The idea behind comparing two future periods is that the variability in near future should be lower than for a later period. Figures 8 and 9 show the change in quantiles and ARFs in Montreal for the near future and far future at a 1 in 20 year return level. While we observe the expected change in variability for quantiles, Figure 9 shows that change in ARF does not significantly vary between near and far projections. This could be explained by looking at the underlying composition of the ARF, where

$$\begin{aligned}\Delta_{ARF} = ARF_{A,R,\vec{\psi}}/ARF_{A,R,\vec{\tau}} &= \frac{\left(\frac{Y_{\vec{\psi},A,R}}{\frac{1}{\mathrm{card}(X)}\sum_{x\in X} Y_{\vec{\psi},x,R}}\right)}{\left(\frac{Y_{\vec{\tau},A,R}}{\frac{1}{\mathrm{card}(X)}\sum_{x\in X} Y_{\vec{\tau},x,R}}\right)} \\ &= \frac{Y_{\vec{\psi},A,R}}{Y_{\vec{\tau},A,R}}\frac{\sum_{x\in X} Y_{\vec{\tau},x,R}}{\sum_{x\in X} Y_{\vec{\psi},x,R}} = \Delta_{quant}\frac{\sum_{x\in X} Y_{\vec{\tau},x,R}}{\sum_{x\in X} Y_{\vec{\psi},x,R}},\end{aligned}$$

such that the first ratio is the change in quantiles, but the second ratio has the current period and future period inverted, suggesting that it will be approximately inversely proportional to the quantile change. As such, the two ratios will cancel out, other than the random noise between different grid cell precipitation, which is what we observe in Figure 9. The fatter tails for Bayesian methods are induced by the distribution of quantile change, which is less centered around a mode, as seen in Figure 8.

The same idea is applied to Quebec in Appendix II, where all methods generally agree, and the Bayesian quantile change projections are more centered around their mode than for Montreal, such that the ARF change projection has smaller tails. The distributions

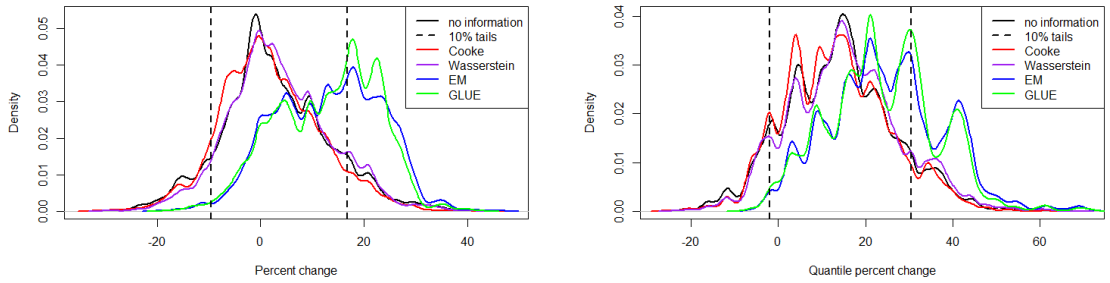

Figure 8 – Distribution of projected quantile change at a 1 in 20 year return level in Montreal between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

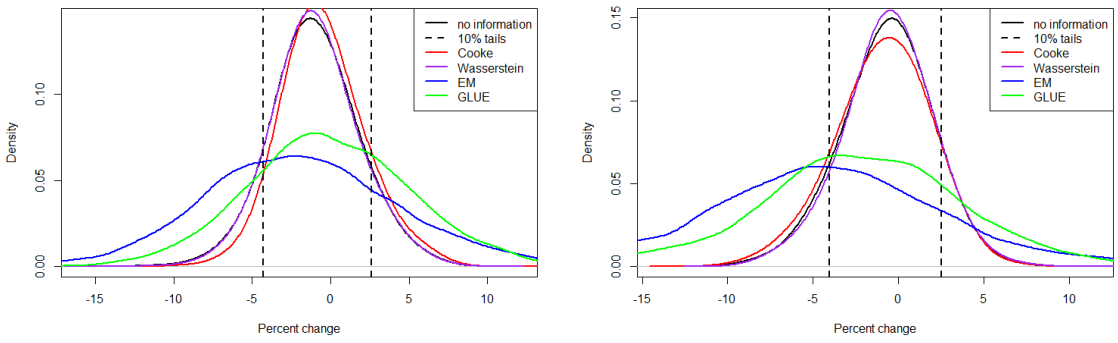

Figure 9 – Distribution of projected ARF change at a 1 in 20 year return level in Montreal between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

resulting from different combination methods can provide valuable information about the uncertainty of projections, where for example in this case the confidence level is higher regarding Quebec projections than Montreal projections. Moreover, compared to the 10% to 90% confidence bounds usually presented, we see that the resulting distributions from combination methods provide alternate bounds based on an evaluation of expert projections. In an actuarial context, this could be very important as it can highlight whether a projection is too conservative or not conservative enough.

Figures 10 and 11 compare the mean percentage change in ARF and quantiles respectively for a 1 in 20 year return level for Montreal between Cooke's method and BMA-EM, divided into approximately 24km x 22km areas. These two methods are chosen to illustrate the substantial variation between a density-based method and a residuals-based method. For example, both methods project increases in quantile, but one projects a 10% increase with little change to the ARF, while the other projects a 22% increase with a reduction to the ARF. From a risk management perspective, this would imply differing scenarios of a moderate increase with similar spatial distribution and a heavier increase with more localised precipitation, which can lead to different losses (see for example Cheng et al. (2012) and American Academy of Actuaries (2020)).

Flood losses provide a particular example of how Cooke's method and Bayesian model averaging with expectation-maximisation would lead to different loss projections. While the link between extreme rainfall and flooding is complex, the difference in scenarios between Cooke's method and BMA-EM allows for a theoretical discussion of its impact for an actuary. Through a combination of hydrological and hydraulic models such as Hydrotel (Fortin et al., 2001), HEC-RAS (Brunner, 2016) or the Hillslope Link Model (Demir and Krajewski, 2013), one can produce discharge flood projections under different rainfall scenarios. Breinl et al. (2021) used elasticity to illustrate the relationship between extreme precipitation and flooding, where depending on ground dampness, an increase in precipitation will have an at least equivalent increase in river discharge, leading to increased flood severity. Supposing that the reduction in ARF will mitigate the impact of an increase in quantiles due to more localised rainfall, such that for example we have an approximately 7% and 19% increase under respectively the Cooke and BMA-EM scenarios, the relationship between discharge and rainfall would clearly imply a greater risk of increased flood losses in the latter case.

Using a hierarchical model such as the one used by Boudreault et al. (2020), flood intensities are associated to different levels of discharge, and their respective probabilities are established from frequency analysis. In their study, the second and third levels of flood intensities have discharges of $1570\text{m}^3/\text{s}$ and of $1740\text{m}^3/\text{s}$ respectively, with occurrence probabilities of 0.01496 and 0.00842. This 10.8% difference in discharge is lower than the projected increase in extreme precipitation using BMA-EM, which is not the case for Cooke's method. All else being equal, the probability of observing more severe flooding in the BMA-EM scenario would increase relative to the Cooke scenario. This change in probability can then be used to calculate premiums and/or reserves for flooding, where BMA-EM would lead to a more conservative estimate than the other method in this case. In a changing climate perspective, the range of scenarios resulting

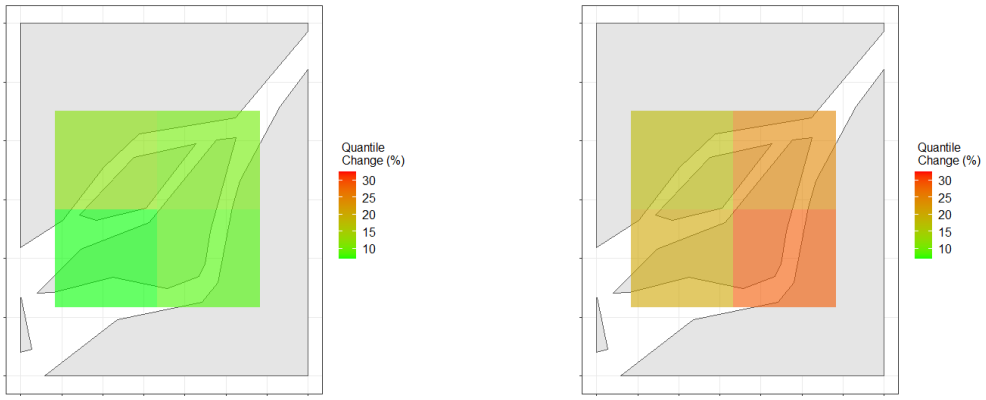

Figure 10 – Percentage change in quantiles for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Montreal using Cooke's method (left) and BMA-EM (right)

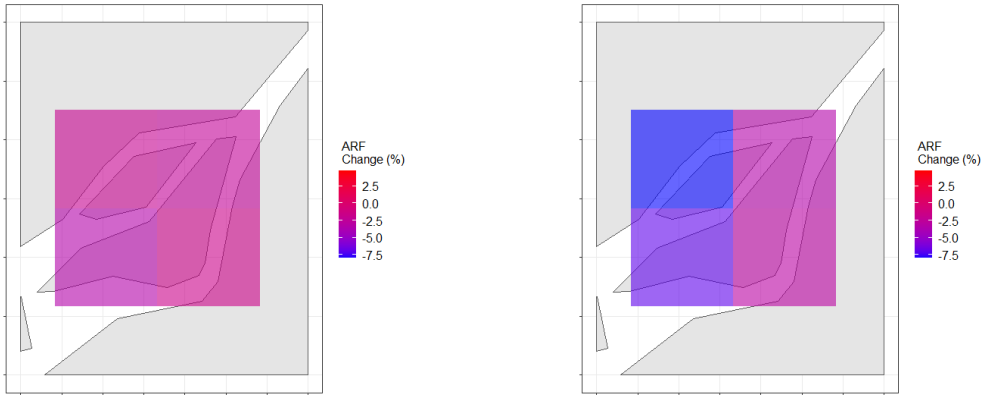

Figure 11 – Percentage change in ARFs for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Montreal using Cooke's method (left) and BMA-EM (right)

from different combination methods becomes even more important to have a fuller understanding of the impact of climate change on insurable losses. An analyst using only one method would fail to obtain a complete picture of projection uncertainty, and may find themselves being overconfident in the result of a single combination method.
Similar graphics are available in Appendix III for Quebec. Projections for this city are much more similar across methods, leading to smaller confidence intervals in this case. In summary, we see that the different combination methods considered can yield varying sets of weights, or probabilities, assigned to each model, which impacts projected quantiles. From the similarities between methods using densities compared to methods using residuals, we see that one only needs to use one method from each approach to obtain a picture of the underlying projection uncertainty, and the difference between the approaches provides a measure of this uncertainty. In cases where methods agree, one

could more confidently reach conclusions about the analysed data, but in cases where methods disagree, using only one method would fail to capture projection uncertainty. Moreover, combination methods can yield alternate confidence bounds based on an evaluation of expert models, and offer an improved pooling projection over considering all expert projections as equally likely.

4 Conclusion

In this paper, we applied model combination methods to the pooling approach used by Innocenti et al. (2019) to highlight the resulting difference in quantile estimation and areal reduction factor (ARF) calculation. More specifically, we compared Cooke's method, an inverse distance weighting approach, and two Bayesian model averaging approaches to equiprobable pooling when considering precipitation annual maxima.
Our main focus was to investigate the impact, if any, of various model combination methods on quantiles obtained through pooling, and therefore on the resulting ARFs. We considered two non-parametric approaches, namely Cooke's method as well as Inverse Distance Weighting using Wasserstein distance, in addition to Bayesian Model Averaging using an Expectation-Maximisation algorithm, and a Generalised Likelihood Uncertainty Estimation algorithm. The choice of these methods was motivated by having an approach not requiring much information, an easy to use and intuitive method, and Bayesian approaches frequently used in recent studies.
We focused on a 1 in 20-year return level in Montreal and Quebec to show that different weighting methods lead to significantly different results for both quantiles and ARF curves. By considering the projected percentage change in quantiles and ARFs from 2001-2020 to 2071-2090, the variability in results offered insight into the uncertainty of future projections, where results seemed to generally agree around Quebec, whereas results varied significantly between methods for Montreal. This suggests that despite past literature demonstrating that combination methods significantly increase accuracy (Clemen, 1989), one should use more than one combination method, given that a single method may lead to overconfidence about projections. Moreover, it may be sufficient to compare a method using densities to another using residuals to obtain alternate confidence bounds instead of the standard bounds used in weather projections. Combination methods can be of particular interest to actuaries in a changing climate context to have a better understanding of the impact of projected changes on potential losses.
A limitation of this study is that the combination methods used ignored the potential dependence between different expert projections by assuming independence between experts. The new method of Bayesian Predictive Synthesis presented in McAlinn and West (2019) would be an interesting extension, as it is a generalisation of Bayesian Model Averaging taking dependence into account in a time-series context.

References

American Academy of Actuaries (2020). Actuaries climate risk index. https://www.actuary.org/sites/default/files/2020-01/ACRI.pdf.

Beven, K. and Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of hydrology*, 249(1-4):11–29.

Boudreault, M., Grenier, P., Pigeon, M., Potvin, J.-M., and Turcotte, R. (2020). Pricing flood insurance with a hierarchical physics-based model. *North American Actuarial Journal*, 24(2):251–274.

Breinl, K., Lun, D., Müller-Thomy, H., and Blöschl, G. (2021). Understanding the relationship between rainfall and flood probabilities through combined intensity-duration-frequency analysis. *Journal of Hydrology*, 602:126759.

Broom, B. M., Do, K.-A., and Subramanian, D. (2012). Model averaging strategies for structure learning in bayesian networks with limited data. *BMC bioinformatics*, 13(13):1–18.

Brunner, G. W. (2016). Hec-ras, river analysis system hydraulic reference manual.

Cheng, C. S., Li, Q., Li, G., and Auld, H. (2012). Climate change and heavy rainfall-related water damage insurance claims and losses in ontario, canada. *Journal of Water Resource and Protection*, 4(2):49–62.

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

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

Climate Data Canada (2018). Climate Data for a Resilient Canada. https://climatedata.ca/.

Cooke, R. et al. (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand.

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

Copernicus Climate Change Service (2022). Era5 hourly data on single levels from 1979 to present. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form.

Darbandsari, P. and Coulibaly, P. (2019). Inter-comparison of different bayesian model averaging modifications in streamflow simulation. *Water*, 11(8):1707.

Demir, I. and Krajewski, W. F. (2013). Towards an integrated flood information system: centralized data access, analysis, and visualization. *Environmental modelling & software*, 50:77–84.

Djeundje, V. B. (2022). On the integration of deterministic opinions into mortality smoothing and forecasting. *Annals of Actuarial Science*, pages 1–17.

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. In *ICML*, volume 747, pages 223–230. Citeseer.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., and Villeneuve, J.-P. (2001). Distributed watershed model compatible with remote sensing and gis data. i: Description of model. *Journal of hydrologic engineering*, 6(2):91–99.

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1):1–28.

Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

Gracianti, G., Zhou, R., and Li, J. (2021). Spatial-temporal modelling of wind speed-a vine copula based approach.

Hammitt, J. K. and Zhang, Y. (2012). Combining experts' judgments: Comparison of algorithmic methods using synthetic data. *Risk Analysis: An International Journal*, 33(1):109–120.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417.

Hu, Q., Li, Z., Wang, L., Huang, Y., Wang, Y., and Li, L. (2019). Rainfall spatial estimations: A review from spatial interpolation to multi-source data merging. *Water*, 11(3):579.

Huang, F. and Browne, B. (2017). Mortality forecasting using a modified continuous mortality investigation mortality projections model for china i: Methodology and country-level results. *Annals of Actuarial Science*, 11(1):20–45.

Innocenti, S., Mailhot, A., Leduc, M., Cannon, A. J., and Frigon, A. (2019). Projected changes in the probability distributions, seasonality, and spatiotemporal scaling of daily and subdaily extreme precipitation simulated by a 50-member ensemble over northeastern north america. *Journal of Geophysical Research: Atmospheres*, 124(19):10427–10449.

Insurance Bureau of Canada (2022). Severe weather in 2021 caused $2.1 billion in insured damage. http://www.ibc.ca/ns/resources/media-centre/media-releases/severe-weather-in-2021-caused-2-1-billion-in-insured-damage.

Jacobs, R. A. (1995). Methods for combining experts' probability assessment. *Neural Computation*, 7:867–888.

Johnson, M. C. (2017). *Bayesian predictive synthesis: Forecast calibration and combination*. PhD thesis, Duke University.

Kantorovitch, L. and Rubinštein, G. (1958). On a space of completely additive functions (in russian). *Vestnik Leningrad Univ.*, 13:52–59.

Kodra, E., Bhatia, U., Chatterjee, S., Chen, S., and Ganguly, A. R. (2020). Physics-guided probabilistic modeling of extreme precipitation under climate change. *Scientific reports*, 10(1):1–11.

Le, P. D., Davison, A. C., Engelke, S., Leonard, M., and Westra, S. (2018). Dependence properties of spatial rainfall extremes and areal reduction factors. *Journal of hydrology*, 565:711–719.

Li, J., Sharma, A., Johnson, F., and Evans, J. (2015). Evaluating the effect of climate change on areal reduction factors using regional climate model projections. *Journal of Hydrology*, 528:419–434.

Massoud, E., Lee, H., Gibson, P., Loikith, P., and Waliser, D. (2020). Bayesian model averaging of climate model projections constrained by precipitation observations over the contiguous united states. *Journal of Hydrometeorology*, 21(10):2401–2418.

McAlinn, K., Aastveit, K. A., Nakajima, J., and West, M. (2020). Multivariate bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association*, 115(531):1092–1110.

McAlinn, K. and West, M. (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210(1):155–169.

Mendel, M. B. and Sheridan, T. B. (1989). Filtering information from human experts. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):6–16.

Pesenti, S. M., Bettini, A., Millossovich, P., and Tsanakas, A. (2021). Scenario weights for importance measurement (swim)–an r package for sensitivity analysis. *Annals of Actuarial Science*, 15(2):458–483.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Richman, R. (2021). Ai in actuarial science–a review of recent advances–part 2. *Annals of Actuarial Science*, 15(2):230–258.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524.

Svensson, C. and Jones, D. A. (2010). Review of methods for deriving areal reduction factors. *Journal of Flood Risk Management*, 3(3):232–245.

Wagner, P. D., Fiener, P., Wilken, F., Kumar, S., and Schneider, K. (2012). Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. *Journal of Hydrology*, 464:388–400.

Zhu, J., Forsee, W., Schumer, R., and Gautam, M. (2013). Future projections and uncertainty assessment of extreme rainfall intensity in the united states from an ensemble of climate models. *Climatic Change*, 118(2):469–485.

Appendix I - Expectation-Maximisation Bayesian Model Averaging algorithm

The following table illustrates the algorithm followed for expectation-maximisation under bayesian model averaging for n experts and M quantiles, where $y^{(e)}_{\vec{\tau},x,m}$ is the m^{th} quantile of vector $\vec{y}^{(e)}_{\vec{\tau},x}$, $y_{\vec{\tau},x,m}$ is the m^{th} quantile of real values, σ^2_e and w_e are respectively the variance and weight for each expert's model, $\phi(y_{\vec{\tau},x,m}|y^{(e)}_{\vec{\tau},x,m},\sigma^2)$ is the density of a normal distribution evaluated at $y_{\vec{\tau},x,m}$ with mean $y^{(e)}_{\vec{\tau},x,m}$ and variance σ^2, and θ is a vector of parameters s.t. $\theta = \{w_e,\sigma^2_e, e = 1,\dots,n\}$.

Algorithm 2: Expectation-Maximisation Bayesian Model Averaging

1: Initialize variance and weights as

$$\sigma^{2(0)} = \frac{1}{nM}\sum_{m=1}^{M}\sum_{e=1}^{n}\left(y_{\vec{\tau},x,m} - y^{(e)}_{\vec{\tau},x,m}\right)^2,$$
$$w_e^{(0)} = 1/n \ \forall e.$$

2: Calculate initial likelihood as

$$l(\theta^{(0)}) = \sum_{m=1}^{M}\log\left(\sum_{e=1}^{n} w_e^{(0)}\phi(y_{\vec{\tau},x,m}|y^{(e)}_{\vec{\tau},x,m},\sigma^{2(0)})\right).$$

3: **while** $|l(\theta^{(j)}) - l(\theta^{(j-1)})| > \beta$, **do**

4: Obtain proportion from normal densities for each expert e and quantile m as

$$z^{(j)}_{e,m} = \frac{w_e^{(j-1)}\phi(y_{\vec{\tau},x,m}|y^{(e)}_{\vec{\tau},x,m},\sigma^{2(j-1)})}{\sum_{e=1}^{n} w_e^{(j-1)}\phi((y_{\vec{\tau},x,m}|y^{(e)}_{\vec{\tau},x,m},\sigma^{2(j-1)})}.$$

5: Update weights and variance to each expert, i.e.

$$w_e^{(j)} = \frac{1}{M}\sum_{m=1}^{M} z^{(j)}_{e,m}$$
$$\sigma^{2(j)}_e = \frac{\sum_{m=1}^{M} z^{(j)}_{e,m}(y_{\vec{\tau},x,m} - y^{(e)}_{\vec{\tau},x,m})^2}{\sum_{m=1}^{M} z^{(j)}_{e,m}}.$$

6: Calculate updated likelihood as

$$l(\theta^{(j)}) = \sum_{m=1}^{M}\log\left(\sum_{e=1}^{n} w_e^{(j)}\phi(y_{\vec{\tau},x,m}|y^{(e)}_{\vec{\tau},x,m},\sigma^{2(j)})\right).$$

7: Update iteration count $j = j + 1$.

8: **end while**

9: Update the probability associated to each expert as $\Pr(\mathcal{M} = \mathcal{M}_e|\vec{y}_{\vec{\tau},x}) = w_e^{(j)}$.

10: Calculate posterior distribution as

$$\Pr(Y_{\vec{\psi},x} = y|\vec{y}_{\vec{\tau},x}) = \textstyle\sum_{e=1}^{n}\Pr(Y_{\vec{\psi},x} = y|\mathcal{M}_e)\Pr(\mathcal{M} = \mathcal{M}_e|\vec{y}_{\vec{\tau},x}).$$

Appendix II - Quantile and ARF changes bootstrap distribution for a 1 in 20 year return level for Quebec

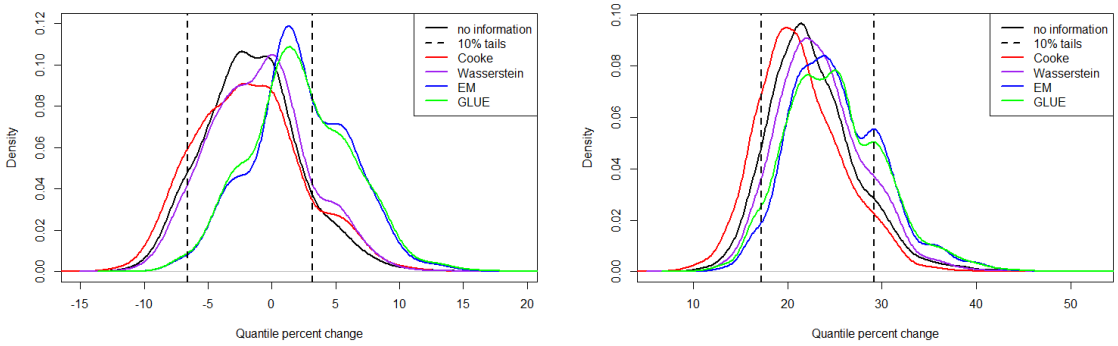

Figure 12 – Distribution of projected quantile change at a 1 in 20 year return level in Quebec between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

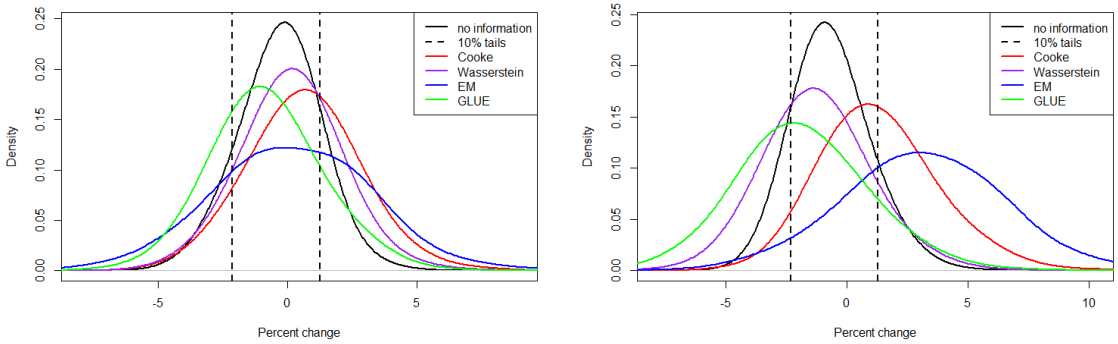

Figure 13 – Distribution of projected ARF change at a 1 in 20 year return level in Quebec between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

Appendix III - Quantile and ARF percent changes for a 1 in 20 year return level for Quebec

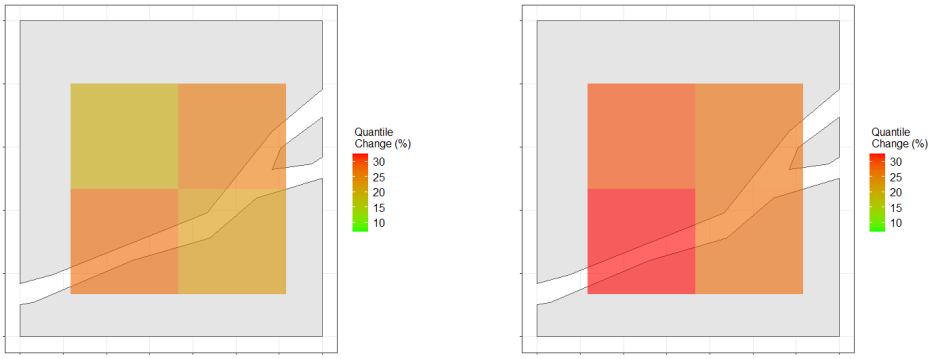

Figure 14 – Percentage change in quantiles for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Quebec using Cooke's method (left) and BMA-EM (right)

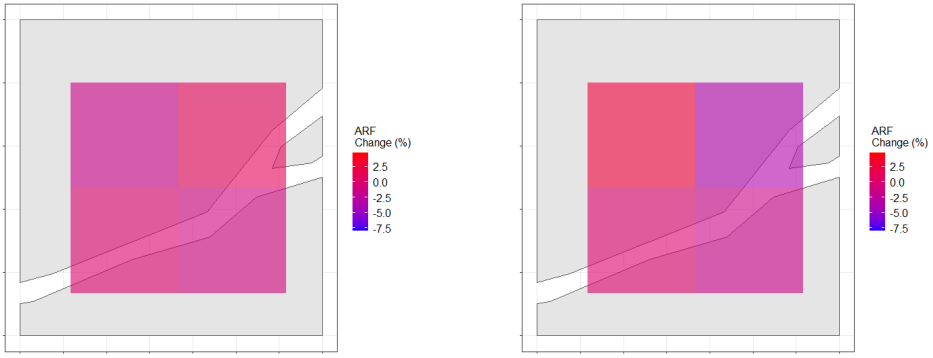

Figure 15 – Percentage change in quantiles for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Quebec using Cooke's method (left) and BMA-EM (right)