Advancements in model combination and uncertainty quantification with applications in actuarial science

Sébastien Jessup

A Thesis

in the Department of

Mathematics & Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Mathematics) at

Concordia University

Montréal, Québec, Canada

October 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Sébastien Jessup**

Entitled: **Advancements in model combination and uncertainty quantification with applications in actuarial science**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Mathematics)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Tatyana Koreshkova

_____ External Examiner
Dr. Anne-Sophie Charest

_____ Examiner
Dr. Frédéric Godin

_____ Examiner
Dr. Yang Lu

_____ Examiner
Dr. Jean-Philippe Boucher

_____ Supervisor
Dr. Mélina Mailhot

_____ Co-supervisor
Dr. Mathieu Pigeon

Approved by _____
Lea Popovic, Graduate Program Director
Department of Mathematics & Statistics

_____ 2024

_____
Pascale Sicotte, Dean
Faculty of Arts and Science

Abstract

Advancements in model combination and uncertainty quantification with applications in actuarial science

Sébastien Jessup, Ph.D.

Concordia University, 2024

In this thesis, we focus on model combination, incorporating elements of uncertainty quantification to address different actuarial science issues. We first tackle the issue of overconfidence from a single model combination approach, highlighting how different combination assumptions can lead to different conclusions about the predicted variable. This is illustrated with an extreme precipitation example for the regions of Montreal and Quebec. We then focus on Bayesian model averaging (BMA), a very popular model combination technique relying on Bayes' theorem to attribute weights to models based on the likelihood that the observed data comes from the models considered. We propose a correction to the classical expectation-maximisation algorithm to account for data uncertainty, where we assume that the observed data is in fact not the only possible observable data. We then generalise our method to include Dirichlet regression, allowing for combination weights to vary depending on risk characteristics. These BMA approaches are applied to a simulation study as well as a simulated actuarial database and are shown to be very promising, as they allow for a more formal model combination framework for combining actuarial reserving methods in a smooth way based on predictive variables. Next, we adapt Bayesian model averaging using Generalised Likelihood Uncertainty Estimation to extreme value mixture models, and show that this modification allows for identifiying the "best" extreme value threshold, although a combination of models will outperform the single best mixture model. This is illustrated using the Danish reinsurance dataset. Finally, we show that the generalised BMA algorithm can be used to identify flexible extreme value thresholds depending on predictive variables. We use this generalised mixture model combination on a recent dataset from a Canadian automobile insurer.

Acknowledgments

First, I would like to thank my supervisors, Professors Mélina Mailhot and Mathieu Pigeon, for all of their advice and mentorship. I cannot count the hours spent with them brainstorming concepts and discussing diverse issues on my mind. They helped shape me into the researcher I am becoming and believed in me in times where I doubted myself. Un énorme merci à vous deux, et au plaisir de continuer à collaborer!

Next, my thanks go out to all the professors and colleagues with whom I discussed research ideas through the years, particularly Professor James A. Goulet, who played a role in forming some of the ideas in Chapter 3. Sharing ideas with people from diverse backgrounds helped me approach problems from different angles, allowing for new perspectives and interesting solutions. I will definitely be carrying this mentality forward.

Thank you to the reviewers of the articles in Chapters 2 and 3 for providing thoughtful and insightful comments that helped to improve our work. Further thanks to the evaluation committee for reviewing this thesis. I am also thankful for the financial support of Concordia University and the Chaire Co-operators en analyse des risques actuariels.

A special thank you goes to my family. To my parents, for their infallible support in everything I do. To my brother, for trying to understand what I was doing and throwing ideas at me. To my partner, for her love and support. Je vous aime!

And finally, thanks to those who will take the time to read this thesis. It is the result of quite a bit of work, hopefully you will find some parts of it interesting!

Contributions of Authors

This thesis is based on three research articles:

I Jessup, S., Mailhot, M., & Pigeon, M. (2023). Impact of combination methods on extreme precipitation projections. *Annals of Actuarial Science*, 17(3), 459-478.

Jessup is responsible for a substantial portion of the analysis and the primary portion of writing, with supervision by Mailhot and Pigeon.

II Jessup, S., Mailhot, M. & Pigeon, M. Uncertainty in heteroscedastic Bayesian model averaging.

This joint work has been resubmitted to the Insurance: Mathematics and Economics journal after revisions. Jessup is responsible for a substantial portion of the analysis and the primary portion of writing, with supervision by Mailhot and Pigeon.

III Jessup, S., Mailhot, M. & Pigeon, M. Flexible extreme thresholds through generalised Bayesian model averaging.

This joint work is still in preparation and constitutes a draft. Jessup is responsible for a substantial portion of the analysis and the primary portion of writing, with supervision by Mailhot and Pigeon.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In most statistical problems, we study data in order to identify patterns and make predictions. This is particularly useful in fields such as insurance, where we face an inverted production cycle. This cycle stems from charging a premium for insurance coverage before knowing the actual cost of claims. Insurers use multiple different methods to anticipate future claims, among which using historical data and statistical models to set premiums accurately. If premiums are set too low, the insurer may not have sufficient funds to cover the claims, leading to financial instability. Conversely, if premiums are set too high, it can drive away potential customers, making the insurance product less competitive in the market. Thus, it is essential to accurately model risk to ensure that the insurer has sufficient funds to cover the insureds' losses, while also providing a fair premium to insureds. Moreover, in the context of climate change, catastrophic risks are becoming more prevalent each year and require special focus. These catastrophic events can include natural disasters like hurricanes, floods, and wildfires, which are occurring with increasing frequency and intensity. The financial impact of these events can be staggering, as only a handful of catastrophes can account for a significant portion of an insurer's losses. For instance, in 2023, 23 events in Canada resulted in over $30 million in damages each, collectively accounting for over a quarter of the $3.1 billion in total Canadian insured losses for that year (CatIQ, 2024).

This heightened need for accurate risk modeling necessitates extensive data analysis and inference. In order to make these inferences, the classical approach is to fit multiple models to the observed data, then evaluate the predictive power of each model, before finally keeping the single

model that best fits the data. This selection implies that we consider the chosen model as fully representative of the true model generating the observed data. However, this approach can be overly simplistic and potentially misleading, as it assumes that the selected model captures all the underlying complexities of the data, which is rarely the case in real-world scenarios.

Natural questions that arise from this process, and that are central to this thesis, are the following: What if the chosen model is not actually the true model? What if the disregarded models contain useful information? Is there then a way to use the information from multiple models simultaneously? These questions highlight the limitations of traditional model selection techniques and suggest the need for more sophisticated methods that can combine information from multiple models.

1.1 Uncertainty quantification

To efficiently combine information from multiple models, addressing notions of uncertainty is essential. Uncertainty can be broadly categorised into two types: random and model error, also called aleatoric and epistemic uncertainty in machine learning (Hüllermeier and Waegeman, 2021). Consider a general model

$$Y = f(x) + \epsilon. \tag{1}$$

In this model, random error is represented by ϵ, capturing the irreducible noise inherent in nearly any dataset. On the other hand, model error can be reduced. This type of error can be decomposed into bias and variance components. It is related to the inaccuracy of $f(x)$ and can be minimised by addressing the well-known bias-variance tradeoff (see, for example, Belkin et al. (2019)). This is illustrated with a classical example in Figure 1.1, where data is generated with a sine wave and random noise around this wave. We can see a high bias model that is not sufficiently adjusted to the data, and an overfitted model that has no bias, but large variance.

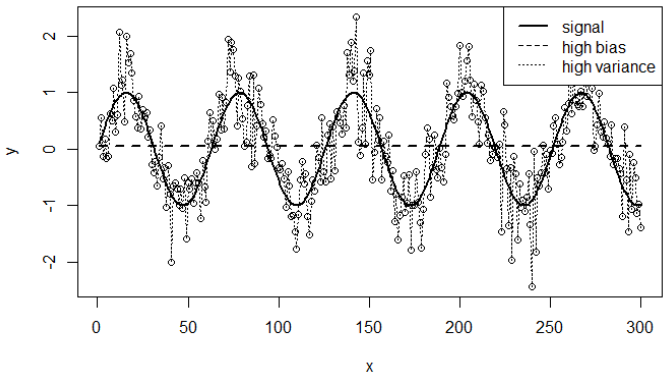

Figure 1.1: The bias-variance tradeoff

In Chapter 2, we first address model error in the context of integrating information from multiple models. Combining models is widely recognised as an effective strategy to reduce bias, thereby enhancing model accuracy (e.g. Webb and Zheng (2004)). However, as will be discussed in 1.2, different combination methods rely on varying assumptions, which can mitigate model error in different ways. We will demonstrate that, when examining extreme precipitation, these differences can lead to substantial variations in projections of both intensity and spatial distribution. As such, Chapter 2 focuses on the uncertainty of model combination methods.

Next, we turn our attention to random error in a heteroscedastic setting. In Figure 1.1, random error is depicted as constant or homoscedastic, where the noise is independent of the variable x. However, in actuarial contexts, losses are influenced by risk characteristics, leading to varying levels of uncertainty. In Chapter 3, we discuss how a commonly used combination method, which integrates information from multiple models using predictive variables, often fails to account for random variation. In this chapter, we propose a way to model residuals in order to quantify random error based on predictive variables.

1.2 Model Combination

With these notions of uncertainty quantification in hand, we can address the central questions of the thesis through model combination. This is a very broad field with numerous different approaches. For example, Cooke et al. (1991) offered a review of early expert combination techniques,

while Kotsiantis et al. (2006) and Mohandes et al. (2018) review machine learning combination techniques. We will focus on linear combination, which aims to assign weights $w_m \in [0, 1]$ to M different models with $\sum w_m = 1$ such that

$$f(y) = \sum_{m=1}^{M} w_m f_m(y), \tag{2}$$

where $f_m(y)$ is the distribution under model $\mathcal{M}_m$. In this kind of setting, we aim to establish weights such that $f(y)$ is most accurate.

1.2.1 Non-parametric methods

The first general family of methods for obtaining combination weights is non-parametric approaches. These methods rely on scoring rules, which broadly aim to evaluate the quality of predictions compared to actual observations. For distributions $\mathbb{P}$ and $\mathbb{Q}$, the expected score of a rule S under $\mathbb{P}$, given the predictive distribution $\mathbb{Q}$, is defined as

$$S(\mathbb{Q}, \mathbb{P}) = \int S(\mathbb{Q}, \omega) d\mathbb{P}(w). \tag{3}$$

A scoring rule S is said to be proper if $S(\mathbb{P}, \mathbb{P}) \geq S(\mathbb{Q}, \mathbb{P}) \; \forall \; \mathbb{P}, \mathbb{Q}$. It is strictly proper if equality holds if and only if $\mathbb{P} = \mathbb{Q}$ (Gneiting and Raftery, 2007).

In practice, we can evaluate the score of a distribution $\mathbb{Q}$ empirically by looking at n realisations of a random variable Y, $\{y_1, \ldots, y_n\}$ from $\mathbb{P}$. For example, one well-known strictly proper scoring rule is the logarithmic score, for which a score S is attributed to each observation y_i under distribution $\mathbb{Q}$ such that

$$S(\mathbb{Q}, y_i) = -\ln(f_{\mathbb{Q}}(y_i)) \tag{4}$$

and the expected score is then approximated as

$$\hat{S}(\mathbb{Q}, \mathbb{P}) = \frac{1}{n} \sum_{i=1}^{n} S(\mathbb{Q}, y_i). \tag{5}$$

This score is minimised if $\mathbb{Q} = \mathbb{P}$, where $\mathbb{P}$ is the real distribution. A variation of this scoring rule is used by Cooke et al. (1991), who attributes a score based on a calibration component and an entropy component. The first component evaluates how well the predictions fit the observations, while the second component evaluates whether the prediction is informative. This method will be explored in more detail in Chapter 2.

Importantly, the method proposed by Cooke et al. (1991) does not require a full distribution, but rather allows for the incorporation of expert opinions about specific values, which is particularly useful in fields such as actuarial science. Indeed, experts often have an idea of what constitutes a low, medium, and large claim, without being able to provide a full loss model. In such cases, non-parametric methods are necessary to attribute weights to expert predictions.

1.2.2 Parametric methods

In contrast to non-parametric methods, the second family of methods, namely parametric approaches, which require specifying a complete model, offer the advantage of Bayesian updating, where weights are revised with new observed data. One such approach is Bayesian Model Averaging (BMA), first proposed by Raftery et al. (1997). BMA assigns weights to each model based on the probability that the data originate from those models. Specifically, for a given dataset $\mathcal{D}$, the weights are determined as

$$w_m = \Pr(\mathcal{M}_m|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\mathcal{M}_m)\Pr(\mathcal{M}_m)}{\sum_{l=1}^{M}\Pr(\mathcal{D}|\mathcal{M}_l)\Pr(\mathcal{M}_l)}, \tag{6}$$

where $\Pr(\mathcal{M}_m)$ represents the prior belief concerning model $\mathcal{M}_m$. BMA has gained significant attention across various scientific fields in recent years (Fragoso et al., 2018). Several algorithms are available to implement BMA, two of which are used in Chapters 2 to 4. While very popular, BMA assumes that one of the models must be correct (Hoeting et al., 1999), which can cause convergence issues, as it considers the observed data as fully representative of its underlying model. This concern is addressed in Chapter 3, where we incorporate uncertainty quantification to acknowledge the inherent randomness in data and prevent convergence to a single model.

Additionally to allowing for Bayesian updating, BMA can be modified to incorporate predictive

variables. We propose a modification to a BMA expectation-maximisation algorithm in Chapter 3 that allows for flexible weights based on vectors of characteristics. This is valuable in diverse actuarial contexts, as will be shown in Chapter 3 for reserving, and Chapter 4 for loss modelling.

It is worth noting that BMA works best when the models are independent. In cases where dependencies exist, such as when a hierarchical approach relies on the output of another model, the more complex structure may be favoured, while the initial one could be disregarded. Although this idea is intriguing, it was not explored further and is left for future research.

Another promising approach to model combination in machine learning that is similar to linear combination is the mixture-of-experts (MoE) method, reviewed by Masoudnia and Ebrahimpour (2014). Unlike the previous approaches that compare models spanning the entire sample space, MoE combines localised models over different parts of the sample space, collectively covering the full space. While MoE presents a very interesting and promising direction, it is beyond the scope of this thesis and is also left for future research.

1.3 Extreme values

As we explore these methodological nuances, it becomes clear that insurers are frequently exposed to substantial risks originating from various sources, such as injuries, high-value property, and, as previously discussed, catastrophic events. These risks can lead to significant financial losses and operational challenges. Revisiting the third central question of our study, another important question arises: If we can leverage the information from multiple models, how can we effectively account for these large risks to which insurers are exposed?

To address this question, we must delve into extreme value theory (EVT). Broadly speaking, EVT is the study of extreme values in the tails of data distributions, focusing on the behaviour of very large or very small values. This area of statistical theory is particularly relevant for insurers because it provides the tools to model and predict the occurrence of rare, yet high-impact events. By studying the distribution of extreme values, insurers can better estimate the probabilities and potential magnitudes of significant losses, which is crucial for risk management and decision-making.

There are two ways of modeling these values. The first approach, mostly used when studying

time series, is to separate the data into blocks, and then use properties of rank statistics to study block maxima. The Fisher-Tippett-Gnedenko theorem (Coles et al., 2001) states that for a sequence of iid random variables $X_1, X_2, \ldots, X_n$ with cumulative distribution function $F(x)$, if there exists a sequence of numbers a_n and b_n such that

$$\lim_{n\to\infty} \Pr\left(\frac{\max(X_1, \cdots, X_n) - b_n}{a_n} \leq x\right) = G(x) \tag{7}$$

for a non-degenerate distribution $G(x)$, then F is in the maximum domain of attraction of G, and G must be among three possible families. The generalised extreme value distribution is then defined as

$$G(x) = \begin{cases} \exp\left(-\left[1+\xi\frac{x-\mu}{\sigma}\right]^{-1/\xi}\right) & \text{if } \xi \neq 0 \\ \exp\left(-\exp\left[-\frac{x-\mu}{\sigma}\right]\right) & \text{if } \xi = 0, \end{cases} \tag{8}$$

where μ, σ and ξ are respectively location, scale, and shape parameters. This approach allows for studying annual maxima, which is useful to obtain risk estimates at very high levels such as a 1 in 100 years frequency. Block maxima are implicitly related to Chapter 2, where we use model combination to study precipitation annual maxima without fitting GEV distributions.

The second approach studies exceedance over a high threshold. It is linked to the maxima approach through the Pickands-Balkema-De Haan theorem (Pickands, 1975), which states that if

$$\lim_{u\to\infty} \Pr\left(\frac{Y-u}{c(u)} \leq x | Y > u\right) = G(x) \tag{9}$$

for a non-degenerate function $G(x)$, then $G(x)$ must follow a generalised Pareto distribution such that

$$G(x) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma_u}\right)^{-1/\xi} & \text{for } \xi \neq 0 \text{ and } x > u \\ 1 - \exp\left(-\frac{x-u}{\sigma_u}\right) & \text{for } \xi = 0 \text{ and } x > u, \end{cases} \tag{10}$$

where σ_u and ξ are again scale and shape parameters.

To apply equation (10), a sufficiently high threshold value u is required. The selection of an appropriate threshold is a challenging problem in the literature, with no universally accepted method. Chapter 4 tackles this problem, demonstrating that model combination can not only identify the

correct threshold but also reduce the dependence on specifying a precise threshold value. This is feasible because the GEV and GPD distributions share the same ξ parameter, and their shape parameters are linearly linked through u. Specifically, for a sufficiently high threshold u,

$$\sigma_u = \sigma + \xi(u - \mu). \tag{11}$$

This link implies that any sufficiently high choice of threshold yields a valid GPD distribution, as will be further explained in Chapter 4.

1.4 Contributions

In Chapter 2, we study the inherent uncertainty associated with model combination to quantify the differences in outputs derived from various combination methods. Specifically, we compare two non-parametric model combination methods, namely Cooke's method and inverse distance weighting, with two Bayesian model averaging (BMA) algorithms when applied to extreme precipitation data. These combination methods are used to analyse the changes in tail quantiles and the spatial distribution of extreme daily rainfall in Montreal and Quebec over two time periods between a current period (2001-2020) and a future period (2071-2090). Our findings indicate that the non-parametric combination methods, which compare distributions, yield significantly different projections compared to Bayesian methods that use quantiles. This disparity suggests that relying on a single combination method could lead to overconfidence in projections, as the choice of combination hypothesis can notably alter the results. Furthermore, the confidence intervals derived from model combination methods contradict the standard confidence intervals obtained with extreme precipitation ensembles where all models are weighted equally, underscoring the critical importance of model combination in assessing uncertainty.

The insights gained from studying uncertainty in model combination in Chapter 2 lead naturally to the developments in Chapter 3, where we focus on enhancing BMA to address similar challenges in actuarial applications. In particular, we propose an enhancement to the classical Expectation-Maximisation (EM) algorithm to address its known tendency to converge to a single model. By incorporating data uncertainty into the EM algorithm, we examine model residuals to condition on

random error, enabling the numerical integration of the error by simulating instances of this uncertainty. Additionally, we generalise our method to allow for flexible weights based on predictive variables through Dirichlet regression. Instead of averaging simulations to numerically integrate random error, we treat the simulations as realisations of a Dirichlet random variable. These approaches are illustrated through simulation studies and applied to a simulated actuarial database. The introduction of the flexible weight combination method is novel in the actuarial literature and shows significant promise, enabling smoother transitions between different actuarial reserving methods across various data segments. Furthermore, our proposed method outperforms an existing combination method for aggregate reserve models, highlighting its potential effectiveness.

Together, these chapters highlight the critical role of model combination in both climate risk assessment and actuarial reserving. By exploring uncertainty in Chapter 2 and proposing methodological enhancements in Chapter 3, we provide a framework that improves the reliability of projections in diverse fields, from extreme precipitation forecasts to insurance reserves. The tools developed in these chapters then allow us to examine combinations of mixture models that allow for extreme values in Chapter 4.

In this last chapter, we propose a modification to the Generalised Likelihood Uncertainty Estimation (GLUE) BMA algorithm to identify the optimal Generalised Pareto Distribution (GPD) threshold. By employing a weighted-likelihood approach that gives more weight to tail quantiles, we demonstrate both mathematically and through an example based on the well-known Danish dataset that comparing the two GLUE combination methods facilitates the identification of the best GPD threshold. We further illustrate how model combination methods are preferable to a single mixture model with the correct threshold, highlighting their lower dependence specifying the correct threshold. We use a similar idea to obtain flexible thresholds depending on predictive variables with an adjustment of the method proposed in Chapter 3, where we modify the algorithm to work with mixture models. Our high-level quantile projections are found to be similar to those obtained using a two-step quantile regression method with an actuarial dataset from a Canadian insurer. This new flexible threshold identification method using model combination surpasses existing variable-dependent threshold methods by enabling the determination of the full distribution, rather than just the excess over a high threshold.

In summary, this thesis explores the applications of model combination techniques in actuarial science, emphasizing their role in quantifying uncertainty and managing extreme events. In Chapter 2, we highlight how different model combination methods yield varying results in extreme precipitation projections, underscoring the importance of addressing uncertainty in model outputs. Building on this, Chapter 3 improves BMA by incorporating heteroscedasticity and flexible weights, enhancing its application in actuarial reserving. Finally, Chapter 4 uses elements from the previous chapters by modifying the GLUE BMA algorithm to identify optimal thresholds for extreme value distributions. This chapter also introduces a flexible threshold method based on predictive variables using the method developed in Chapter 3, integrating the advancements in model combination from earlier chapters. Through these advancements, we aim to enhance the accuracy of actuarial models in addressing uncertainty and extreme events.

Chapter 2

Impact of combination methods on extreme precipitation projections

2.1 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 Averaging (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 2.2 provides details regarding parametric and non-parametric model combination methods, Section 2.3 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 2.4 provides concluding comments. Additional material can be found in Appendices A.1 to A.3.

2.2 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 2.3, 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 M experts providing a model $\mathcal{M}_m$ allowing for projections of annual maxima for site x and area A, $\vec{y}^{(m)}_{\vec{\tau},x}$ and $\vec{y}^{(m)}_{\vec{\tau},A}$ respectively, where $m \in \{1, \ldots, M\}$, over a period $\vec{\tau}$ as described above. With a certain weight w_m attributed to each expert, the objective is then to obtain a precipitation projection with a weighted sum of the experts' projections, that is,

$$\tilde{\vec{y}}_{\vec{\tau},x} = \sum_{m=1}^{M} w_m \vec{y}^{(m)}_{\vec{\tau},x}.$$

The goal of each method is then to obtain these w_m 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 Q percentages for which each one of M experts provides corresponding quantiles $v^{(m)}_{k,q}$, $k = 1, \ldots, K$; $q = 1, \ldots, Q$; and $m = 1, \ldots, M$. In the context of extreme precipitation projection, we would have card(X) calibration variables corresponding to $\vec{Y}_{\vec{\tau},x}$ for a calibration period $\vec{\tau}$.

2.2.1 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 M over time period $\vec{\tau}$ at location x is defined as

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

with $F_{Y_{\vec{\tau},x}}$ the real cumulative distribution function and $F_{Y^{(m)}_{\vec{\tau},x}}$ the expert's CDF.

With this distance, the weight attributed to each expert's projection is then

$$w_m = \frac{1/D^{(m)}}{\sum_{l=1}^{M} 1/D^{(l)}}.$$

2.2.2 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 $v_{k,Q+1}$ such that

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

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

$\vec{z} = \{z_1, \ldots, z_{Q+1}\}$ that we can compare to the theoretical proportion $\vec{p} = \{p_1, \ldots, p_{Q+1}\}$. As shown by Cooke et al. (1991), we can obtain the calibration and entropy components, $C(m)$ and $O(m)$ respectively, as

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

where

$$I(z,p) = \sum_{q=1}^{Q+1} z_q \ln\left(\frac{z_q}{p_q}\right)$$

is the relative information component, and

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

It can readily be shown that the relative information component $I(z,p)$ 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

$$w'_m = C(m)O(m)I_{\{C(m)>\alpha\}}$$

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 $C(m)$. We can then recalibrate the weights to make their sum equal to 1 by dividing w'_m by the sum over all experts:

$$w_m = \frac{w'_m}{\sum_{l=1}^{M} w'_l}.$$

These w_m 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.

2.2.3 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 $\{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. An analyst has some prior belief about the probability that each expert's model is right, $\Pr(\mathcal{M} = \mathcal{M}_m)$, which we will denote $\Pr(\mathcal{M}_m)$, normalised such that $\sum_{m=1}^{M} \Pr(\mathcal{M}_m) = 1$. In the absence of prior information, then $\Pr(\mathcal{M}_m) = 1/M$, $\forall\, m$. Given data $\vec{y}_{\vec{\tau},x}$, the analyst can update these probabilities through Bayesian updating, that is

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

where $\Pr(\vec{y}_{\vec{\tau},x}|\mathcal{M}_m)$ is the probability of observing $\vec{y}_{\vec{\tau},x}$ under model $\mathcal{M}_m$. Since we divide by $\sum_{l=1}^{M}\Pr(\vec{y}_{\vec{\tau},x}|\mathcal{M}_l)\Pr(\mathcal{M}_l)$, it follows that $\sum_{m=1}^{M}\Pr(\mathcal{M}_m|\vec{y}_{\vec{\tau},x}) = 1$, and posterior probabilities

$\Pr(\mathcal{M}_m|\vec{y}_{\vec{\tau},x})$ 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 2.4. 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 $\vec{y}^{(m)}_{\vec{\tau},x}$, representing an expert's projection generated from model $\mathcal{M}_m$ 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 A.1. 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 2.3.

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

Algorithm 1: Generalised Likelihood Uncertainty Estimation

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

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

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

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

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

4: Update the probability of each expert as

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

5: Calculate posterior distribution as

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

2.3 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 $\vec{y}_{\vec{\psi},x}^{(m)}$ 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

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

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

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

with $y \in \vec{y}_{\vec{\psi},x}$, M 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 $1 - 1/20 = 0.95$.

2.3.1 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 ($\Pr(\mathcal{M}_m) = 1/M$), we can update our belief about the probability of each model with observed data. By defining

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

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.

2.3.2 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 2.2, 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. In words, this can be thought of as the ratio of the quantile of the area's average precipitation to the average of the individual pointwise quantiles across the area.

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 2.2, we calculate shifted densities reflecting these adjusted weights, as can be observed in Figures 2.4 and 2.5. 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 2.1 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 2.1.

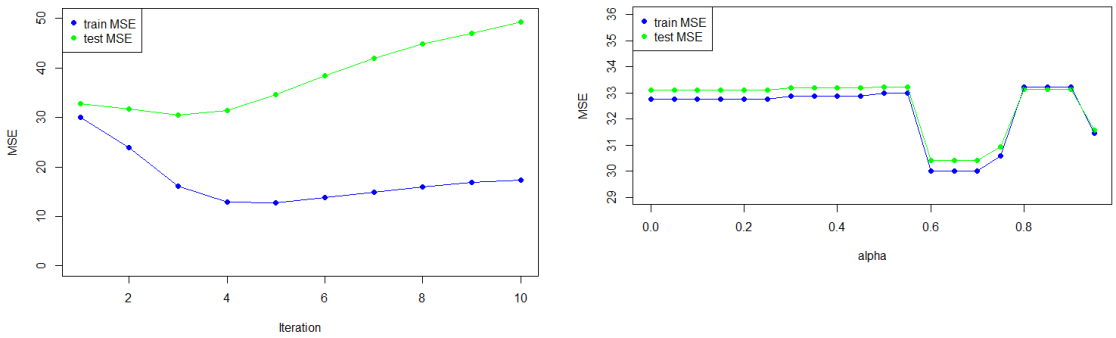

Figure 2.1: Grid cell MSE of the expectation-maximisation algorithm (left) and Cooke's method (right) in the Montreal region from 2001 to 2020

We first note that different combination methods can yield very different weights attributed to each model. Figure 2.2 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 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 2.3. 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.

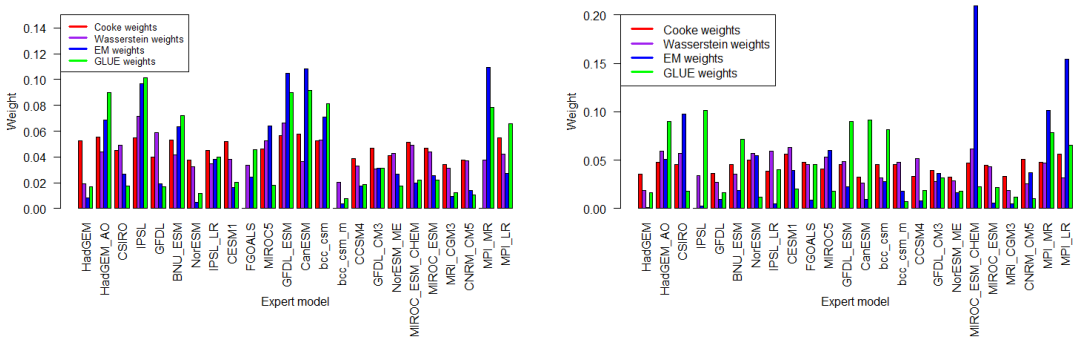

Figure 2.2: Model weight by method for Montreal (left) and Quebec (right) for precipitation from 2001 to 2020

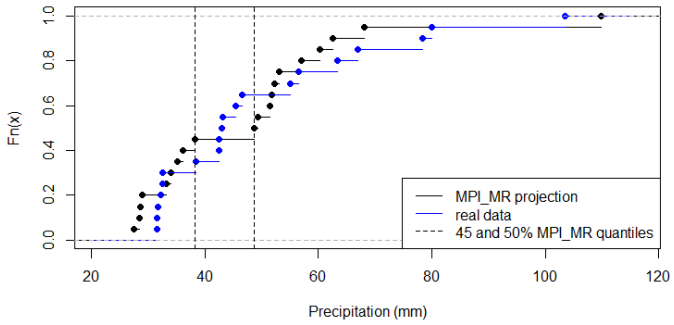

Figure 2.3: Cumulative distribution for model MPI_MR and real data in Montreal for a grid cell between 2001 and 2020

Figures 2.4 and 2.5 illustrate the upper tail of the resulting empirical cumulative distribution 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.

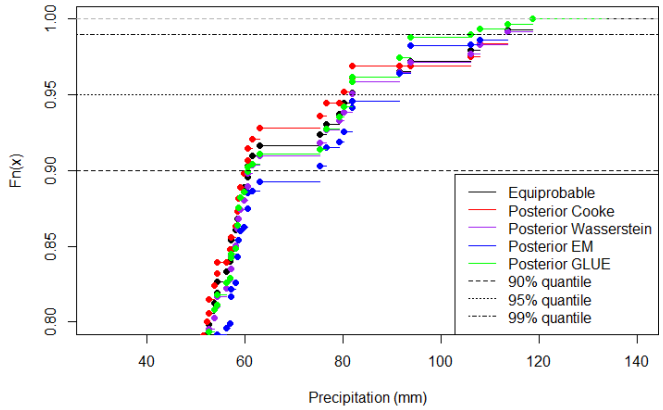

Figure 2.4: Upper tail of empirical cumulative distribution functions of pooled annual maximum daily rainfall (mm) for Montreal from 2016 to 2021 with different weighting methods

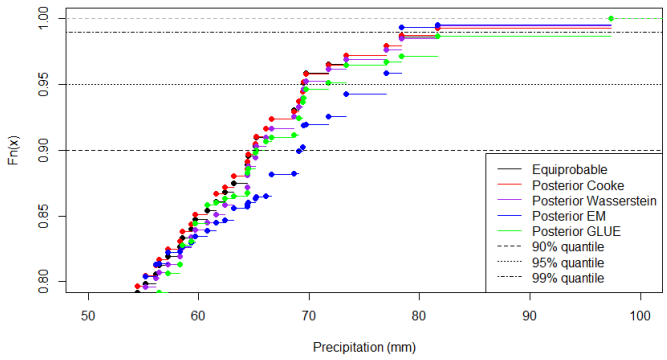

Figure 2.5: Upper tail of empirical cumulative distribution functions of pooled annual maximum daily rainfall (mm) for Quebec from 2016 to 2021 with different weighting methods

Since different combination methods yield different results, one may be interested in the variability induced by attributing weights to each expert. Let $F^{(m)}_{\vec{\tau},x}$ be the cumulative distribution function corresponding to model $\mathcal{M}_m$. We define the CDF of $Y_{\vec{\tau},A}$ as

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

where $w_1, \ldots, w_M$ are the weights attributed to each expert (which correspond to probabilities $\Pr(\mathcal{M}_m|\vec{y}_{\vec{\tau},A})$). 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 $\{Y^{(1)}_{\vec{\tau},A,R}, \ldots, Y^{(M)}_{\vec{\tau},A,R}\}$. Indeed, we have

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

Now suppose $F^{(i)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq F^{(j)}_{\vec{\tau},A}(Y_{\vec{\tau},A})$ for some $i \in \{1, \ldots, M\}$ and $\forall\, j \in \{1, \ldots, M\}$. Then it follows that $F^{(i)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq w_1 F^{(1)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) + \ldots + w_M F^{(M)}_{\vec{\tau},A}(Y_{\vec{\tau},A}) \geq 1 - 1/R$, provided that $w_1, \ldots, w_M \in [0, 1]$ 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.

From this reasoning, Figure 2.6 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.

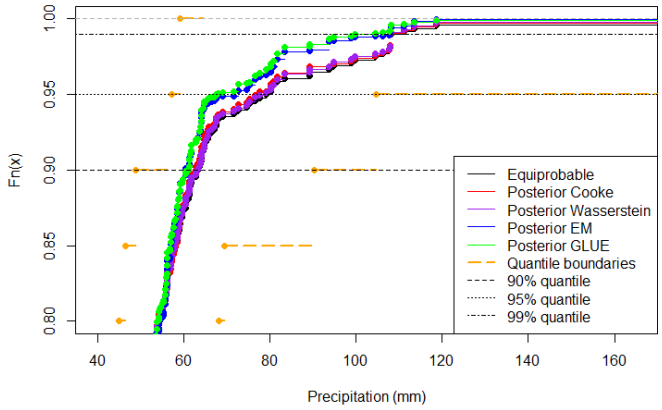

Figure 2.6: Upper tail of empirical cumulative distribution functions of pooled annual maximum daily rainfall (mm) for Montreal from 2001 to 2020 with different weighting methods, and minimum and maximum boundaries

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 combination scenario. Figure 2.7 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. 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 2.6. Given the similarity in results between the non-parametric methods using densities, and the BMA methods using residuals as in Figure 2.7, 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 2.1 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.

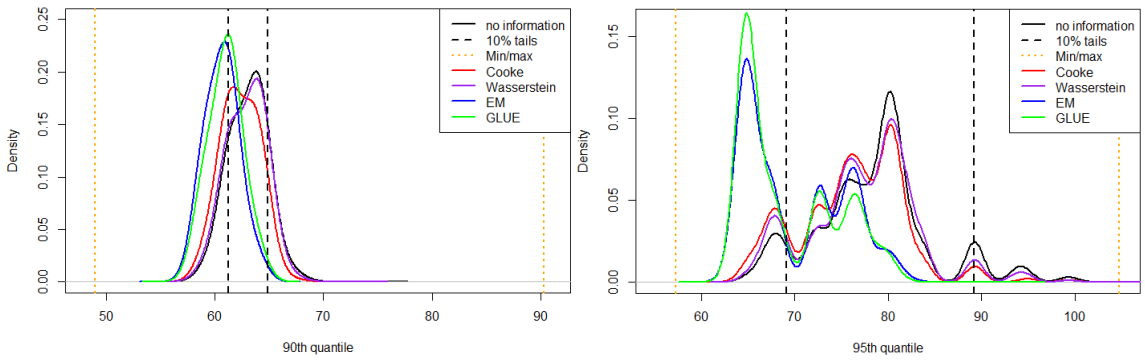

Figure 2.7: Comparison of bootstrap densities under different combination methods for the 90th quantile (left) and 95th quantile (right) in Montreal between 2001 and 2020 for 10000 iterations

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

Table 2.1: 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

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 $y^{(m)}_{\vec{\tau},x,R}$ and $y^{(m)}_{\vec{\tau},A,R}$, 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 2.8 and 2.9 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 2.9 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

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

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 2.9. 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 2.8.

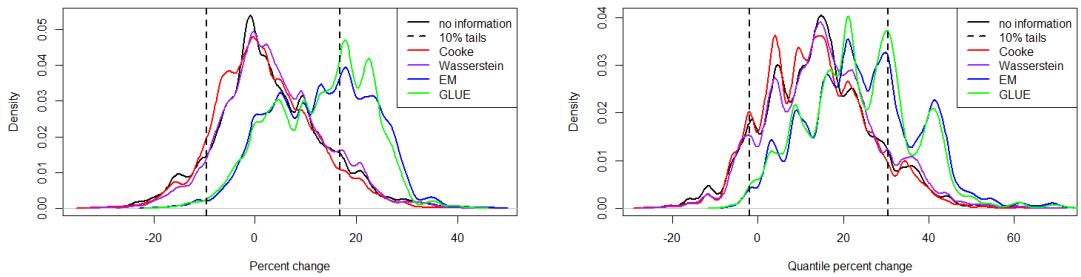

Figure 2.8: Distribution of projected quantile change at a 1 in 20 year return level in Montreal between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

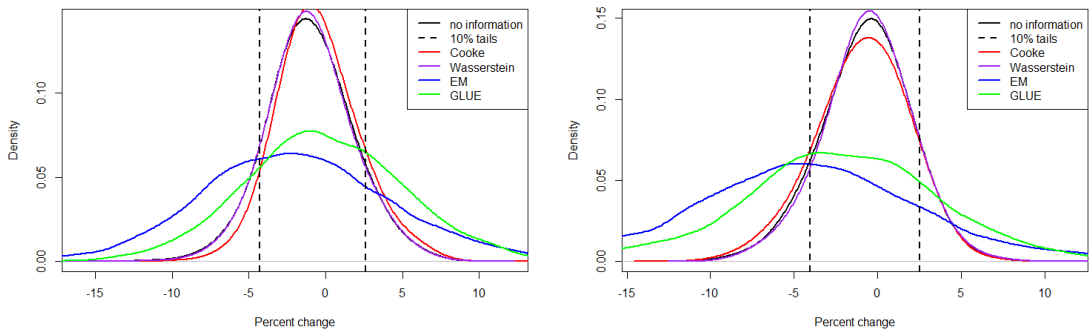

Figure 2.9: Distribution of projected ARF change at a 1 in 20 year return level in Montreal between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

The same idea is applied to Quebec in Appendix A.2, 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 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 2.10 and 2.11 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 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.

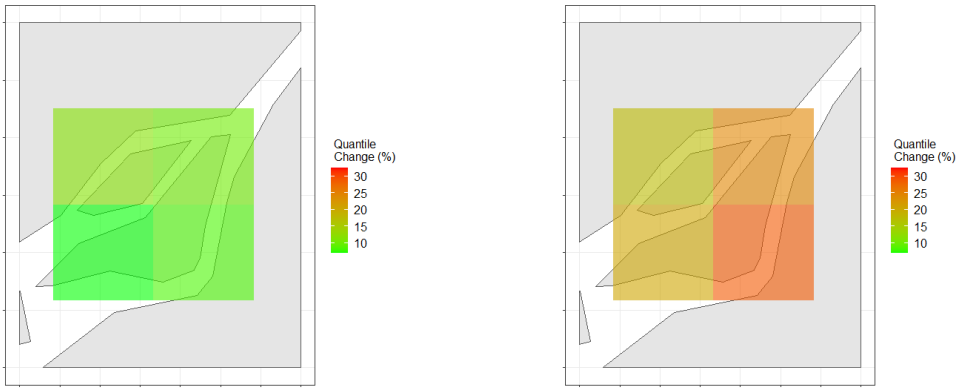

Figure 2.10: Percentage change in quantiles for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Montreal using Cooke's method (left) and BMA-EM (right)

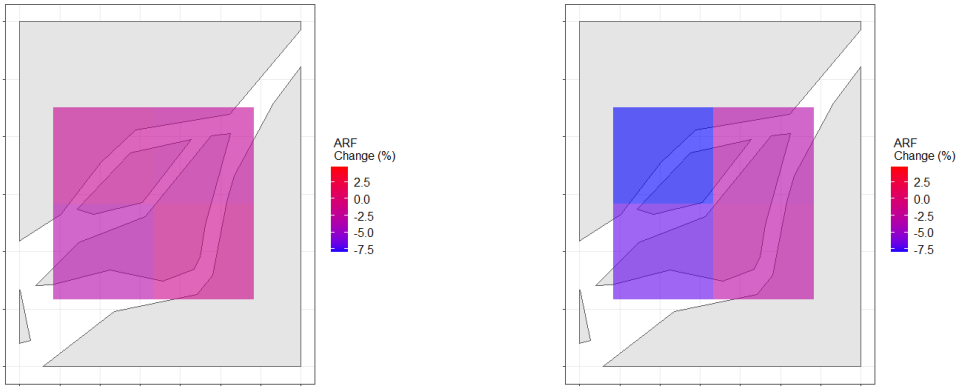

Figure 2.11: Percentage change in ARFs for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Montreal using Cooke's method (left) and BMA-EM (right)

Similar graphics are available in Appendix A.3 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. While we did not test the relative performance of different methods, given that non-parametric methods require very little input from experts, they may be better suited to low-data environments, while parametric methods should perform better with more data.

2.4 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.

Chapter 3

Uncertainty in heteroscedastic Bayesian model averaging

3.1 Introduction

Evaluating outstanding claim liabilities is of central importance to insurance companies for solvency purposes. Regulators impose constraints that insurers must respect relative to capital requirements and risk measures, as outlined in the ORSA guidelines for North America and Solvency II guidelines in Europe. These guidelines help to ensure that insureds' claims will be covered up to a high risk level. As such, much research is devoted to the development of stochastic models to evaluate claims liabilities, as outlined in Wüthrich and Merz (2008).

This research can generally be divided into two main categories: aggregate (collective) reserve models and granular (individual) reserve models. Classical collective models, such as the stochastic Chain-Ladder proposed by Mack (1993), have seen many developments, such as Generalised Linear Models (GLM, Taylor and McGuire (2016)) and Generalised Additive Model for Location, Scale and Shape (GAMLSS, Spedicato et al. (2014)). Meanwhile, individual models have grown in popularity in the last 15 years as computational power has increased. These models include semi-parametric approaches (e.g. Zhao et al. (2009), Antonio and Plat (2014)), dependence modeling with copulas (e.g. Zhao and Zhou (2010) Pešta and Okhrin (2014)), and more recently approaches based on machine learning (e.g. Duval and Pigeon (2019), Wüthrich (2018)).

While the literature devoted to claim liability estimation is very well developed, it is almost exclusively devoted to the performance of a single best model. Recently, Avanzi et al. (2024) instead proposed to use linear pooling to exploit the fact that different reserve models have different strengths (Taylor (2012), Friedland (2010)). This idea is new to the actuarial reserving literature, where combination was otherwise done based on ad-hoc rules, except for a proposal by Taylor (2012) of attributing weights to minimise the variance of total reserves. Avanzi et al. (2024) adapted a standard linear pool to combine collective reserve models, taking into consideration particularities of reserve triangles.

Their method is equivalent to Bayesian Model Averaging (BMA), which was initially proposed by Raftery et al. (1997), and has gained much popularity in recent years. Fragoso et al. (2018) reviewed the application of BMA in 587 articles published between 1994 and 2014 with applications in environmental studies, biology, social sciences, and finance. In actuarial science, Jessup et al. (2023a) compared BMA to other combination methods when projecting extreme precipitation, and a related concept of Mixture-of-Experts has been used in pricing (e.g. Bladt and Yslas (2023), Tseung et al. (2022)).

Despite BMA's recent popularity, it has a well-known issue of overfitting the observed data (Domingos, 2000) and converging to a single model. This convergence can be problematic in cases where BMA does not converge to the true model. A practical solution to this problem is to use an Expectation-Maximization (EM) algorithm (Raftery et al., 2005) and to stop after a certain optimal number of iterations to prevent convergence to a single model. The user chooses this number through a score function such as RMSE, a distance measure, or a divergence measure. Different score functions do not necessarily yield the same optimal numbers of iterations, making this choice highly subjective.

The overfitting problem may arise from the algorithm considering the observed data as the only data, leading to the neglect of random error, defined as the irreducible uncertainty in data, which is a key type of uncertainty in modelling (Abdar et al., 2021). As noted by Hüllermeier and Waegeman (2021), the distinction between aleatoric uncertainty (or random error) and epistemic uncertainty is becoming increasingly significant in a machine learning context. We propose integrating over

random error using Bayesian conditioning, which allows BMA to consider more than just the observable data. By assuming knowledge of the distribution of random error, we can condition on this error and integrate over it, resulting in a single sweep update of weights that removes random error. This method prevents the convergence to a single model when multiple models are plausible, thus preserving model diversity.

The next main issue we address is the assignment of a unique weight per model across the entire data distribution in BMA, which can be suboptimal as different models may better describe different parts of the data. We extend the idea of integrating over random error by treating the weights as Dirichlet random variables, enabling flexible weight adjustments based on predictive variables. Such flexibility is particularly valuable for actuarial reserves, where it is well known that different reserve models perform differently depending on accident years and development periods. Our error integration approach allows for weights to vary across different parts of the data.

In light of these considerations, we propose a new error integration approach that mitigates the issue of converging to a single model and introduces the concept of weights as random variables for greater flexibility. The paper is divided as follows: Section 3.2 explains Bayesian Model Averaging, Section 3.3 explains the proposed Error Integration approach and its generalisation, Section 3.4 illustrates the approaches in a simulation study, Section 3.5 applies the proposed methods to simulated loss reserving data, and Section 3.6 finally provides concluding comments.

3.2 Bayesian Model Averaging

A popular approach in statistics is selecting the model that best fits the data, then considering it as true and disregarding the uncertainty in model selection. Ensemble learning literature offers multiple approaches to model combination, allowing us to consider this model selection uncertainty. In the context of linear pooling, for M different models, we seek to attribute weights $w_m \in [0, 1]$, where $\sum w_m = 1$, to each model $\mathcal{M}_m$, such that

$$f(y^{(k)}|\mathbf{X}^{(k)}) = \sum_{m=1}^{M} w_m f_m(y^{(k)}|\mathbf{X}^{(k)}), \tag{12}$$

where f_m is the distribution under model $\mathcal{M}_m$, and the k^{th} response variable $y^{(k)}$ depends on a vector of characteristics $\mathbf{X}^{(k)}$. One simple approach is to weigh all models equally, i.e. setting $w_m = 1/M$. While this approach does not dismiss any models, it does not evaluate which models offer a better fit. Another approach we are interested in is determining weights w_m for each model to optimise a score function.

A common approach to establishing these weights is to use Bayesian Model Averaging. In the pure statistical sense of Bayesian model averaging, the weights w_m to each model are defined as

$$w_m = \Pr(\mathcal{M}_m|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\mathcal{M}_m)\Pr(\mathcal{M}_m)}{\sum_{l=1}^{M}\Pr(\mathcal{D}|\mathcal{M}_l)\Pr(\mathcal{M}_l)}, \tag{13}$$

where $\Pr(\mathcal{D}|\mathcal{M}_m)$ is the probability of observing data $\mathcal{D}$ under model $\mathcal{M}_m$. In cases where the density can be evaluated, finding this probability is straightforward. However, the density function is not necessarily always available, such as with an overdispersed Poisson distribution or with recent machine learning models. In these cases, we need an assumption to evaluate the likelihood of each model. While it would certainly be possible to use MCMC (see for example Geyer (1992)), such a method is computationally expensive, and a simpler approach is desirable.

Raftery et al. (2005) defined a dynamic expectation-maximisation (EM) algorithm supposing that residuals follow a normal distribution. This allowed for obtaining a vector of weights $\hat{\mathbf{w}}$ for all M models as

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{y^{(k)}\in\mathcal{D}} \log\left(\sum_{m=1}^{M} w_m f_m(y^{(k)}|\sigma_m^2)\right), \tag{14}$$

where each model was assumed normal with unknown homoscedastic variance σ_m^2, a common assumption used for example by Darbandsari and Coulibaly (2019) working with streamflow simulation. Most articles applying BMA are variations of the EM algorithm.

When heteroscedasticity is present, a modification to this idea is necessary to account for changing variance. One approach is to combine heteroscedastic models for which density can directly be evaluated (e.g. Avanzi et al. (2024) and Liu and Maheu (2009)). Also, using bootstrapping to calculate a variance for each quantile is conducted in (Zhu et al., 2013). Our suggested approach is to

suppose that for the random variable $Y^{(k)}$ corresponding to the k^{th} observation,

$$Y^{(k)} - E(\hat{Y}^{(k)}) \sim N(0, \sigma_k^2), \tag{15}$$

where σ_k^2 varies for each observation k. In other words, we assume that the random error around each observation is normally distributed with varying variance. This alternative approach allows us to use the benefits of the EM algorithm, while also considering heteroscedasticity when density cannot be directly evaluated. Such heteroscedastic cases can arise when considering climate-related variables such as extreme precipitation, where climate change affects the uncertainty around projections, or in an insurance context, where the variance depends on the expected loss amount.

The distributional assumption in (15) allows equation (14) to be adapted to heteroscedastic variance. Let

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{y^{(k)} \in \mathcal{D}} \log \left(\sum_{m=1}^{M} w_m \tilde{f}_m(y^{(k)}) \right) \tag{16}$$

where $\tilde{f}_m(y^{(k)}) = \phi\left(\frac{y^{(k)} - E(\hat{Y}^{(k)})}{\sigma_{k,m}}\right)$, with ϕ the density of a standard normal distribution and $\sigma_{k,m}$ the m^{th} model's standard deviation for observation k. It can readily be shown that assuming normality of residuals as in equation (16) does not affect the EM algorithm. The proof is in Appendix A.4 and follows a logic similar to Conflitti et al. (2015). Note that we approximate the residual error as a normal distribution, but make no assumption about the shape of the initial model, including model error, implying that the residual hypothesis should be applicable to most models.

3.3 Error integration

The EM algorithm has the well-known issue that it converges to a single model, as demonstrated by Le and Clarke (2022). Indeed, unless a stopping criterion is set, successive updates of the weights will lead to a single model receiving full weight. This is related to BMA supposing that the true model is among the candidate models. However, since the true model is often not one of the proposed ones, a combination can outperform any individual proposed model. To mitigate convergence to a single model, we can use cross-validation to determine the optimal number of iterations to achieve the lowest mean squared error (MSE). Determining this optimal number can

however be computationally demanding. A different approach is to exploit another source of error by incorporating data uncertainty into BMA.

When evaluating $\Pr(\mathcal{M}_m|\mathcal{D})$ in the usual BMA algorithm, we suppose that the dataset $\mathcal{D}$ is fully representative of its underlying model. This can lead to cases where, given two equally valid models, one would receive a weight of 0 while the other would receive a weight of 1, depending on the data. For example, consider a case where data comes from a normal distribution with a certain mean μ and variance σ^2. Two normal models with the same variance are compared, one model with mean $\mu - 1$, and the other with mean $\mu + 1$. Suppose the observed data has mean $\bar{X} = \mu - 0.1$. Then, the first model will have higher likelihood, such that iterative weight updates will converge to it, despite both models being equally distant from the true distribution. Data has inherent uncertainty, which needs to be considered to avoid alternating between a model receiving full weight or no weight.

To address this issue, we know that real data follows some unknown distribution $\mathbb{P}$ such that $Y^{(k)} \sim \mathbb{P}(\mathbf{X}^{(k)}; \mathbf{\Theta})$, depending on covariates $\mathbf{X}^{(k)}$ with parameters $\mathbf{\Theta}$. The observed data $\mathcal{D}$ represents one iteration from $\mathbb{P}$, from which we fit models with known distribution $\mathbb{Q}_m$ and estimated parameters such that $\hat{Y}_m^{(k)} \sim \mathbb{Q}_m(\mathbf{X}^{(k)}; \hat{\mathbf{\Theta}}_m)$.

3.3.1 Symmetric uncertainty

Suppose the uncertainty around $Y^{(k)}$ can be approximated reasonably well by a normal distribution, such that

$$y^{(k)} = E(Y^{(k)}) + \epsilon_k, \tag{17}$$

where $\epsilon_k \sim N(0, \sigma_k^2)$. Then,

$$\begin{aligned} y^{(k)} - E(\hat{Y}_m^{(k)}) &= y^{(k)} - E(Y^{(k)}) + E(Y^{(k)}) - E(\hat{Y}_m^{(k)}) \\ &= \epsilon_k + E(Y^{(k)}) - E(\hat{Y}_m^{(k)}), \end{aligned}$$

where $E(\hat{Y}_m^{(k)})$ should converge to its true value when observing more data, s.t.

$$y^{(k)} - E(\hat{Y}_m^{(k)}) \sim N(E(Y^{(k)}) - E(\hat{Y}_m^{(k)}), \sigma_k^2).$$

If we further suppose that

$$E(\hat{Y}_m^{(k)}) = E(Y^{(k)}), \tag{18}$$

that is, that models are unbiased, then this hypothesis is equivalent to equation (15). While supposing equality of means is a strong hypothesis, for a model for which this does not hold, we have $\Pr(\mathcal{M}_m|\mathcal{D}) \to 0$. A misspecified model tending to 0 is due to the hypothesis of the distribution of residuals. If residuals are not centered at 0, then the model likelihood will certainly be lower than that of unbiased models. The mismatch in variance will also lead to a lower likelihood. Using Bayes' theorem will thus result in a small weight for the misspecified model.

We want to take data uncertainty, which can be thought of as random error, into account when combining models. Let $\boldsymbol{\epsilon}$ be a vector of random variables representing this uncertainty, such that $y^{(k)} = E(Y^{(k)}) + \epsilon_k$. Consider its distribution $\pi(\boldsymbol{\epsilon}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements σ_k^2 (assuming each observation's random error is considered independent). We then have

$$\begin{aligned}\Pr(\mathcal{M}_m|\mathcal{D}) &= \int \Pr(\mathcal{M}_m|\mathcal{D}, \boldsymbol{\epsilon})\pi(\boldsymbol{\epsilon}|\mathcal{D})d\boldsymbol{\epsilon} \\ &= \int \cdots \int \frac{\Pr(\mathcal{D}|\mathcal{M}_m, \boldsymbol{\epsilon}) \Pr(\mathcal{M}_m)}{\sum_{l=1}^{M} \Pr(\mathcal{D}|\mathcal{M}_l, \boldsymbol{\epsilon}) \Pr(\mathcal{M}_l)} \pi(\epsilon^{(1)}|\mathcal{D}) \cdots \pi(\epsilon^{(K)}|\mathcal{D}) d\epsilon^{(1)} \cdots d\epsilon^{(K)},\end{aligned}$$

which, supposing no prior information about each model, such that $\Pr(\mathcal{M}_m = 1/M)$, reduces to

$$= \int \cdots \int \frac{\Pr(\mathcal{D}|\mathcal{M}_m, \boldsymbol{\epsilon})}{\sum_{l=1}^{M} \Pr(\mathcal{D}|\mathcal{M}_l, \boldsymbol{\epsilon})} \pi(\epsilon^{(1)}|\mathcal{D}) \cdots \pi(\epsilon^{(K)}|\mathcal{D}) d\epsilon^{(1)} \cdots d\epsilon^{(K)}. \tag{19}$$

This formula cannot be evaluated analytically, but it can be approximated by simulating S draws of random errors to each observation such that $\hat{y}_{m,s}^{(k)} = E(\hat{Y}_m^{(k)}) + \epsilon_s^{(k)}$ and approximating the

resulting weight as

$$\Pr(\mathcal{M}_m|\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\frac{1}{|\mathcal{D}|}\sum_{y^{(k)}\in\mathcal{D}}\frac{\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_m)}{\sum_{l=1}^{M}\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_l)}, \tag{20}$$

where $\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_m) = \phi((y^{(k)} - \hat{y}_{m,s}^{(k)})/\sigma_{k,m})$. Equation (20) however implies that variance of random errors must be known, which is generally not the case in practice.

To solve this issue, we use an approach initially proposed by Harvey (1976). Using the equality in means in equation (18), we can measure σ_k^2 by further supposing that $\sigma_k^2 = g(\mathbf{X}^{(k)}\boldsymbol{\theta}_m)$ for a link function g which is convex and differentiable on its domain, and a vector of parameters $\boldsymbol{\theta}_m$, such that

$$g^{-1}(\boldsymbol{\sigma}^2) = \mathbf{X}\boldsymbol{\theta}_m + \boldsymbol{\upsilon} \tag{21}$$

for an error term $\boldsymbol{\upsilon}$ and matrix $\mathbf{X}$ grouping vectors of covariates $\mathbf{X}^{(k)}$ for all K observations. Taking a standard loss function such as quadratic loss, we then need to solve

$$\hat{\boldsymbol{\theta}}_m = \arg\min_{\boldsymbol{\theta}_m}\sum_{k=1}^{K}\left(g^{-1}((y^{(k)} - E(\hat{Y}_m^{(k)}))^2) - \mathbf{X}^{(k)}\boldsymbol{\theta}_m\right)^2, \tag{22}$$

which has the well-known solution

$$\hat{\boldsymbol{\theta}}_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'g^{-1}(\boldsymbol{\sigma}^2). \tag{23}$$

Although there are other possible choices, we use the quadratic loss function due to its popularity in fitting procedures and its properties, especially in our linear framework, e.g. Judge and Mittelhammer (2004).

Take the link function g to be the exponential function. This ensures positivity of the result, which is desirable for a variance function. Then, $g^{-1}(\boldsymbol{\sigma}^2) = \mathbf{R}_m$, where $R_{m,k} = \ln((y^{(k)} - E(\hat{Y}_m^{(k)}))^2)$.

Notice that this estimator is different from each model's error. Call this modelled variance $\tilde{\sigma}_{m,k}^2 = e^{\mathbf{X}^{(k)}\hat{\boldsymbol{\theta}}_m}$ to distinguish it from $\sigma_{m,k}^2$. We can then simulate using the modelled variance $\tilde{\sigma}_{m,k}^2$ and calculate resulting weights with normal densities using each model's error $\sigma_{k,m}^2$, which

can be evaluated with a model-specific function $h_m(\mathbf{X}^{(k)})$. Then, averaging according to equation (20), we obtain weights numerically integrated over random error.

3.3.2 Asymmetric uncertainty

The approach discussed so far supposes that error is symmetric around $E(Y^{(k)})$, which is not always true. In cases where skewness is observed, we need to simulate uncertainty from a skewed distribution. Depending on the value γ of skewness, different options are possible. If $|\gamma| < 1$, an alternative to the normal distribution would be to consider the skew-normal distribution, which has density

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right), \tag{24}$$

with ϕ and Φ, respectively the standard normal density and cumulative normal density, and $\xi \in \mathbb{R}$, $\omega > 0$ and $\alpha \in \mathbb{R}$ the location, scale and shape parameters. It is easy to verify that when $\alpha = 0$, this is equal to the normal distribution. If $|\gamma| \geq 1$, a better alternative would be to consider extreme value theory with a Generalized Extreme Value (GEV) distribution (see for example Hosking et al. (1985)).

Suppose γ_k depends on the k^{th} observation. Then, similarly to variance, we wish to model skewness based on covariates, such that $\gamma_k = g(\mathbf{X}^{(k)}\boldsymbol{\zeta}_k)$ for a vector of parameters $\boldsymbol{\zeta}_k$ and a link function g. Just like equation (22), we have

$$\hat{\boldsymbol{\zeta}}_m = \arg\min_{\boldsymbol{\zeta}_m} \sum_{k=1}^{K}\left(g^{-1}\left(\frac{(y^{(k)} - E(\hat{Y}_m^{(k)}))^3}{\tilde{\sigma}^3_{m,k}}\right) - \mathbf{X}^{(k)}\boldsymbol{\zeta}_m\right)^2.$$

It would be desirable to take g^{-1} as $\ln$ to bring all values to a similar scale, but $(y^{(k)} - E(\hat{Y}_m^{(k)}))^3$ can take negative values. To deal with this issue, we can separate the positive and negative terms, then obtain a prediction for both. Consider $K^+ = \{k : y^{(k)} - E(\hat{Y}_m^{(k)}) > 0\}$, and $K^- = \{k : y^{(k)} - E(\hat{Y}_m^{(k)}) < 0\}$. Then,

$$\hat{\boldsymbol{\zeta}}^+_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^+_m \text{ and } \hat{\boldsymbol{\zeta}}^-_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^-_m, \tag{25}$$

where $S^+_{m,k} = \ln((y^{(k)} - E(\hat{Y}^{(k)}))^3/\tilde{\sigma}^3_{m,k})$ and $S^-_{m,k} = \ln(-(y^{(k)} - E(\hat{Y}^{(k)}))^3/\tilde{\sigma}^3_{m,k})$ for $k \in K^+$

and K^- respectively. With this equation, we finally have

$$\tilde{\gamma}_k = e^{\mathbf{X}^{(k)}\zeta_k^+} - e^{\mathbf{X}^{(k)}\zeta_k^-}, \tag{26}$$

where we subtract the negative skewness that we artificially made positive. Note that taking a log function can potentially introduce bias through Jensen's inequality, but this step is necessary to obtain stable results.

Algorithm 2 explains this procedure, where we have S simulations of a skewed normal, with $\alpha = 0$, or equivalently $\gamma = 0$, the special case of the normal distribution. This algorithm is appropriate if $-1 < \gamma_k < 1$, and would otherwise need to be adapted using for example a GEV distribution. Note that our algorithm updates the weights only once, thus guaranteeing weights to be positive, and arbitrarily close to 0 for a model that is poorly adjusted to the data.

In the case where a GEV distribution is more appropriate, MLE parameters are not as readily available as with a skew-normal, but can be obtained through iterative algorithms, or the parameters can be estimated using the method of moments. The remainder of the algorithm then follows the same steps, where step 4 of Algorithm 2 is a GEV simulation instead of a skew-normal simulation.

Algorithm 2: BMA with Numerical Error Integration

1: Set initial weights, variance and skewness as

$$w_m^{(0)} = 1/M \ \forall m,$$
$$\tilde{\sigma}_{k,m}^2 = e^{\mathbf{X}\hat{\boldsymbol{\theta}}_m},$$
$$\tilde{\gamma}_{k,m} = e^{\mathbf{X}\hat{\boldsymbol{\zeta}}_m^+} - e^{\mathbf{X}\hat{\boldsymbol{\zeta}}_m^-},$$
$$\sigma_{k,m}^2 = h_m(\mathbf{X}^{(k)}).$$

2: Calculate the skew-normal parameters as

$$\delta_{k,m} = \sqrt{\frac{\pi|\gamma_{k,m}^{1.5}|}{2(|\gamma_{k,m}|^{1.5} + ((4-\pi)/2)^{2/3})}}$$
$$\alpha_{k,m} = \frac{\delta_{k,m}}{\sqrt{1-\delta_{k,m}^2}}$$
$$\omega_{k,m} = \sqrt{\frac{\pi \cdot \tilde{\sigma}_{m,k}^2}{\pi - 2\alpha_{k,m}^2/(1+\alpha_{k,m}^2)}}$$
$$\xi_{k,m} = E(\hat{Y}^{(k)}) - \omega_{k,m}\sqrt{\frac{2\alpha_{k,m}}{\pi(1+\alpha_{k,m}^2)}}$$

3: **for** s **in** $1:S$, **do**

4: Simulate error-adjusted $\hat{y}_{m,s}^{(k)}$ for each model m and claim k as

$$\hat{y}_{m,s}^{(k)} = SN(\xi_{k,m}, \omega_{k,m}, \alpha_{k,m})$$

5: Obtain proportion from normal densities for each expert m and claim k as

$$z_{s,m,k} = \frac{w_m^{(0)}\phi\left(\frac{y^{(k)}-\hat{y}_{m,s}^{(k)}}{\sigma_{k,m}}\right)}{\sum_{l=1}^{M} w_l^{(0)}\phi\left(\frac{y^{(k)}-\hat{y}_{m,s}^{(k)}}{\sigma_{k,l}}\right)}$$

6: Update the probability associated to each model as

$$\Pr(\mathcal{M} = \mathcal{M}_m|\mathcal{D}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{S}\sum_{s=1}^{S} z_{s,m,k}.$$

7: Calculate posterior distribution as

$$\Pr(y|\mathcal{D}) = \sum_{m=1}^{M}\Pr(y|\mathcal{M}_m)\Pr(\mathcal{M} = \mathcal{M}_m|\mathcal{D}).$$

3.3.3 Desirable model properties for optimal performance

By comparing the Kullback-Leibler (KL) divergence of the BMA algorithm with the divergence of the error integration (EI) algorithm, we can establish desirable properties for models that we wish to combine for EI to perform optimally. The KL divergence between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = \int f_{\mathbb{P}}(x) \log\left(\frac{f_{\mathbb{P}}(x)}{f_{\mathbb{Q}}(x)}\right) dx. \tag{27}$$

Consider the case where $\mathbb{Q} = \mathbb{P}$. Then, it directly follows that

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = D_{KL}(\mathbb{P}||\mathbb{P}) = 0,$$

such that if the true model $\mathcal{M}_{m^*}$ is in the set $\mathcal{H} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ and BMA converges to this model, then $D_{KL}(\mathbb{P}||\mathbb{Q}^{(\text{BMA})}) = 0$. For EI however, from equation (20), model weight can tend towards 0 for bad models, but is strictly greater than 0. We then have

$$\begin{aligned} D_{KL}(\mathbb{P}||\mathbb{Q}^{(\text{EI})}) &= \int f_{\mathbb{P}}(x) \log\left(\frac{f_{\mathbb{P}}(x)}{\sum_{m=1}^{M} \int \Pr(\mathcal{M}_m|\mathcal{D},\epsilon) \Pr(\epsilon|\mathcal{D}) d\epsilon f_{\mathbb{Q}_{\geq}}(x)}\right) dx \\ &> \int f_{\mathbb{P}}(x) \log\left(\frac{f_{\mathbb{P}}(x)}{f_{\mathbb{Q}_{m^*}}(x)}\right) dx \\ &= 0, \end{aligned}$$

such that if BMA converges to the true model, then BMA will outperform EI. Note, however, that a model can have higher likelihood than the true model for the observed data, such that BMA could converge to the wrong model.

Next, consider the case where the true model is not in $\mathcal{H}$. While there is no closed form allowing us to compare KL divergence in terms of general distributions, a closed form exists for normal distributions (Hershey and Olsen, 2007). By taking normally distributed approximations $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$ to $\mathbb{P}$ and $\mathbb{Q}$ respectively, we know that

$$D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}) = \frac{1}{2}\left(\log\frac{|\Sigma_{\hat{\mathbb{Q}}}|}{|\Sigma_{\hat{\mathbb{P}}}|} - n + Tr(\Sigma_{\hat{\mathbb{Q}}}^{-1}\Sigma_{\hat{\mathbb{P}}}) + (\boldsymbol{\mu}_{\hat{\mathbb{P}}} - \boldsymbol{\mu}_{\hat{\mathbb{Q}}})^T \Sigma_{\hat{\mathbb{Q}}}^{-1}(\boldsymbol{\mu}_{\hat{\mathbb{P}}} - \boldsymbol{\mu}_{\hat{\mathbb{Q}}})\right). \tag{28}$$

Proposition 1.

$$\textit{If} \quad \sum_{k=1}^{n} \left| \sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2 - \sigma_k^2 \right| \leq \sum_{k=1}^{n} \left| \sigma_{m^*,k}^2 - \sigma_k^2 \right| \quad \textit{and} \quad \sum_{k=1}^{n} \log \left(\frac{\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2}{\sigma_{m^*,k}^2} \right) \geq 0,$$

then $D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}^{(EI)}) \leq D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}^{(BMA)})$.

Proof. Suppose BMA converges to a single model s.t. $\mathbb{Q}^{(\text{BMA})} = \mathbb{Q}_{m^*}$ for the model $\mathcal{M}_{m^*}$ with highest likelihood, while EI yields

$$\mathbb{Q}^{(\text{EI})} = \sum_{m=1}^{M} w_m \mathbb{Q}_m,$$

where $w_m = \int \Pr(\mathcal{M}_m|\mathcal{D}, \epsilon)\pi(\epsilon|\mathcal{D})d\epsilon$. Further suppose that $D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}^{(\text{EI})}) \leq D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}^{(\text{BMA})})$. Then,

$$\sum_{k=1}^{n} \left[\log \left(\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2 \right) + \frac{\sigma_k^2}{\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2} + \frac{(\mu_k - \sum_{m=1}^{M} w_m \mu_{m,k})^2}{\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2} \right] \leq \sum_{k=1}^{n} \left[\log(\sigma_{m^*,k}^2) + \frac{\sigma_k^2}{\sigma_{m^*,k}^2} + \frac{(\mu_k - \mu_{m^*,k})^2}{\sigma_{m^*,k}^2} \right].$$

Under the same assumption as equation (18), the last terms on either side of the equality fall to 0, and we can rearrange terms to obtain

$$\sum_{k=1}^{n} \log \left(\frac{\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2}{\sigma_{m^*,k}^2} \right) \leq \sum_{k=1}^{n} \frac{\sigma_k^2 (\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2 - \sigma_{m^*,k}^2)}{(\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2)\sigma_{m^*,k}^2}.$$

With this simplified inequality, the result follows by induction, as shown in Appendix A.5. □

In a case where some models have $\sigma_{m,k}^2 > \sigma_k^2$ and others have $\sigma_{m,k}^2 < \sigma_k^2$, then there exists a combination of weights $\tilde{w}_m$ s.t. $\sum_{m=1}^{M} \tilde{w}_m^2 \sigma_{m,k}^2 = \sigma_k^2$, while $\sigma_{m^*,k}^2 \neq \sigma_k^2$. Even though $\tilde{w}_m \neq w_m$, it follows that under the unbiased hypothesis, we can reasonably expect the EI algorithm to outperform BMA, provided that some models are overdispersed and others are underdispersed.

We can further develop this intuition by looking at the divergence measure proposed by Bhattacharyya (1946), defined as

$$D_B(\mathbb{P}, \mathbb{Q}) = -\ln\left(\int \sqrt{p(x)q(x)}dx\right). \tag{29}$$

Again under the assumption of equation (18) and supposing normal distributions,

$$D_B(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) = \frac{1}{2}\ln\left(\frac{\sigma^2_{\hat{\mathbb{P}}} + \sigma^2_{\hat{\mathbb{Q}}}}{2\sigma_{\hat{\mathbb{P}}}\sigma_{\hat{\mathbb{Q}}}}\right). \tag{30}$$

Then, solving a similar inequality as with Kullback-Leibler, and knowing that the $\ln$ terms are non-negative, we obtain the following conditions for the inequality to hold.

$$\text{If } \sum_{k=1}^{K}\sigma^2_{m^*,k} > \sum_{k=1}^{K}\sum_{m=1}^{M} w_m^2\sigma^2_{m,k}, \text{ then } \sum_{k=1}^{K}\sigma_k^2 \leq \sum_{k=1}^{K}\sigma^2_{m^*,k}.$$
$$\text{If } \sum_{k=1}^{K}\sigma^2_{m^*,k} < \sum_{k=1}^{K}\sum_{m=1}^{M} w_m^2\sigma^2_{m,k}, \text{ then } \sum_{k=1}^{K}\sigma_k^2 \geq \sum_{k=1}^{K}\sigma^2_{m^*,k}.$$

These conditions tell us that EI will perform at least as well as BMA if we have a mix of overdispersed and underdispersed models, where the first condition corresponds to the best model being overdispersed, while the second corresponds to the model being underdispersed.

As such, this discussion allows us to conclude that for the error integration algorithm to work optimally and potentially outperform BMA, one needs a mix of models with some overdispersed models and some underdispersed models. This conclusion will be further illustrated in the simulation study in Section 3.4.

3.3.4 Generalised error integration

The combination methods considered so far assume a single weight to be applied to each model, as per equation (12). While this is certainly useful, cases can arise where one model can fit a part of the data better, whereas another model might be best for another part of the data. This idea has been explored by Kapetanios et al. (2015), who proposed a generalised density combination where

weights depend on the variable of interest, such that

$$f(y) = \sum_{m=1}^{M} w_m(y) f_m(y). \tag{31}$$

Due to the nearly infinite possibilities for functions $w_m(y)$, the authors focus on piecewise linear weights, which is similar to Mixture-of-Experts, an idea used in pricing consisting of selecting one model for each part of the data using logit gating functions (e.g. Tseung et al. (2020)).

We propose to consider w_m not as a function of a random variable, but as a random variable itself. Suppose $w_m^{(k)}$ varies for the k^{th} observation, such that

$$f(y^{(k)}) = \sum_{m=1}^{M} w_m^{(k)} f_m(y^{(k)}), \tag{32}$$

where $w_m^{(k)} \in [0,1]$ and $\sum_m w_m^{(k)} = 1 \; \forall k$, such that $f(y^{(k)})$ is a proper probability distribution integrating to 1. We can then think of $\{W_1, \ldots, W_M\}$ as a random M-tuple from a Dirichlet distribution (see for example Frigyik et al. (2010)), defined as

$$g(w_1, \ldots, w_M | \iota_1, \ldots, \iota_M) = \frac{\Gamma\left(\sum_{m=1}^{M} \iota_m\right)}{\prod_{m=1}^{M} \Gamma\left(\iota_m\right)} \prod_{m=1}^{M} w_m^{\iota_m - 1}, \tag{33}$$

for $\iota_m > 0$, where $0 \leq w_m \leq 1 \; \forall m$ and $\sum_{m=1}^{M} w_m = 1$. Using this distribution, we then have

$$E(W_m) = \frac{\iota_m}{\sum_{m=1}^{M} \iota_m}. \tag{34}$$

We can then establish a loglink for the k^{th} observation, similarly to GLMs, such that

$$\log(\iota_m^{(k)}) = \boldsymbol{\beta}_m \mathbf{X}^{(k)}, \tag{35}$$

allowing us to use maximum likelihood estimation to determine $\hat{\boldsymbol{\beta}}_m$, and so obtain

$$\hat{\iota}_m^{(k)} = \exp(\hat{\boldsymbol{\beta}}_m \mathbf{X}^{(k)}). \tag{36}$$

Notice that in the case of non-informative characteristics $\mathbf{X}^{(k)}$, we recover constant weights, and we can recover convergence to a single model by letting all $\iota_m \to 0$ except for the best model with $\iota_{m^*} > 0$. In such a way, with this more general approach, we can still obtain the results presented in the previous section.

Including a Dirichlet regression modifies steps 6 and 7 of Algorithm 2. Instead of averaging over the K observations, we fit a Dirichlet regression (see package VGAM (Yee, 2010)), then use the predicted weights to obtain the density generated by Equation (32). Note that the choice of a linear structure is not the only possible approach, and was chosen mainly for its strong interpretability.

3.4 Simulation Study

3.4.1 Single weight per model

To illustrate the single weight methods proposed in Section 3.3, consider the following simulation study. Suppose we have data such that $Y \sim LN(\mu = 0.0009968x, \sigma = 0.008\sqrt{x})$. This data has a mean, or signal, of $\exp(0.01x)$, is heteroscedastic, where the variance depends on x, and has positively skewed uncertainty around the signal. Suppose a Poisson Generalised Linear Model (GLM) and an inverse Gaussian (IG) GLM are fitted to this data, as well as a misspecified Poisson model set as half of the Poisson GLM. Figure 3.1 illustrates the data, as well as a random draw from the fitted models. Here, in order to match mean and variance, thus providing underdispersed models, the Poisson GLMs are fitted to continuous data despite being discrete. For large amounts, this discretisation is frequent in practice. Note that the equal mean and variance of the Poisson GLM provides a classical benchmark for the chain-ladder reserving method, such that it is a standard actuarial choice.

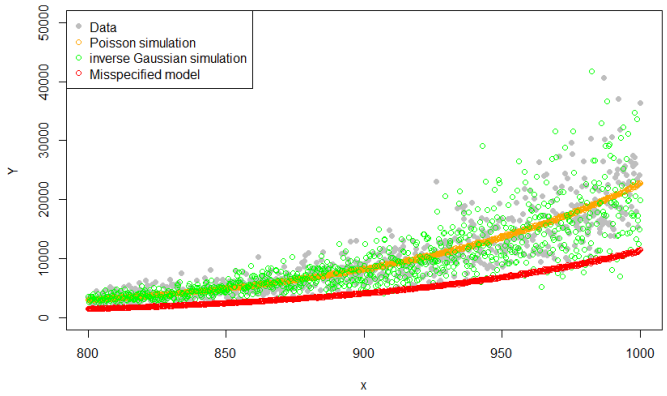

Figure 3.1: Simulation study random draw

We see that the first two models accurately predict the signal of $\exp(0.01x)$, as seen by the fitted parameters of these models, namely $\hat{\boldsymbol{\lambda}}$ for the Poisson, as well as $\hat{\boldsymbol{\nu}}$ and $\hat{\delta}$ for the inverse Gaussian in Table 3.1, but have very different uncertainties around the prediction, while the misspecified Poisson model is a poor fit. The Poisson model is clearly largely underdispersed, while the inverse Gaussian model is slightly overdispersed. If we are interested in accurately predicting quantiles, we need to match the data's uncertainty.

Table 3.1: Poisson and inverse Gaussian fitted parameters

	Intercept	x
Poisson	$\hat{\lambda}_0 = -0.10446$	$\hat{\lambda}_1 = 0.010136$
IG	$\hat{\nu}_0 = 0.175029$	$\hat{\nu}_1 = 0.009826$
	$\hat{\delta} = 8.35078 * 10^{-6}$	NA

With $\hat{\theta}$ and $\tilde{\gamma}$ obtained respectively with equations (23) and (26), we simulate 10,000 draws as $E(\hat{Y}^{(k)}) + SN(0, \tilde{\sigma}_k, \tilde{\gamma}_k)$, where

$$E(\hat{Y}^{(k)}_{\text{Pois}}) = \exp\left(\hat{\lambda}_0 + \hat{\lambda}_1 * x\right)$$

and

$$E(\hat{Y}_{\text{IG}}^{(k)}) = \exp\left(\hat{\nu}_0 + \hat{\nu}_1 * x\right),$$

then calculate normal densities using the Poisson and inverse Gaussian variances, evaluated as

$$Var(\hat{Y}_{\text{Pois}}^{(k)}) = \exp\left(\hat{\lambda}_0 + \hat{\lambda}_1 * x\right)$$

and

$$Var(\hat{Y}_{\text{IG}}^{(k)}) = \hat{\delta} * \left(\exp\left(\hat{\nu}_0 + \hat{\nu}_1 * x\right)\right)^3.$$

Averaging over the $10,000$ draws, we obtain a weight of 88.5% for the inverse Gaussian model and 8.3% for the Poisson model, while the wrong model receives 3.2%. The algorithm successfully avoids convergence to a single model and recognises the bad model. To compare with the BMA expectation-maximisation algorithm, we need to determine the optimal number of iterations to avoid convergence to a single model, which is the inverse Gaussian model in this case. To this end, we use the Diebold-Mariano (DM) test, which allows for comparing distributions and obtaining significance levels (Diebold and Mariano, 2002). Looking at Figure 3.2, we see that the lowest DM statistic is achieved after a single iteration, where the 0 iteration corresponds to equal weights to each model.

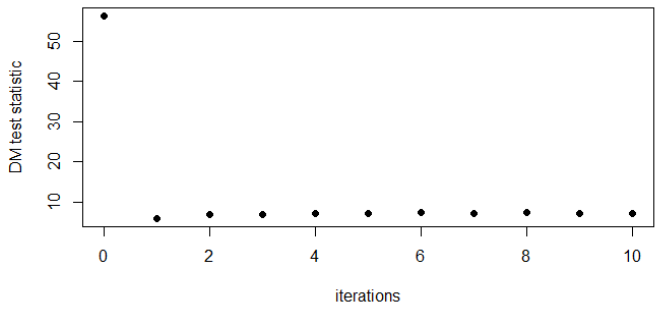

Figure 3.2: Diebold-Mariano test statistic
by number of iterations of the EM algorithm

Figure 3.3 shows the real data and random draws from the weighted distributions obtained through EI as well as BMA stopped after one iteration. We see that both methods visually closely match the data's uncertainty, which is further supported by the zoomed in densities for specific values of x, where both methods have similar predicted densities. Looking at Table 3.2, this matching of uncertainty is confirmed, where both methods outperform BMA without a stopping point, have similar Kullback-Leibler (KL) divergence, and error integration significantly outperforms the other two methods in terms of DM test statistics, achieving a p-value of 0.2. Note that the requirement of mixing over and under-dispersed models is respected here, where if the inverse Gaussian model had not been overdispersed, then EI could not have outperformed BMA.

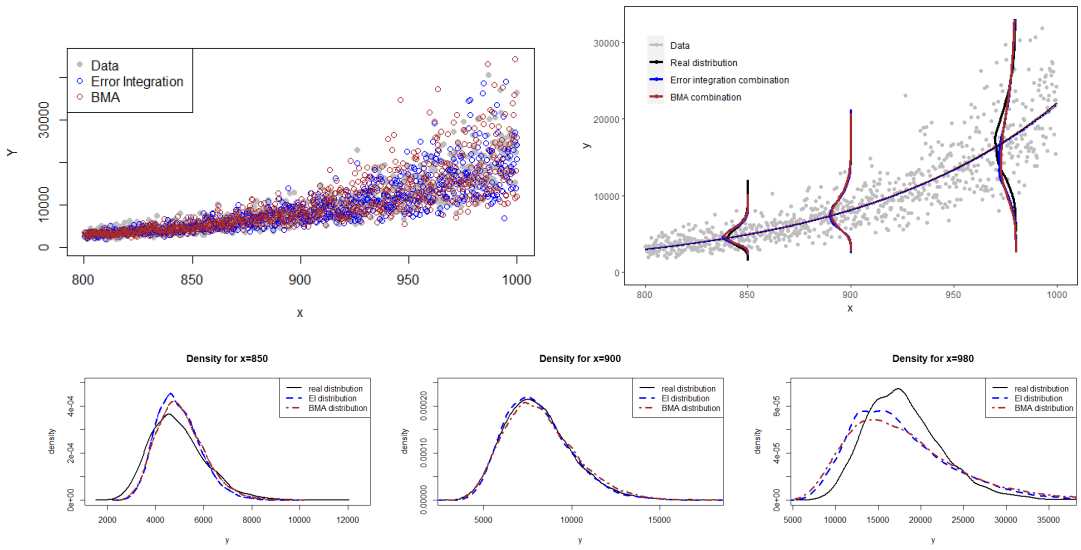

Figure 3.3: Simulation study combination random draw (left) and densities (right), and zoomed in densities for $x = 850, 900, 980$

Table 3.2: Weighted average Diebold-Mariano test statistic and Kullback-Leibler divergence between combined distributions and real distribution

Method	Diebold-Mariano		Kullback-Leibler
	Statistic	p-value	Divergence
EI	2.35	0.20	0.070
BMA - until convergence	7.49	$2.06 * 10^{-6}$	0.081
BMA - optimal iterations	6.86	$1.85 * 10^{-5}$	0.072

Given the similar results between the two methods, the main advantage of error integration compared to BMA is that it removes the need for determining the optimal number of iterations, which can be computationally intensive, and subjective.

3.4.2 Generalised weights

To illustrate the generalisation proposed in subsection 3.3.4, consider a slightly different case, where data comes from a mixture of a Poisson with mean $\exp(0.01x)$ and a lognormal with $\mu = 0.00995x$ and $\sigma = 0.01\sqrt{x}$, with weights linearly changing from 100% for the Poisson to 100% for

the lognormal. Once again, Poisson and inverse Gaussian GLM are fitted to this data, as well as a misspecified Poisson model. Instead of averaging the 10,000 draws to obtain a single weight, we average the draws for each x, then use a Dirichlet regression to obtain a prediction for the weights that depend on x. Figure 3.4 illustrates the underlying data and a random draw from the fitted distributions, while Figure 3.5 illustrates the weights obtained, where we see that the algorithm recognises that the data initially follows a Poisson, while the inverse Gaussian is a better fit for larger values of x. Note that due to the nature of the Dirichlet distribution, no link function can allow for predicted weights to change linearly. While this is suboptimal in this specific example, for larger datasets with categorical explanatory variables, this limitation is not expected to cause issues.

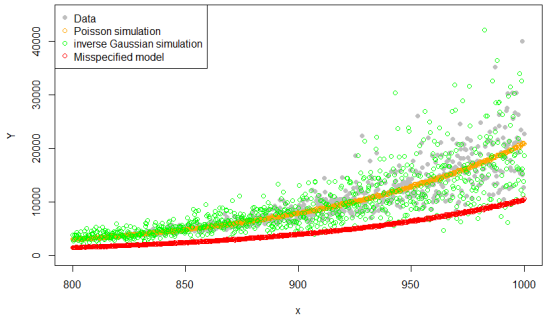

Figure 3.4: Second simulation study random draw

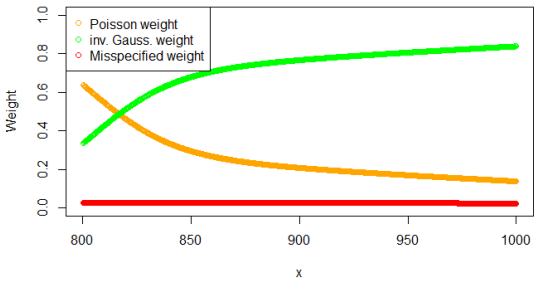

Figure 3.5: Dirichlet regression weights for each model

Comparing the usual BMA approach to the generalised EI approach with flexible weights, in this case the generalised approach outperforms the BMA approach. This can be seen in both Figure 3.6 and Table 3.3. Indeed, the density of the generalised combination is closer to the real distribution except for high values of x, where BMA and the generalised approach return similar outputs. Moreover, the generalised approach has lower DM test statistic and KL divergence than BMA.

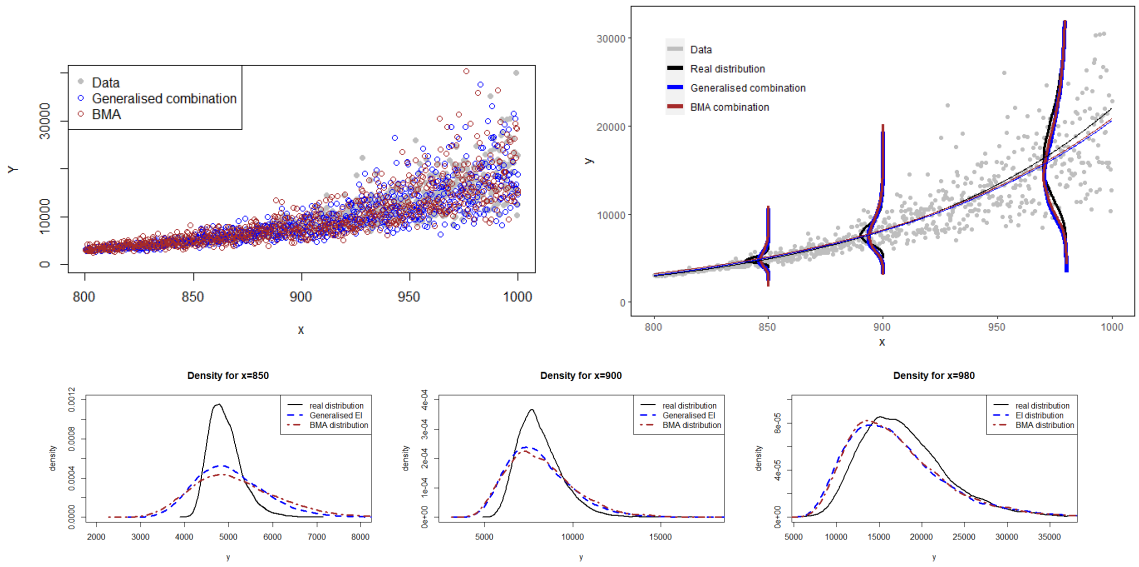

Figure 3.6: Second simulation study combination random draw (left) and combination densities (right), and zoomed in densities for $x = 850, 900, 980$

Table 3.3: Weighted average Diebold-Mariano test statistics and Kullback-Leibler divergence between combined distributions and real distribution

Method	Diebold-Mariano		Kullback-Leibler
	Statistic	p-value	Divergence
Single weight EI	0.67	0.03	0.170
Generalised EI	0.21	0.03	0.158
BMA - optimal iterations	6.09	0.06	0.205

3.5 Analysis

In this section, we present a case study based on a simulated Property and Casualty insurance dataset. While the example is actuarial in nature, the proposed algorithms are applicable to any model combination when faced with heteroscedastic data.

3.5.1 Data

We work with the Simulation Machine created by Gabrielli and Wüthrich (2018) to obtain a dataset $\mathcal{D}$ with a single line of business with approximately 10,000 claims for which we know the claim code, accident year/quarter, the age of the insured, the injured body part, the amounts paid each year, and the ultimate amount paid. We add random calendar year inflation following a normal distribution with mean 2% and standard deviation 0.5%. We also create a strong predictive variable for large losses by setting this variable equal to 1 if the ultimate claim is above 1000, and 0 otherwise. In such a way, this variable has a correlation of approximately 0.3 with ultimate losses.

$\mathcal{D}$ is separated into a training set $\mathcal{D}_{\text{train}}$ and a prediction set $\mathcal{D}_{\text{pred}}$ by splitting payment information before and after a certain valuation date. The prediction set represents outstanding claims that actuaries must evaluate, also known as the actuarial reserve. This separation ensures that all available information is considered, which is necessary in practice, and enables comparisons with the results obtained by Avanzi et al. (2024), who also use model combination to project reserves.

This approach can lead to leakage of information, such that a claim can have payments in the training and prediction datasets. To avoid such a situation, we can exclude a subset of claims from the training set, and use this excluded subset for testing, as in Gabrielli et al. (2020). In our case, this methodology provides comparable reserve sufficiency projections compared to using the full dataset.

The training set is further separated to obtain a calibration set $\mathcal{D}_{\text{calib}}$ by randomly selecting 30% of the payment information from the most recent calendar year. This calibration set is necessary in order to determine the weights attributed to each model. The choice of the most recent calendar year for calibration is justified by its practical applicability. Additionally, the calibration set needs to be further separated for the EM algorithms to determine how many iterations are optimal.

Table 3.4 presents a short development triangle, grouping data between accident year and development year. Paid claims represent the training data. These claims are known on the valuation date. This dataset is further separated by randomly sampling 30% of the claims in the latest calendar year to calibrate the ensembling weights. Outstanding claims occur after the valuation date and form the

prediction period. Note that when using aggregate models, the latest accident year and development period cannot be used for calibration, as the information in the claims triangle is necessary for training models.

Table 3.4: Loss development triangle

	Development period					
AY	**1**	**2**	**3**	**4**	**5**	**6**
1	P	P	P	P	P	O
2	P	P	P	P	O	O
3	P	P	P	O	O	O
4	P	P	O	O	O	O
5	P	O	O	O	O	O

Note: paid claims (P), outstanding claims (O), calibration period (P)

3.5.2 Underlying models

As explained by Fragoso et al. (2018), the model space for any model selection problem can be vast, and fitting all possible models is not realistic. In the context of actuarial reserving, we face a similar issue, where given the recent literature on granular reserving, many models are available for modelling purposes. Avanzi et al. (2024) describe three criteria to select models in an aggregate combination scenario: models that can be fitted automatically, models with different strengths and limitations, and models that are easily identifiable and interpretable.

An efficiency argument justifies the first argument; if component models are hard to adjust, the combination algorithm would be inefficient and have little practical interest.

The second criterion allows all potential data patterns to be covered. Moreover, it is known that independent models yield optimal combination results (Jacobs, 1995). Models relying on different hypotheses should have lower correlation in their predictions and provide more information than dependent models. From our previous discussion, we also wish to have models with varying levels of dispersion.

The third criterion limits the use of machine learning models, which makes sense in an aggregate context due to the low availability of data. In a granular data setting, sufficient data is available for

these methods, but the first criterion implies that we need machine learning methods that are efficient computationally and do not require user input.

In their article, the authors use these criteria to select multiple generalised linear models with different effects, smoothing splines (Green and Silverman, 1993), and generalised additive models for location, scale, and shape (GAMLSS, Rigby and Stasinopoulos (2005)).

With the same criteria in mind, we choose Gamma and overdispersed Poisson (ODP) generalised linear models (GLM, see De Jong et al. (2008) for more details on GLMs), as well as an overdispersed Poisson double GLM (see Smyth and Jørgensen (2002)). We also consider an aggregate overdispersed Poisson GLM (see Wüthrich and Merz (2008)) and an aggregate generalised additive model for location, scale, and shape (GAMLSS, see Rigby and Stasinopoulos (2005)) applied to individual data. The models are fitted to payments by accident year and development year, with the individual models also using claim-specific covariates.

While recent actuarial literature on reserve models includes interesting developments in machine learning methods, such methods require intensive adjustments and are not necessarily transparent. We have therefore chosen not to consider them for combination purposes for now.

3.5.3 Model combination

With a single simulated database, Figure 3.7 shows the reserve distributions obtained for the five models that are combined, and the vertical black line represents the total amount of outstanding claims of the simulated loss dataset. In the absence of a strong predictive variable, models give similar outputs, slightly underestimating the real amount due to calendar year inflation, except for the Gamma GLM. With a strong predictor, the granular GLMs slightly overestimate the real amount, while the DGLM performs better than without this predictive variable.

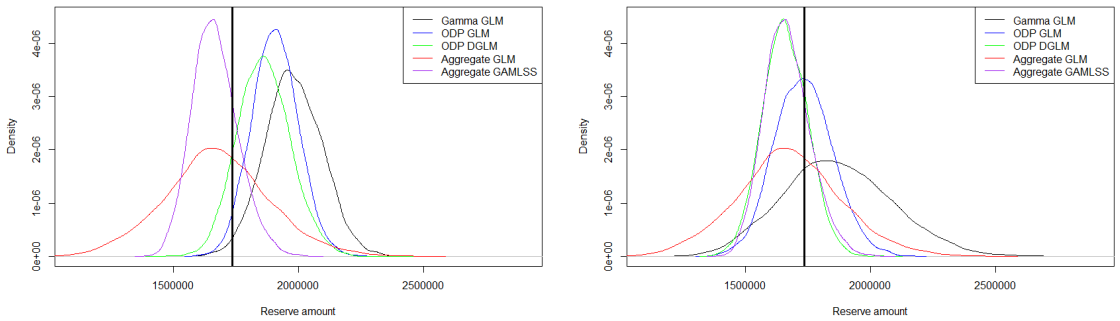

Figure 3.7: Underlying reserve model distributions with (left) and without (right) strong predictor

Tables 3.5 and 3.6 list the resulting weights from the classical BMA-EM algorithm, the heteroscedastic version, and the EI procedure, where the first table is obtained without a strong predictive variable, while the second table is obtained with this variable.

Table 3.5: Weights obtained from each method without strong predictor

Model	Weights (%) from BMA based on		
	Homoscedastic	Heteroscedastic	Error integration
Gamma GLM	28.9	0.1	17.9
ODP GLM	18.0	0.3	18.5
ODP DGLM	20.6	99.3	27.4
Aggregate GLM	15.6	0.3	18.0
Aggregate GAMLSS	16.9	0.0	18.2

Table 3.6: Weights obtained from each method with strong predictor

Model	Weights (%) from BMA based on		
	Homoscedastic	Heteroscedastic	Error integration
Gamma GLM	21.1	9.0	25.0
ODP GLM	40.4	3.0	17.4
ODP DGLM	29.3	73.5	30.0
Aggregate GLM	5.0	14.0	13.5
Aggregate GAMLSS	4.2	0.1	14.1

We note that the homoscedastic algorithm favours the granular GLM models, the heteroscedastic algorithm puts almost all weight on the DGLM, while the EI procedure favours the DGLM, but gives weight to all models. The homoscedastic algorithm avoids convergence to a single model because of the variance assumption, leading to large variance even for low losses, such that all models seem to perform similarly. The heteroscedastic algorithm corrects the issue of misstating the variance of individual losses, but faces the problem of rapid convergence to a single model. We can note here that the algorithm is not converging to the model most centered around the true reserve amount, highlighting how convergence to a single model can be problematic when the algorithm does not converge to the true model. This further accentuates the usefulness of our proposed method, since EI successfully avoids convergence to a single model, and spreading weights across multiple models yields a better output than any individual model. We can see that in the presence of a strong predictor, the individual models receive more weight than without this predictor under homoscedastic BMA and EI, which makes sense since the individual models use this variable while the aggregate models do not.

Using the generalised EI algorithm, we can analyse the impact of characteristics on the weights attributed to each model. We find that the fitted log-coefficients ($\hat{\boldsymbol{\beta}}_m$) become increasingly negative for more recent accident years, suggesting greater uncertainty about which models are best suited for those years. Additionally, there is some fluctuation in the log-coefficients; however, this variation has only a minor effect on the weights themselves. This indicates a notable level of uncertainty

regarding the optimal models. Despite this, a strong predictive variable leads to aggregate models receiving less weight for more recent accident years compared to granular models. This result is intriguing, as it aligns with the intuition that older years with more data are better suited to aggregate models, while granular reserve models can better predict more recent accident years. The fitted log-coefficients are available in Appendix A.6.

Note that the data generator that we use in R shows that the idea of Dirichlet regression works well even with a simple case. Indeed, with the Simulation Machine proposed by Gabrielli and Wüthrich (2018), the transition is smoother than the ADLP method proposed in Avanzi et al. (2024). This method splits the reserve triangle in two (or more), which, when used on a small triangle, leaves little information for calibration, causing weights to converge to a single model for the upper portion of the triangle. This creates a large jump in weights from one year to the next, which does not happen with our method.

With a more complex data generator such as SynthETIC (Avanzi et al., 2021), we expect the idea to yield even better results, which we leave as an area for future research. However, if little data is available, we believe a Dirichlet regression could lead to overfitting, which could cause jumps in weights. This could be mitigated by penalising the regression to ensure smoothness in weights between different parts of the data.

Figure 3.8 illustrates the distributions obtained using the proposed combination methods, as well as using the standard linear pool aggregate method. As expected from the similarity between models in the absence of a strong predictive variable, the combination methods yield similar results in this case, with the EI algorithm performing slightly better than its counterparts. With a strong predictive variable, where models were more different, generalised EI yields clearly better results, where it is centered around the true outstanding loss amount.

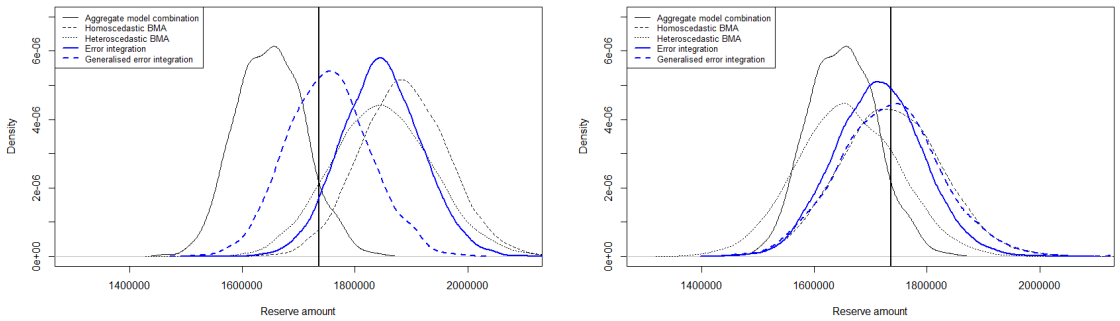

Figure 3.8: Result of BMA combination using different approaches with (left) and without (right) strong predictor

While EI outperforms standard linear pooling of aggregate models in this particular example, we did not compare it to the accident and development year adjusted pooling proposed by Avanzi et al. (2024), other than observing that splitting the triangle causes convergence to a single model for older accident years. Moreover, in a limited data context where aggregate methods are usually preferred to individual models, it is expected that aggregate model combination would perform better than individual model combination. Performing both combinations would allow for avoiding overconfidence in a combination method.

3.5.4 Performance and runtime

Given that the database is simulated, we can investigate how often a reserve set at the 99^{th} quantile is sufficient by simulating 200 different databases and calculating reserves for each database. We then look at whether the real amount from the simulated database is lower than the 99^{th} quantile to evaluate reserve sufficiency, which should happen 99% of the time. Simulating multiple databases allows for mitigating the potential bias from a single simulated database as in Section 3.5.3. Table 3.7 illustrates the percentage of sufficiency, the mean reserve, and the runtime for each reserve simulation. We see that in the absence of a strong predictive variable, generalised weights return similar results to a single weight, only slightly outperforming the latter. This makes sense, as both approaches should be equivalent with non-informative predictive variables. In the presence of a strong predictor, generalised weights outperform the single weight approach. In terms of computation time, the generalised weight approach is however significantly more expensive, where the

computation time more than doubles using variable weights compared to single weights.

Note that the generated databases' losses were limited at 50,000 to be closer to reality. Given the simplicity of the chosen models despite the large losses, this explains the slight underperformance of model combinations in terms of sufficiency. We see that in the presence of a strong predictive variable, generalised weights allow for a less conservative approach than the classical bootstrap Chain-Ladder method, which overestimates the necessary reserves.

	Method	Sufficiency (%)	Mean reserve (M)	Runtime
Without predictor	Single weight	91.5	1.418	4.5min
	Generalised weight	94.5	1.421	10.5min
With predictor	Single weight	94.5	1.447	4.5min
	Generalised weight	99	1.608	10.5min
	Bootstrap CL	100	1.746	0.5min

Table 3.7: Proportion of sufficient reserves and runtime with & without predictive variable at a 99th level quantile

3.6 Conclusion

In this paper, we proposed two model combination algorithms based on Bayesian Model Averaging taking heteroscedasticity into account in an actuarial reserving context. More specifically, we adjusted the classical Expectation-Maximisation algorithm to account for heteroscedasticity, we proposed a numerical error integration approach to take data uncertainty into account, and we proposed a generalisation to this approach allowing for flexible weights through a Dirichlet regression.

Through a simulation study, we showed that the proposed error integration algorithm successfully identifies better models while avoiding convergence to a single model, and performs at least as well as the BMA algorithm adapted to heteroscedastic data. We demonstrated that to perform optimally, model combination using error integration requires a mix of overdispersed and underdispersed models.

We applied the proposed algorithms to a simulated dataset using the simulation machine created by Gabrielli and Wüthrich (2018). We found that in the presence of a strong predictive variable, the proposed error integration approach outperformed other approaches. Without this strong predictor, all methods performed similarly, suggesting that error integration will perform at least as well as its more classical BMA counterpart. Generalised EI performed much better with a strong predictive variable, but was also found to be computationally much more expensive than single weight approaches.

It would be interesting to relax the assumption of independence in random error, where dependence could be induced by an unobservable variable confounding causal links (Liu et al., 2023). This could allow for better representation of the distribution over which we integrate. It would further be interesting to allow for more complex models with better predictive variables, which would be expected to improve the performance of model combination.

Chapter 4

Flexible extreme thresholds through generalised Bayesian Model Averaging

4.1 Introduction

Insurers are often exposed to large losses resulting from diverse sources such as injuries following an accident, crashes of high value vehicles, or catastrophic events. In 2023, 23 of these events in Canada each resulted in over $30 million in damages, contributing to nearly a quarter of the $3.1 billion in total insured losses (CatIQ, 2024). As such, insurers need models capable of accounting for these large losses as well as for the bulk of losses composed of smaller claims in order to account for all possible events.

4.1.1 Extreme value theory

One approach to dealing with large losses is to use Extreme Value Theory (EVT). Initially developed by Fréchet (1927) and Fisher and Tippett (1928), then later by Gumbel (1958), this theory deals with the tail of a distribution. There are two broad categories to extreme value analysis: block maxima and Peak-over-Threshold. The first approach studies the largest observations from successive blocks of independent and identically distributed data. This is well explained in Coles et al. (2001) and has many applications in insurance and finance, such as flood risk modelling (Boudreault et al., 2020), catastrophe risk (Embrechts et al., 2013), climatic extremes (Cheng et al.,

2014), and risk management (McNeil et al., 2015).

The second approach examines values that exceed a specified level, known as the Peak-over-Threshold method. For a sufficiently high limit, the excess values can be demonstrated to follow a generalised Pareto distribution (GPD), developed by Pickands (1975), with cumulative density function

$$G(x) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi} & \text{for } \xi \neq 0 \text{ and } x > u \\ 1 - \exp\left(-\frac{x-u}{\sigma}\right) & \text{for } \xi = 0 \text{ and } x > u, \end{cases} \tag{37}$$

with location, scale, and shape parameters u, σ, and ξ. This approach can be used in a similar variety of contexts as block maxima, such as extreme daily precipitation (Thiombiano et al., 2017), operational risk (Chavez-Demoulin et al., 2006), stock returns (He et al., 2022), as well as catastrophe risk (Li et al., 2016), and has the main advantage over block maxima that it allows for more data. The main challenge with this method is identifying a suitable threshold beyond which losses follow a GPD. Threshold selection, however, is an open problem with no universally accepted method. Caeiro and Gomes (2015) explain a few of the available methods based on a heuristic choice or minimization of mean squared error.

Recently, in a homogeneous setting, automatic threshold selection methods have been proposed such as using L-moments (Silva Lomba and Fraga Alves, 2020), parameter stability (Curceac et al., 2020), goodness-of-fit (Bader et al., 2018), and other methods partly reviewed by Benito et al. (2023), who find that different thresholds can yield similar market risk measures. These methods generally require establishing potential threshold values through a range of quantiles, from which a "best" threshold is chosen.

This range of values can be fully automated, such as choosing a standard set of quantiles, or it can be chosen graphically using extreme value theory. It is well known that for a sufficiently large threshold u_0, the mean excess loss is a linear function of $u > u_0$, such that

$$E(X - u | X > u) = \frac{\sigma_{u_0} + \xi(u - u_0)}{1 - \xi} + \frac{\xi}{1 - \xi} u. \tag{38}$$

We can then look at a mean residual life plot and select a point where the plot becomes linear, choosing as a function of the bias-variance trade-off, where a higher threshold reduces bias, but

increases variance. Many other methods are available, and can be implemented using the "tea" package in R (Ossberger, 2020).

4.1.2 Mixture model combination

While there is a significant body of literature devoted to identifying a single best threshold, a question central to this article is whether it could be better to use model combination to simultaneously consider multiple potential threshold values. Such a combination would be less affected by threshold misspecification than selecting a single best threshold. In fact, Northrop et al. (2017) recently used Bayesian Model Averaging (BMA), a method initially proposed by Raftery et al. (1997) that gained significant popularity across many scientific fields (Fragoso et al., 2018), to reduce sensitivity to threshold selection when studying the number of exceedances over a high threshold in ocean storms.

Oppositely to Northrop et al. (2017) who focus on the tail of the distribution, in an actuarial context, we are often interested in all the potential values that data can take rather than only the exceedances above a threshold. In such cases, each model must account for the full range of values that the data can take. Therefore, we require a model that can efficiently represent both the bulk and the tail of the data. Laudagé et al. (2019) proposed a mixture of a generalised linear model (GLM) and a GPD for a known threshold, which was expanded on by Ghaddab et al. (2023) in an excess-of-loss reinsurance context, where the threshold was estimated using MRL plots and Hill plot estimators. Li and Liu (2023) proposed a three-part mixture model dividing low, medium, and high claims according to a 20%/60%/20% rule of thumb. These proposed approaches are subject to threshold misspecification.

To reduce the issue of threshold selection, we will focus on a mixture model proposed by MacDonald et al. (2011). The authors suggested using a non-parametric distribution with kernels for the bulk of the data and a GPD for the tail beyond a threshold u. Replacing the kernel distribution by a parametric distribution, this has density

$$f(y|\mathbf{\Lambda}, u, \sigma_u, \xi) = \begin{cases} (1-\phi_u) \times \frac{h(y|\mathbf{\Lambda})}{H(u|\mathbf{\Lambda})} & y \leq u \\ \phi_u \times g(y|u, \sigma, \xi) & y > u, \end{cases} \tag{39}$$

where $h(y|\mathbf{\Lambda})$ is the bulk density with parameters $\mathbf{\Lambda}$, while $g(y|u,\sigma,\xi)$ is a GPD density with parameters σ and ξ. Note that ϕ_u is simply the proportion of data above u and is not a parameter as such. This can cause a point discontinuity at u, which could be reduced by optimising both the bulk and tail parameters simultaneously using MCMC (MacDonald et al., 2011). Another less computationally demanding method is to adjust ϕ_u to ensure continuity (Pigeon and Denuit, 2011). Such an approach does however remove the intuition of ϕ_u being a proportion.

By considering mixture models following equation (39), we can apply combination methods such as BMA to allocate weights to each model. This enables the use of a range of thresholds, thus reducing the impact of threshold misspecification.

4.1.3 Tail-Weighted GLUE for Threshold Selection in BMA

We propose using a tail-weighted version of Generalised Likelihood Uncertainty Estimation (GLUE, see Beven and Freer (2001)) within BMA to identify the "best" threshold, which outperforms single threshold selection methods. We will show that by placing more weight on tail likelihood compared to the bulk of the data, we can identify this threshold using extreme value theory principles. This idea will be compared to the results obtained through the ForwardStop methodology proposed by Bader et al. (2018). Furthermore, our proposed approach offers an improvement over the BMA method proposed by Northrop et al. (2017) since it allows for studying the full distribution instead of focusing on the tail.

Moving beyond the homogeneous setting, Coles et al. (2001) argued that the threshold can depend on several covariates. Recent approaches to flexible threshold selection using covariates often rely on quantile regression. For example, Youngman (2019) used quantile regression to calculate thresholds at 0.90 to 0.99 levels and selected the threshold with the lowest RMSE. Similarly, Fu and Sayed (2023) selected thresholds based on parameter stability.

Our proposed approach relies on an idea similar to weighted GLUE which allows us to identify thresholds based on predictive variables without using quantile regression. First, the mixture model proposed by MacDonald et al. (2011) is generalised by considering generalised additive model (GAM) versions of the mixed models (Hastie, 2017). We then combine these models to account for predictive variables using a modified version of the BMA method proposed by Jessup et al. (2023b),

which models residual uncertainty and integrates out random error. This approach can be applied to a mixture model, modifying the algorithm to account for changing uncertainty between the bulk of data and the extreme value tail. Heavier tail weights help obtain thresholds varying with predictive variables.

This method is preferable to quantile regression approaches for identifying flexible thresholds, such as Youngman (2019), because it provides immediate results over the entire distribution. While homogeneous threshold selection methods like Bader et al. (2018) run very fast, they require considering each predictive variable value separately, given that the Anderson-Darling test is not well-defined for categorical variables.

We thus propose a method for identifying the "best" threshold through model combination in a homogeneous setting and demonstrate that a combination of mixture models performs better than a single model with the right threshold while reducing dependence on choosing this threshold. We extend this idea to Bayesian model averaging of GAM models in the presence of predictive variables to obtain flexible thresholds based on risk characteristics. Our method proves preferable to other threshold selection methods, both with and without predictive variables, in terms of high quantiles while providing distributions for the full data.

The paper is divided as follows: Section 4.2 establishes the main theoretical results and applies them to the Danish dataset, Section 4.3 generalises the results in a regression setting and applies them to an actuarial dataset, and Section 4.4 concludes the article.

4.2 Homogeneous setting

4.2.1 Theory

Due to the frequent scenario of limited data for analysis, we often need to study only the variable of interest, without any explanatory variables. In such cases, we usually consider data to be independent and identically distributed. We can attempt to model this data by proposing M different models, then using model combination to accurately model the data by combining these models.

In particular, Bayesian Model Averaging (BMA) is a popular model combination method used in many fields of science (Fragoso et al., 2018) relying on Bayesian updating in a context of linear

pooling. The broad category of linear combination attributes weights $w_m \in [0,1]$ to M different models, where $\sum w_m = 1$, to each model $\mathcal{M}_m$, such that

$$f(y) = \sum_{m=1}^{M} w_m f_m(y), \tag{40}$$

where f_m is the distribution under model $\mathcal{M}_m$. While there are many ways of establishing these weights, BMA sets the weights w_m as the probability that each model is the true model given the observed data, or

$$w_m = \Pr(\mathcal{M}_m|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\mathcal{M}_m)\Pr(\mathcal{M}_m)}{\sum_{l=1}^{M}\Pr(\mathcal{D}|\mathcal{M}_l)\Pr(\mathcal{M}_l)}, \tag{41}$$

where $\Pr(\mathcal{D}|\mathcal{M}_m)$ is the likelihood of data $\mathcal{D}$ under model $\mathcal{M}_m$.

We know that the data comes from an unknown distribution $\mathbb{P}$, such that the observed data $\mathcal{D}$ is in fact not the only possible observable data. Generalised Likelihood Uncertainty Estimation (GLUE, see Beven and Freer (2001)) can be used within BMA to take this uncertainty into account. This method uses the asymptotic normality of quantile estimates (Van der Vaart, 2000) to evaluate the likelihood of each model, where the variance is quantile-dependent to take into account quantile heteroscedasticity (e.g. Jessup et al. (2023a), Zhu et al. (2013)). A limitation of this approach is that it assumes the sample size is sufficiently large for asymptotic properties to hold. This assumption is often invalid, especially when dealing with large values, as data for such cases is typically limited. Smaller sample size can cause bias and skewness, such that a distribution taking skewness into account may be more appropriate. We propose to adjust the GLUE algorithm to consider skewness as well as variance in a skew-normal distribution. Details concerning the skew-normal are available in Appendix A.7, and the GLUE algorithm is described in Algorithm 3.

Note that the normal distribution is a special case of the skew-normal distribution (with skewness 0), and so, with sufficient data, using a skew-normal version of GLUE will asymptotically converge to the standard GLUE methodology. Proof of equivalence when there is no skewness is shown in Appendix A.7.

In the standard GLUE algorithm, we put equal weight $1/Q$ to all quantiles when calculating

data likelihood under each model, such that

$$L(\mathcal{D}|\mathcal{M}_m) = \frac{1}{Q}\sum_{q=1}^{Q} L(\hat{y}_m^{(q)}), \tag{42}$$

where $\hat{y}_m^{(q)}$ is the predicted q^{th} quantile of model $\mathcal{M}_m$. In the context of extreme values, although these values account for only a fraction of the events, they represent the scenarios where important damage or losses can happen. As such, we want to ensure that this portion of the distribution is well-modeled. By taking a weighted mean such that

$$L^*(\mathcal{D}|\mathbf{M}_m) = \sum_{q=1}^{Q} \frac{\hat{y}_m^{(q)}}{\sum_{j=1}^{Q} \hat{y}_m^{(j)}} L(\hat{y}_m^{(j)}), \tag{43}$$

the weight assigned to the m^{th} model will depend more on the tail of the distribution. Note that the choice of weights is motivated by the intuition of losses receiving a weight corresponding to their percentage of total loss, and that other reweighting methods are possible. As such, we propose to use weighted-GLUE by replacing equation (42) with equation (43). Comparing the results of these two equations in the context of mixture models allows for Proposition 2.

Proposition 2. *Let M different models $\mathcal{M}_m$, $m \in \{1, \ldots, M\}$, be mixture models as defined by equation (39) $\forall$ m, with each model having a different predetermined threshold u_m s.t. the models cover a wide range of possible thresholds. Let w_m and w_m^* be the weights to the m^{th} model under respectively the mean and weighted mean GLUE. Further let u be the best threshold. Then $w_m^* \geq w_m$ if $u_m \geq u$, and $w_m^* \leq w_m$ if $u_m \leq u$.*

Proof. From the Fisher-Tippett-Gnedenko and Pickands-Balkema-De Haan theorems (Coles et al., 2001), if a variable has a distribution function belonging to a maximum domain of attraction, then its maximum distribution belongs to a GEV and there is an equivalence between the GEV and GPD parameters. This implies that beyond a sufficiently high threshold, all threshold choices yield valid GPD distributions.

Consider a combination with only two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, where $u_1 < u$ and $u_2 \geq u$. Given that both models are mixture models defined by Equation (39), it follows that $\mathcal{M}_2$ must be better

adjusted to the tail beyond u_2 than $\mathcal{M}_1$, seeing as the Pickands-Balkema-De Haan theorem implies that $\mathcal{M}_2$ will have lower bias for the tail beyond u_2 than $\mathcal{M}_1$.

This further suggests that for quantiles beyond u_2, $|y^{(q)} - \hat{y}_1^{(q)}| > |y^{(q)} - \hat{y}_2^{(q)}|$, where $y^{(q)}$ is the observed q^{th} quantile, and $\hat{y}_m^{(q)}$ is the q^{th} quantile from the m^{th} model. Then, from the skew-normal density, $L(\hat{y}_1^{(q)}) < L(\hat{y}_2^{(q)})$ for q s.t. $F^{-1}(p_q) > u_2$.

Then, comparing Equations (42) and (43), for p_q large enough, we must have

$$\frac{y_m^{(q)}}{\sum_{j=1}^{Q} y_m^{(j)}} > \frac{1}{Q}.$$

Since more weight is placed on tail quantiles, and $L(\hat{y}_1^{(q)}) < L(\hat{y}_2^{(q)})$ for q s.t. $F^{-1}(p_q) > u_2$, it follows that the weight to $\mathcal{M}_2$ should increase, which means the weight to $\mathcal{M}_1$ must decrease, since $\sum w_m = 1$.

For each additional model, from the same argument, if the model has $u_m \geq u$ ($u_m < u$), then it will have higher (lower) tail likelihood than the models with threshold under (over) u. With the quantile-weighted version of GLUE, the weight to the additional model must thus increase (decrease). Generalising to M models, the result directly follows. □

Note that the proof of Proposition 2 assumes a combination of values both below and above the optimal threshold. It is not immediately clear what would occur when all models have values either exclusively below or above the best one, emphasizing the importance of a wide range of selected thresholds to ensure that the model with the optimal value is included within the combination.

Algorithm 3 describes how to identify the threshold using Proposition 2, where $y^{(q)}$ is the observed q^{th} quantile, $y_b^{(q)}$ is the q^{th} quantile of the b^{th} bootstrap resampling of data $\mathcal{D}$ and $\hat{y}_{m,b}^{(q)}$ is a similar quantile for the m^{th} model, with B bootstrap iterations and Q quantiles.

Algorithm 3: Skewed Generalised Likelihood Uncertainty Estimation

1: Resample $\mathcal{D}$ to obtain B bootstrap iterations $y_b^{(q)}$ of the q^{th} quantile.

2: Calculate the variance for quantile q as $\sigma_q^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(y_b^{(q)} - \frac{1}{B}\sum_{i=1}^{B} y_i^{(q)}\right)^2$.

3: Calculate the skewness for quantile q as $\gamma_q = \frac{1}{B}\frac{\sum_{b=1}^{B}\left(y_b^{(q)} - \frac{1}{B}\sum_{i=1}^{B} y_i^{(q)}\right)^3}{\sigma_q^3}$.

4: Calculate the skew-normal parameters as:

$$\delta_q = \sqrt{\frac{\pi|\gamma_q^{1.5}|}{2(|\gamma_q|^{1.5} + ((4-\pi)/2)^{2/3})}}$$

$$\alpha_q = \frac{\delta_q}{\sqrt{1-\delta_q^2}}$$

$$\omega_q = \sqrt{\frac{\pi * \sigma_q^2}{\pi - 2\alpha_q^2/(1+\alpha_q^2)}}$$

$$\xi_q = y^{(q)} - \omega_q\sqrt{\frac{2\alpha_q}{\pi(1+\alpha_q^2)}}$$

5: Calculate the likelihood and weighted-likelihood assuming residuals follow a skew-normal distribution, with ϕ and Φ the standard normal density and cumulative function respectively:

$$L(\hat{y}_m^{(q)}) = \frac{2}{\omega_q}\left(\prod_{b=1}^{B}\phi\left(\frac{\hat{y}_{m,b}^{(q)} - \xi_q}{\omega_q}\right)\Phi\left(\alpha_q\left(\frac{\hat{y}_{m,b}^{(q)} - \xi_q}{\omega_q}\right)\right)\right)^{1/B} \tag{44}$$

$$L(\mathcal{D}|\mathcal{M}_m) = \frac{1}{Q}\sum_{q=1}^{Q} L(\hat{y}_m^{(q)}).$$

$$L^*(\mathcal{D}|\mathcal{M}_m) = \sum_{q=1}^{Q}\frac{\hat{y}_m^{(q)}}{\sum_{j=1}^{Q}\hat{y}_m^{(j)}} L(\hat{y}_m^{(j)})$$

6: Update the probability of each model as

$$w_m = \frac{L(\mathcal{D}|\mathcal{M}_m)\Pr(\mathcal{M}_m)}{\sum_{l=1}^{M} L(\mathcal{D}|\mathcal{M}_l)\Pr(\mathcal{M}_l)},$$

$$w_m^* = \frac{L^*(\mathcal{D}|\mathcal{M}_m)\Pr(\mathcal{M}_m)}{\sum_{l=1}^{M} L^*(\mathcal{D}|\mathcal{M}_l)\Pr(\mathcal{M}_l)}.$$

7: Identify the optimal threshold u^* with corresponding model $\mathcal{M}_{m^*}$ as

$$u^* = \underset{m}{\operatorname{argmin}}(u_m : w_m^* \geq w_m).$$

8: Calculate posterior distributions as

$$\Pr(y|\mathcal{D}) = \sum_{m=1}^{M}\Pr(y|\mathcal{M}_m)\Pr(\mathcal{M}_m|\mathcal{D}),$$

$$\Pr{}_{u^*}(y|\mathcal{D}) = \Pr(y|\mathcal{M}_{m^*}).$$

Proposition 2 and Algorithm 3 allow us to identify the best threshold. While this is certainly desirable, a natural question in the context of model combination is whether combined models can outperform the model fitted with the right threshold. When fitting the mixture model in equation (39), the threshold automatically implies truncation of the left part of the data. This truncation in turn means that the parameters obtained through MLE for the bulk of the data will be biased. Note that while this bias can be reduced by using censored MLE (see for example Zeng and Lin (2007)), since the truncated observations affect the evaluation of parameters in a finite data setting, some bias will remain. Model combination can reduce this bias by considering multiple parameters simultaneously. In theory, a combination will thus be closer to the true distribution than a model fitted with the best threshold.

4.2.2 Application

To illustrate Proposition 2, consider the well-known Danish reinsurance dataset (McNeil, 1997). There are 2,167 losses between 1 million and 263 million Danish krones, expressed in millions. Embrechts et al. (2013) suggested that a threshold of 10 or 18 is appropriate for this dataset based on the mean residual life (MRL) plot shown in Figure 4.1. We can see that around the suggested points, the mean excess is approximately linear from 10 to 18, then from 18 to 30 with a different slope. This method of selecting a threshold is highly subjective, but gives a reasonable idea of where the threshold might be.

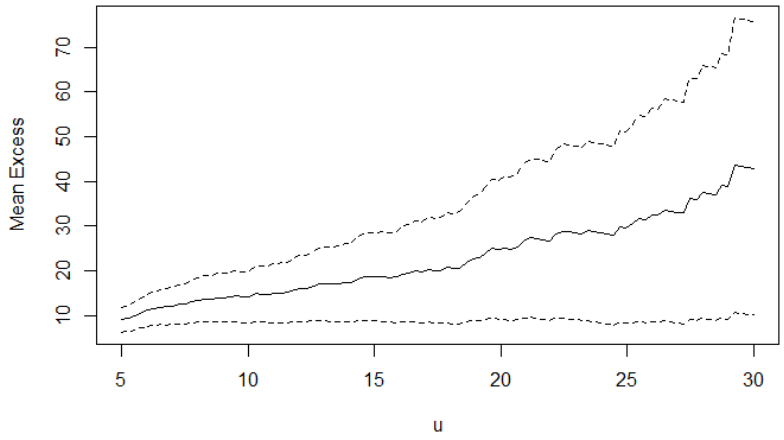

Figure 4.1: Danish mean residual life plot

We randomly split the data into a training and testing set, where both sets have approximately the same size. We then fit mixture models where we suppose that the bulk of the data follows a lognormal distribution and the tail follows a GPD. Figure 4.2 illustrates the weights obtained by combining models with thresholds ranging from 6 to 15 under both GLUE and weighted-GLUE. As expected from Proposition 2, the weight reversal happens at 10, the suggested threshold.

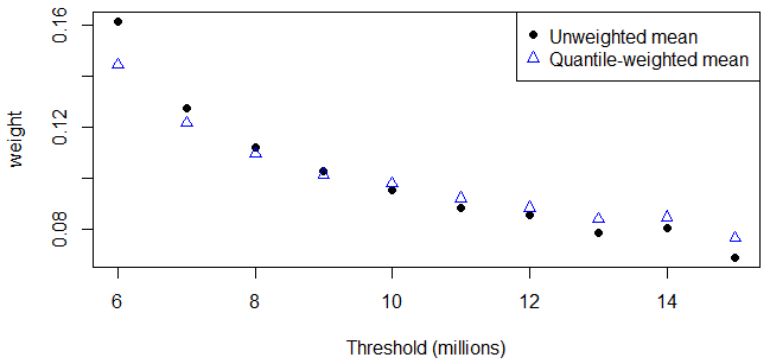

Figure 4.2: Model weights for different thresholds

In order to support the argument that a combination can outperform a fitted model with the right threshold, consider the Hellinger distance (Beran, 1977) and the Kullback-Leibler (KL) divergence (Van Erven and Harremos, 2014). Hellinger distance is defined (under Lebesgue measure) as

$$H^2(f,g) = \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx, \tag{45}$$

which can easily be shown to be equivalent to

$$= 1 - \int \sqrt{f(x)g(x)}dx,$$

where $H^2(f,g) = 0$ if $f = g$ and $H^2(f,g) = 1$ is the case where f and g have entirely different supports.

KL divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = \int f_{\mathbb{P}}(x) \log\left(\frac{f_{\mathbb{P}}(x)}{f_{\mathbb{Q}}(x)}\right) dx. \tag{46}$$

When adjusting the model with a sufficiently high threshold, the parameters for the tail of the distribution are accurate, but the estimated parameters for the lognormal bulk of the data are systematically biased. Indeed, right truncation leads to the location parameter being underestimated, where

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln(x_i) I(x_i < u)}{\sum_{i=1}^{n} I(x_i < u)} < \mu.$$

This should in turn lead to the distance and divergence between this model and the data being higher than with a combination of multiple thresholds.

Table 4.1 shows the Hellinger distance and KL divergence comparing the empirical distribution of the test set to the distributions obtained using GLUE and weighted-GLUE combinations, a simple mean, and the thresholds of 10 and 18 proposed by Embrechts et al. (2013). The results indicate that the GLUE combinations perform similarly to each other and outperform both the simple mean and the single threshold model. Additionally, the mean model outperforms both single thresholds, reinforcing the argument that using combinations is preferable to identifying a single best threshold if we are interested in the full distribution.

			Threshold		
	GLUE	weighted-GLUE	Mean	10	18
Hellinger distance	8.01	8.05	8.23	10.15	13.7
KL divergence	0.166	0.167	0.170	0.228	0.320

Table 4.1: Hellinger distance ($\times 10^{-5}$) and KL divergence by combination method

In addition to fitting the overall distribution, we can compare high-level quantiles obtained through model combination with those derived using an automated threshold selection method, specifically the Anderson-Darling ForwardStop (FS) algorithm proposed by Bader et al. (2018). This method involves fitting a GPD to data above increasing thresholds, using the Anderson-Darling p-value to identify the first threshold that exceeds a certain significance level. By applying the FS

method with a 5% significance level across the 85th to 99th quantiles, we obtain a threshold of 16.55, compared to a threshold of 10 using our approach.

We then compare the absolute differences between projections and observed values in the test dataset for three cases: the FS threshold, our model combination, and our identified threshold at the 99th and 99.5th quantiles. As shown in Table 4.2, when the threshold is misspecified, the FS method produces less accurate results than our identified threshold, which aligns with thresholds commonly accepted in the literature. The model combination yields results similar to those of the correct threshold for high quantiles, as illustrated in Figure 4.3. In the QQ-plots of the random test set from the Danish data, the single threshold and model combination produce very similar outcomes, particularly for the tail. This outcome is theoretically sound, as using the correct threshold reduces bias in the tail, while the combination approach mitigates bias in the bulk of the data.

	Threshold method		
	FS	Identified	Combination
99th quantile	5.5	0.1	2.4
99.5th quantile	10.2	2.1	0.5

Table 4.2: Absolute error (%) of fitted distributions on the test dataset

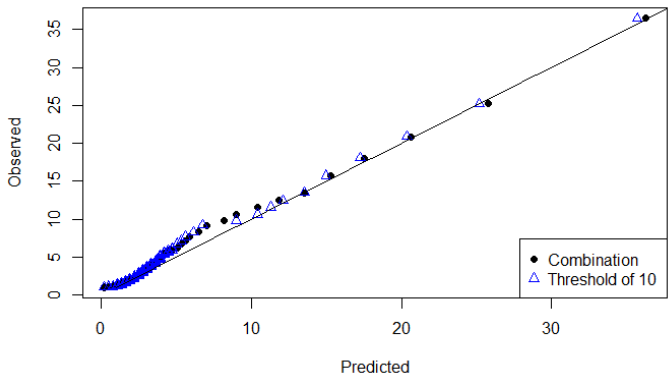

Figure 4.3: QQ-Plot of model combination and the identified threshold for the Danish test set

In a general setting with no predictive variables, comparing the usual GLUE BMA algorithm with a quantile-weighted version thus allows for identifying the correct threshold, and this method models the full distribution instead of only the tail. While the identified threshold provides similar

tail projections as a combination, model combination provides a better Hellinger distance than a single threshold.

4.3 Heterogeneous setting

4.3.1 Theory

Extending beyond a homogeneous setting, the methodology in Section 4.2 can be generalised to incorporate a vector of explanatory variables $\mathbf{X}$ in a regression setting. In climate risk studies, geographical and environmental factors are known to influence risk levels. Similarly, in insurance, predictive variables are used to differentiate between various risks. It is natural to assume that risk levels depend on problem-specific predictive variables, causing the parameters for the extreme portion of risks to vary accordingly. We thus want a mixture model capable of accounting for this dependence. However, a varying threshold becomes complex from a parameter adjustment point of view. As such, consider a generalised mixture model with a fixed threshold such that

$$f(y^{(k)}|\mathbf{X}^{(k)}, \mathbf{\Lambda}(\mathbf{X}^{(k)}), u, \sigma(\mathbf{X}^{(k)}), \xi(\mathbf{X}^{(k)})) = \begin{cases} (1-\phi_u) \times \frac{h(y|\mathbf{X}^{(k)}, \mathbf{\Lambda}(\mathbf{X}^{(k)}))}{H(u|\mathbf{X}^{(k)}, \mathbf{\Lambda}(\mathbf{X}^{(k)}))} & y \leq u \\ \phi_u \times g(y^{(k)}|u, \sigma(\mathbf{X}^{(k)}), \xi(\mathbf{X}^{(k)})) & y > u, \end{cases} \tag{47}$$

where the prediction depends on the characteristics $\mathbf{X}^{(k)}$ for the k^{th} observation. Bulk distributions where parameters depend on predictive variables can be modeled by Generalised Additive Models for Location, Scale, and Shape (GAMLSS, see Rigby and Stasinopoulos (2005)). For the excess over the threshold, we can use Generalised additive extreme value models (evgam, see Youngman (2020)). Note that similarly to the homogeneous case, (47) can lead to point discontinuities, but in this situation there is no direct solution to this issue.

In the homogeneous setting, we used the GLUE algorithm to compare quantiles across the full dataset. However, when predictive variables are present, the quantiles vary depending on these variables, preventing the use of the same approach. Specifically, we lack empirical quantiles for each possible combination of predictive variables, making it impractical to apply the GLUE algorithm. As an alternative, Jessup et al. (2023b) propose an approach that involves working with residuals

instead of quantiles, which we can use in this context. In this framework, suppose that

$$y^{(k)} = E(Y^{(k)}) + \epsilon^{(k)}, \tag{48}$$

where $\epsilon^{(k)}$ is a normally distributed random error. The weights are then approximated as

$$\Pr(\mathcal{M}_m|\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\frac{1}{|\mathcal{D}|}\sum_{y^{(k)}\in\mathcal{D}}\frac{\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_m)}{\sum_{l=1}^{M}\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_l)}, \tag{49}$$

where S is the number of simulations, $\epsilon_s^{(k)}$ is the s^{th} simulation of $\epsilon^{(k)}$, and $|\mathcal{D}|$ is the cardinality of the data.

For a mixture model incorporating extreme values, this approach needs to be adjusted. In the GLUE homogeneous approach, uncertainty is assumed to depend on quantiles. It is reasonable to suppose that similarly, residuals will behave differently between the extreme tail and the bulk of the data. As such, the approach proposed by Jessup et al. (2023b) can be modified to separate the components of the mixture model.

Consider $\mathcal{D}_{B,m}$ and $\mathcal{D}_{T,m}$, the data for the bulk and the tail for each model, where $\mathcal{D}_{B,m} = \{\mathcal{D} : y^{(k)} \leq u_m\}$ and $\mathcal{D}_{T,m} = \mathcal{D}\backslash\mathcal{D}_{B,m}$ depending on the model's specified threshold u_m. For both of these datasets, we want to find parameters $\boldsymbol{\theta}_m$ and $\boldsymbol{\zeta}_m$ for each model $\mathcal{M}_m$ such that $\sigma_m^2 = g_1(\mathbf{X}^{(k)}\boldsymbol{\theta}_m)$ and $\gamma_m = g_2(\mathbf{X}^{(k)}\boldsymbol{\zeta}_m)$. Similarly to Jessup et al. (2023b), we choose an exponential link function g_1 as it ensures positivity of results for the variance estimator and minimise under quadratic loss, such that

$$\hat{\boldsymbol{\theta}}_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}_m, \tag{50}$$

where $\mathbf{X}$ is a matrix of covariates $\mathbf{X}^{(k)}$ and $R_{m,k} = \ln((y^{(k)} - E(\hat{Y}_{m,k}))^2)$, to finally obtain $\tilde{\sigma}_{m,k} = e^{\mathbf{X}^{(k)}\hat{\boldsymbol{\theta}}_m}$.

We can further model skewness under the hypothesis that residual uncertainty might be skewed. Using the same logic as for variance, we set

$$\tilde{\gamma}_k = e^{\mathbf{X}^{(k)}\zeta_k^+} - e^{\mathbf{X}^{(k)}\zeta_k^-}, \tag{51}$$

where

$$\hat{\boldsymbol{\zeta}}_m^+ = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}_m^+ \text{ and } \hat{\boldsymbol{\zeta}}_m^- = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}_m^-, \tag{52}$$

with the second term in equation (51) being subtracted to correct for the negative skewness that was artificially made positive. Full details are available in Jessup et al. (2023b).

In order to modify Algorithm 1 in Jessup et al. (2023b) to account for the difference in residuals stemming from a mixture model, we need to obtain estimators of variance and skewness for both the bulk and the tail, where the bulk uncertainty is assumed skew-normal while the tail uncertainty can be assumed to follow a GEV. The main difference is that variance and skewness are estimated separately for the bulk and the tail, and error-adjusted $\hat{y}_{m,s}^{(k)}$ depend on whether the observation is in the bulk or the tail.

Similarly to Section 4.2, we want to focus mostly on the tail to determine the weights to each model. We can again use a weighted mean instead of an average, such that equation (49) becomes

$$\Pr(\mathcal{M}_m|\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\sum_{y^{(k)}\in\mathcal{D}} \frac{y^{(k)}}{\sum\mathcal{D}} \frac{\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_m)}{\sum_{l=1}^{M}\Pr(y^{(k)}|\epsilon_s^{(k)},\mathcal{M}_l)}. \tag{53}$$

We then obtain a result very similar to Proposition 2, where instead of considering quantiles, we consider the k^{th} observation.

This result is particularly promising when combined to the Dirichlet regression component proposed by Jessup et al. (2023b). We suppose that the weights for the k^{th} observation follow a Dirichlet distribution depending on covariates, such that

$$f(y^{(k)}) = \sum_{m=1}^{M} w_m^{(k)} f_m(y^{(k)}). \tag{54}$$

This allows for flexible weights depending on covariates, which combined to the previous result, leads to Corollary 1.

Corollary 1. *Let M different models $\mathcal{M}_m$, $m \in \{1,\dots,M\}$, be mixture models as defined by equation (47) $\forall$ m, with each model having a different predetermined threshold u_m s.t. the models cover a wide range of possible thresholds. Let $w_m^{(k)}$ and $w_m^{*(k)}$ be the weights to the m^{th} model under respectively the mean and weighted mean Dirichlet regression for the k^{th} observation. Further*

let $u^{(k)}$ *be the best threshold for the* k^{th} *observation. Then* $w_m^{*(k)} \geq w_m^{(k)}$ *if* $u_m \geq u^{(k)}$*, and* $w_m^{*(k)} \leq w_m^{(k)}$ *if* $u_m \leq u^{(k)}$.

Proof. The proof is nearly identical to the proof of Proposition 2, where we consider observations k instead of quantiles q. □

Corollary 1 allows for identifying flexible thresholds based on covariates. Additionally, this approach provides the entire distribution alongside the thresholds. This represents an improvement over methods like quantile regression, where, for instance, Youngman (2019) outlines a two-step procedure: using quantile regression to identify thresholds and then fitting a GAM version of GPD to the excesses over those thresholds. Such methods do not model the full distribution.

Moreover, once again, through bias reduction, flexible model combination can outperform a single model with the right threshold.

4.3.2 Application

To illustrate this generalised idea, we work with an automobile claims dataset from a Canadian insurer. We have data from over one million claims from 2015 to 2021 for multiple coverages. We choose to study only Vehicle Damage claims in Ontario, which has claims between 2 and 561,000 dollars.

We separate data into a training and testing set by taking historical data from 2015 to 2019 as the training data and the more recent 2020 and 2021 losses as the testing data. The training data is further separated to include a calibration component from which combination weights can be calculated by randomly sampling 30% of claims in the training set.

We set potential thresholds as the 50th to 97.5th quantiles by jumps of 2.5%, a common approach in automatic threshold selection algorithms. To use model combination, for each model with a different threshold, we fit a GAMLSS lognormal model on the bulk and a GAM version of the GPD on the tail. For ease of interpretability, Figure 4.4 presents the Dirichlet results of the error integration (EI) BMA algorithm, comparing weight variation when gender is available versus unavailable. The weights are nearly identical between males and females, so only one figure is presented for both. To reduce calculation time, we regroup quantiles to have 10 weights instead of

20. The threshold identified by our method varies based on the predictive variable, showing a lower threshold when gender is not available, that is, approximately 7000 compared to 8700 for male and female.

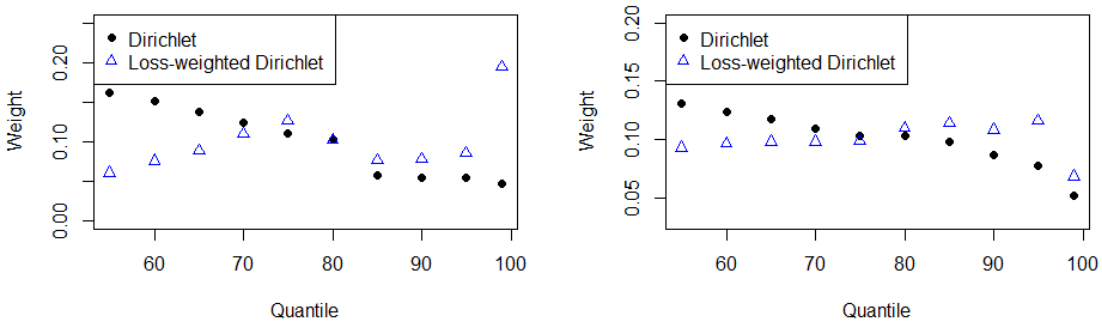

Figure 4.4: Weights by threshold quantile when gender is unavailable (left) and available (right)

This difference in threshold can be explained by the quartile values and variance shown in Table 4.3. We see that while losses are mostly lower when gender is not available, there is significantly more tail uncertainty than when gender is known, as reflected by a larger variance despite the 75^{th} quantile being the lowest.

	Male	Female	Not available
25%	745	809	217
50%	2591	2467	1485
75%	5808	5384	4717
Variance	$7.02 * 10^7$	$4.74 * 10^7$	$1.06 * 10^8$

Table 4.3: Quartile and variance values by gender

Figure 4.5 illustrates the MRL plots obtained by gender, along with the identified threshold values. We see that male and female have similar MRL curves, while the other curve behaves differently. This can offer further insight into why the threshold is different, where the distribution seems significantly different when gender is not available.

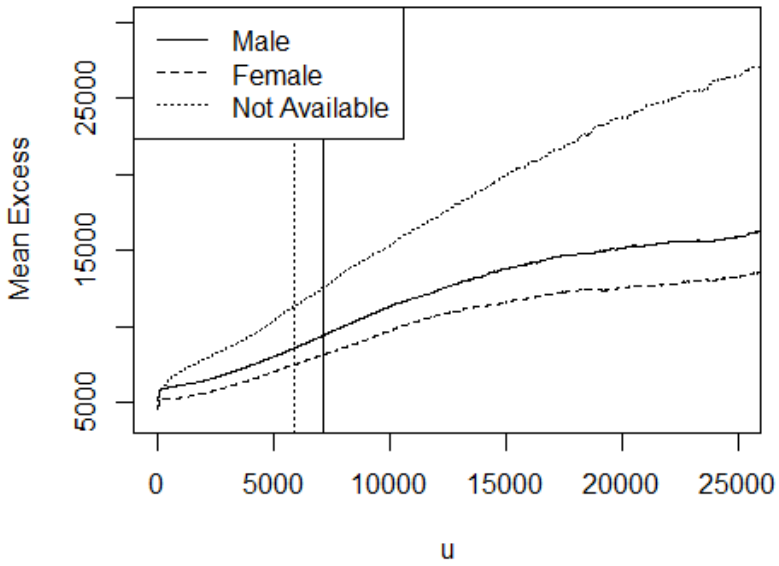

Figure 4.5: MRL plot by gender with the identified thresholds

Using the weights in Figure 4.4, we can obtain densities and predict high quantiles by gender. Table 4.4 shows the Hellinger distance between the empirical density function of the test dataset and the densities obtained with weighted-EI, compared to the identified threshold. Across the full distribution, the combined approach is closer to the true distribution than a single threshold. Only the tail-weighted results are presented, as we are focused on high quantiles, where these weights are expected to perform better than a simple mean of the likelihood of each observation.

	weighted-EI	Single threshold
Hellinger distance	1.35	1.85
KL divergence	0.10	0.14

Table 4.4: Hellinger distance $(\times 10^{-5})$ and KL divergence by combination method

Next, Table 4.5 compares the mean absolute error (MAE) as a percentage of quantiles obtained with model combination and with the two-step procedure of quantile regression with a fixed level of 0.97, then fitting a GAM version of GPD to the excess, as suggested by Youngman (2019). Given the significant uncertainty for high level quantiles, the two-step quantile regression and model combination offer similar performance, where the combination is only slightly better than quantile

regression for the 99.5th and 99.9th quantiles, while the single identified threshold performs better for the 99.5th quantile, but not for the 99.9th quantile. Again, the similarity between the identified threshold and the combination for tail quantiles makes sense from a theoretical point of view, where we expect the tail to be unbiased, such that model combination does not offer an improvement, while it reduces bias for the bulk of the data. We can further compare the fit over the full tail using a normalised root mean squared error (NRMSE, see Curceac et al. (2020)). We find that the identified thresholds and full combination yield nearly identical values of 0.167 and 0.168, compared to 0.248 for the quantile regression approach. Since these values are close to 0, we can conclude that all methods yield good results.

	Threshold method		
	QR	Identified	Combination
99th quantile	19.1	16.9	15.1
99.5th quantile	16.3	12.0	15.3
99.9th quantile	13.7	13.7	12.4

Table 4.5: MAE (%) of fitted distributions for the test dataset

Figure 4.6 shows the QQ-plots over the full distribution for the model combination as well as the mixture model with the identified threshold. We can see that the results are quite similar, with the combination performing slightly better, especially for the bulk of the data, and the large MAE being explained by the higher quantiles when gender is not available.

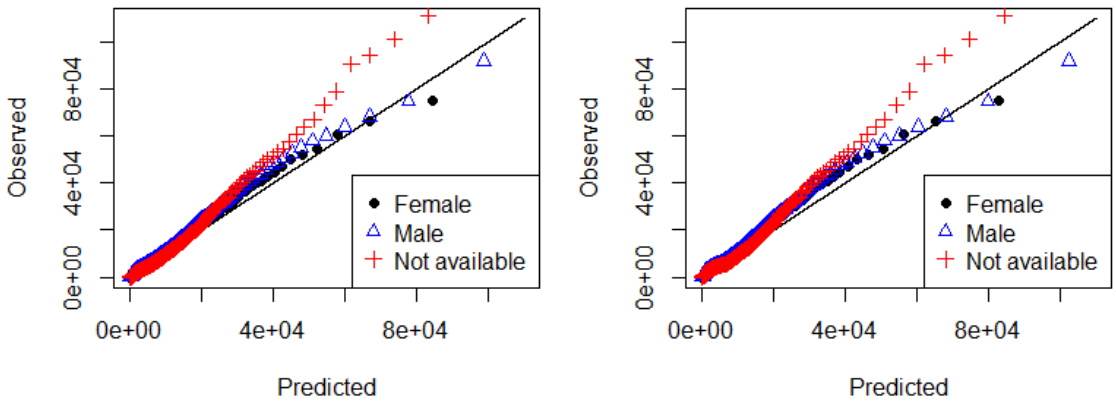

Figure 4.6: QQ-Plots for combined mixture models (left) and the identified threshold (right)

In light of these results, the main advantage of model combination over quantile regression is that it allows for obtaining a full distribution of the data depending on predictive variables. This shows a promising application of Proposition 1, where the idea could be used when studying climate risk to have a combined model with stations and climate factors as variables to obtain the full distribution instead of studying only high quantiles. Moreover, from the similarity in tail quantiles between the identified thresholds and the combinations in both the general and predictive setting, if the goal is only to predict tail quantiles, only the models with the identified threshold could be kept, which would allow for faster computing.

4.4 Conclusion

In this paper, we proposed a modification of the Bayesian model averaging GLUE algorithm to identify the "best" threshold in a homogeneous setting. By combining mixture models, we demonstrated the effectiveness of our approach using the Danish reinsurance dataset. We showed that model combination can outperform a single model with the correct threshold and is preferable for forecasting high quantiles.

Additionally, we modified the error integration (EI) BMA algorithm proposed by Jessup et al. (2023b) to combine mixture models that include predictive variables. When combined with the Dirichlet regression component of the EI algorithm, this approach allows for identifying flexible thresholds based on predictive variables. We compared our method to a two-step quantile regression procedure and found that it provides similar high quantile predictions. Our method's main advantage is producing a full distribution rather than only modelling the excess over the thresholds. This is particularly useful in insurance contexts, where extreme values are important, but the bulk of the data is also of interest.

For future work, it would be interesting to consider multiple coverages simultaneously, such as different car insurance coverages known to be correlated. Dependence could affect threshold values, requiring multiple coverages to be considered together. While this does not affect marginal distributions, it could improve overall loss projections. Additionally, applying our flexible threshold approach to datasets like extreme precipitation, where different stations are treated as variables rather than separately, could increase data availability and enhance analysis.

Chapter 5

Conclusion

This Chapter concludes the thesis, which is based on three manuscripts focused on advancements in model combination and uncertainty quantification applied to actuarial science.

In Chapter 2, we explore various model combination methods to illustrate the uncertainty inherent in model combination. Using two non-parametric methods and two Bayesian model averaging algorithms, we derive weights for an ensemble of 24 experts. These methods are applied to generate predictive densities for the annual maxima of daily rainfall in Montreal and Quebec City. By employing areal reduction factors and quantile projected changes, we demonstrate that non-parametric combination methods produce significantly different outcomes compared to parametric combination methods. We emphasize the importance for actuaries to consider multiple combination methods to avoid overconfidence in their projections.

In Chapter 3, we introduce a novel approach to Bayesian model averaging that accounts for data uncertainty. We critique the BMA Expectation-Maximisation algorithm for treating the data as the only observable data, leading to convergence on a single model unless a stopping criterion is specified. Our approach uses residuals to model data uncertainty and integrates random error numerically, enabling a single-sweep weight update that does not converge on one model. We illustrate the conditions under which this error integration algorithm performs optimally and propose treating combination weights as Dirichlet variables, allowing weights to vary with predictive variables. These methods are validated through simulation studies and a Property & Casualty simulated insurance dataset, demonstrating that flexible weights are particularly advantageous for reserving purposes,

where Dirichlet regression facilitates a smooth transition between different reserving methods for different data segments.

In Chapter 4, we demonstrate that BMA can be employed to identify extreme value thresholds in both homogeneous and heterogeneous settings. In the homogeneous setting, by combining mixture models where each model shares the same structure but has a different threshold, we show that placing more weight on tail quantiles allows for a reversal in combination weights at the correct threshold. This approach outperforms an automated threshold selection method on the well-known Danish reinsurance dataset. In the heterogeneous setting, a similar concept of weighting larger losses is applied, with a slight modification to the algorithm proposed in Chapter 3 to achieve flexible thresholds depending on predictive variables. Our method yields comparable projections to a two-step quantile regression procedure, with the added advantage of projecting the full distribution rather than just the tail.

The ideas presented in Chapters 3 and 4 pave the way for numerous future research opportunities. First, by integrating the approaches from both chapters, we can develop a reserve model that takes extreme values into account, addressing the exclusion of extreme losses in Chapter 3. Additionally, flexible weights can be used to identify change-point models. Lastly, thresholds that depend on predictive variables enable large-scale studies of extreme weather events, allowing for the comprehensive use of all available data instead of analysing each location individually.

Appendix A

Appendices

A.1 Expectation-Maximisation Bayesian Model Averaging algorithm

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

Algorithm 5: Expectation-Maximisation Bayesian Model Averaging

1: Initialize variance and weights as

$$\sigma^{2(0)} = \frac{1}{QM}\sum_{q=1}^{Q}\sum_{m=1}^{M}\left(y_{\vec{\tau},x,q} - y_{\vec{\tau},x,q}^{(m)}\right)^2,$$

$$w_m^{(0)} = 1/M \;\forall m.$$

2: Calculate initial likelihood as

$$l(\theta^{(0)}) = \sum_{q=1}^{Q}\log\left(\sum_{m=1}^{M} w_q^{(0)}\phi(y_{\vec{\tau},x,q}|y_{\vec{\tau},x,q}^{(m)}, \sigma^{2(0)})\right).$$

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

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

$$z_{m,q}^{(j)} = \frac{w_m^{(j-1)}\phi(y_{\vec{\tau},x,q}|y_{\vec{\tau},x,q}^{(m)}, \sigma^{2(j-1)})}{\sum_{m=1}^{M} w_m^{(j-1)}\phi((y_{\vec{\tau},x,q}|y_{\vec{\tau},x,q}^{(m)}, \sigma^{2(j-1)})}.$$

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

$$w_m^{(j)} = \frac{1}{Q}\sum_{q=1}^{Q} z_{m,q}^{(j)}$$

$$\sigma_m^{2\,(j)} = \frac{\sum_{q=1}^{Q} z_{m,q}^{(j)}(y_{\vec{\tau},x,q} - y_{\vec{\tau},x,q}^{(m)})^2}{\sum_{q=1}^{Q} z_{m,q}^{(j)}}.$$

6: Calculate updated likelihood as

$$l(\theta^{(j)}) = \sum_{q=1}^{Q}\log\left(\sum_{m=1}^{M} w_m^{(j)}\phi(y_{\vec{\tau},x,q}|y_{\vec{\tau},x,q}^{(m)}, \sigma^{2(j)})\right).$$

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

8: **end while**

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

10: Calculate posterior distribution as

$$\Pr(Y_{\vec{\psi},x} = y|\vec{y}_{\vec{\tau},x}) = \sum_{m=1}^{M}\Pr(Y_{\vec{\psi},x} = y|\mathcal{M}_m)\Pr(\mathcal{M} = \mathcal{M}_m|\vec{y}_{\vec{\tau},x}).$$

A.2 Quantile and ARF changes bootstrap distribution for a 1 in 20 year return level for Quebec

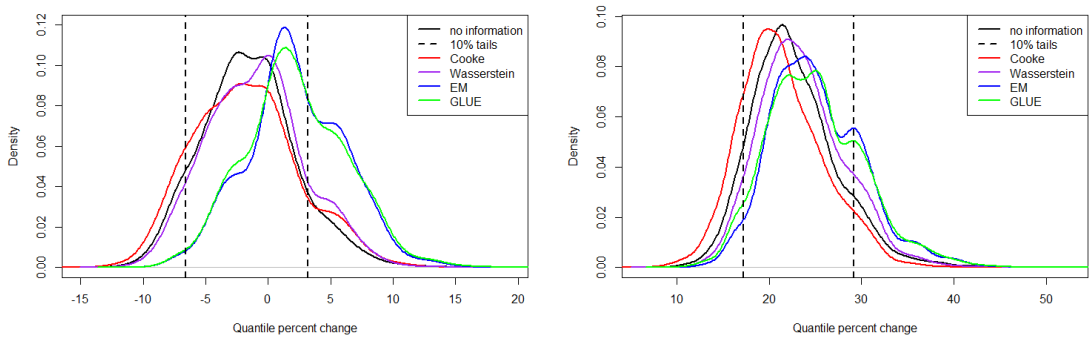

Figure A.1: Distribution of projected quantile change at a 1 in 20 year return level in Quebec between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

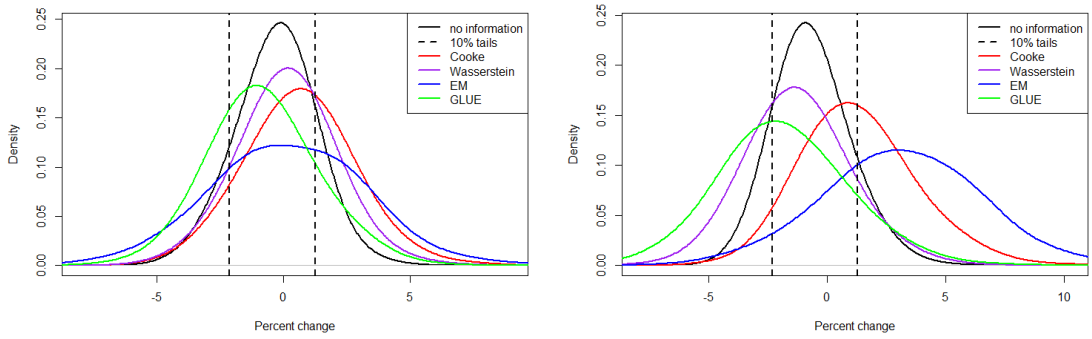

Figure A.2: Distribution of projected ARF change at a 1 in 20 year return level in Quebec between 2001-2020 and 2011-2030 (left) or 2071-2090 (right)

A.3 Quantile and ARF percent changes for a 1 in 20 year return level for Quebec

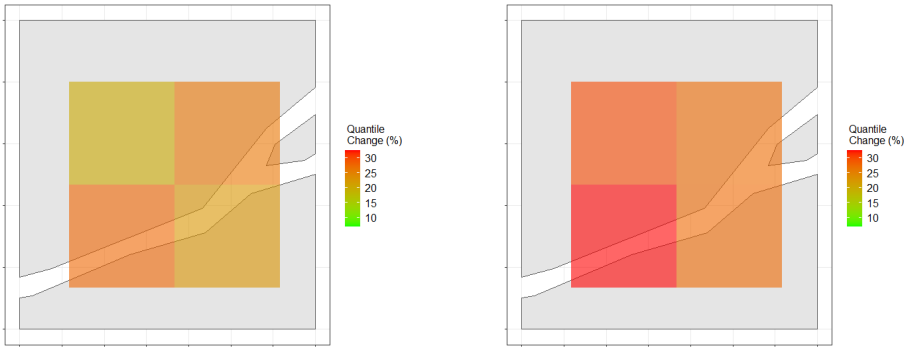

Figure A.3: Percentage change in quantiles for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Quebec using Cooke's method (left) and BMA-EM (right)

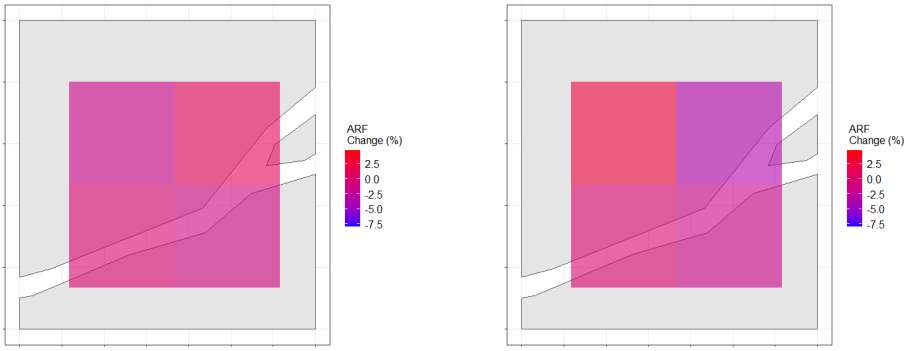

Figure A.4: Percentage change in quantiles for a 1 in 20 year return level between 2001-2020 and 2071-2090 for the region of Quebec using Cooke's method (left) and BMA-EM (right)

A.4 Proof of heteroscedastic BMA weights

Following a logic similar to Conflitti et al. (2015), for the j^{th} iteration, take

$$\eta(\mathbf{w}^{(j)}) = \sum_{y^{(k)} \in \mathcal{D}} \log \left(\sum_{m=1}^{M} w_m^{(j)} \tilde{f}_m(y^{(k)}) \right) - \lambda \sum_{m=1}^{M} w_m^{(j)},$$

where λ is a Lagrange multiplier, subject to the constraint that $\sum_{m=1}^{M} w_m = 1$. We cannot directly optimise this function, seeing as deriving with respect to a particular $w_n^{(j)}$ would not allow for

isolating this weight. Instead, we consider an alternative function

$$\psi(\mathbf{w}^{(j)}, \mathbf{a}) = \sum_{y^{(k)} \in \mathcal{D}} \sum_{m=1}^{M} \frac{\tilde{f}_m(y^{(k)}) a_m}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l} \ln \left(\frac{w_m^{(j)}}{a_m} \sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l \right) - \lambda \sum_{m=1}^{M} w_m^{(j)},$$

with $\mathbf{a} = \{a_1, \ldots, a_M\}$. We have that $\psi(\mathbf{a}, \mathbf{a}) = \eta(\mathbf{a})$ for any $\mathbf{a}$, and $\psi(\mathbf{w}^{(j)}, \mathbf{a}) \leq \eta(\mathbf{w}^{(j)})$ for any $\mathbf{a}$ and $\mathbf{w}^{(j)}$, so $\psi(\mathbf{w}^{(j)}, \mathbf{a})$ is an appropriate alternative function (Conflitti et al., 2015). Indeed,

$$\begin{aligned}
\psi(\mathbf{a}, \mathbf{a}) &= \sum_{y^{(k)} \in \mathcal{D}} \sum_{m=1}^{M} \frac{\tilde{f}_m(y^{(k)}) a_m}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l} \ln \left(\frac{a_m}{a_m} \sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l \right) - \lambda \sum_{m=1}^{M} a_m \\
&= \sum_{y^{(k)} \in \mathcal{D}} \ln \left(\sum_{l=1}^{M} a_l \tilde{f}_l(y^{(k)}) \right) - \lambda \sum_{m=1}^{M} a_m \\
&= \eta(\mathbf{a}),
\end{aligned}$$

and the inequality condition follows from the concavity of the $\ln$ function, where

$$\begin{aligned}
\sum_{m=1}^{M} \frac{\tilde{f}_m(y^{(k)}) a_m}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l} \ln \left(\frac{w_m^{(j)}}{a_m} \sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l \right) &\leq \ln \left(\sum_{m=1}^{M} \frac{\tilde{f}_m(y^{(k)}) a_m}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l} \frac{w_m^{(j)}}{a_m} \sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l \right) \\
&= \ln \left(\sum_{m=1}^{M} \tilde{f}_m(y^{(k)}) w_m^{(j)} \right).
\end{aligned}$$

Then, deriving with respect to the n^{th} weight,

$$\frac{d\psi(\mathbf{w}^{(j)}, \mathbf{a})}{dw_n^{(j)}} = 0 = \frac{1}{w_n^{(j)}} \sum_{y \in \mathcal{D}} \frac{\tilde{f}_n(y^{(k)}) a_n}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l} - \lambda$$

from which we obtain

$$\hat{w}_n^{(j)} = \frac{1}{\lambda} \sum_{y^{(k)} \in \mathcal{D}} \frac{\tilde{f}_n(y^{(k)}) a_n}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l}$$

and using the constraint on $\mathbf{w}^{(j)}$, we can replace λ s.t.

$$\sum_{m=1}^{M} w_m^{(j)} = 1 = \sum_{m=1}^{M} \frac{1}{\lambda} \sum_{y^{(k)} \in \mathcal{D}} \frac{\tilde{f}_m(y^{(k)}) a_m}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l}$$

$$\Leftrightarrow \lambda = |\mathcal{D}|,$$

and finally

$$\hat{w}_m^{(j)} = \frac{1}{|\mathcal{D}|} \sum_{y^{(k)} \in \mathcal{D}} \frac{\tilde{f}_m(y^{(k)}) a_m}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) a_l}.$$

Now, set a_m as $w_m^{(j-1)}$, i.e. the weight from the previous iteration, and so

$$\hat{w}_m^{(j)} = \frac{1}{|\mathcal{D}|} \sum_{y^{(k)} \in \mathcal{D}} \frac{\tilde{f}_m(y^{(k)}) w_m^{(j-1)}}{\sum_{l=1}^{M} \tilde{f}_l(y^{(k)}) w_l^{(j-1)}}.$$

We thus obtain the same update formula as the standard algorithm since this phase of EM does not depend on σ_k.

A.5 Proof of Kullback-Leibler conditions

We want to show that if $\sum_{k=1}^{n} | \sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2 - \sigma_k^2 | \leq \sum_{k=1}^{n} |\sigma_{m^*,k}^2 - \sigma_k^2|$ and $\sum_{k=1}^{n} \log \left(\frac{\sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2}{\sigma_{m^*,k}^2} \right) \geq 0$, then $D_{KL}(\hat{\mathbb{P}} || \hat{\mathbb{Q}}^{(\text{EI})}) \leq D_{KL}(\hat{\mathbb{P}} || \hat{\mathbb{Q}}^{(\text{BMA})})$.
Consider the case with $n = 1$, and let $a_k = \sum_{m=1}^{M} w_m^2 \sigma_{m,k}^2$, $b_k = \sigma_{m^*,k}^2$, and $c_k = \sigma_k^2$ for brevity of notation. Let $|a_1 - c_1| \leq |b_1 - c_1|$ and let $a_1 \geq b_1$, s.t. $\log(a_1/b_1) \geq 0$.

If $a_1 \leq c_1$, then $b_1 \leq c_1$, so by the Taylor expansion of $\log(a_1/b_1)$,

$$\log \left(\frac{a_1}{b_1} \right) \leq \frac{a_1 - b_1}{b_1} \leq \frac{c_1}{a_1} \frac{a_1 - b_1}{b_1},$$

and $D_{KL}(\hat{\mathbb{P}} || \hat{\mathbb{Q}}^{(\text{EI})}) \leq D_{KL}(\hat{\mathbb{P}} || \hat{\mathbb{Q}}^{(\text{BMA})})$.

If $a_1 \geq c_1$, then $b_1 \leq c_1$ and $c_1 - b_1 \geq a_1 - c_1$. Again using the Taylor expansion of $\log(a_1/b_1)$,

$$\frac{c_1(a_1 - b_1)}{a_1 b_1} \geq \frac{(a_1 + b_1)(a_1 - b_1)}{2a_1 b_1} \geq \log(a_1/b_1),$$

and again $D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}^{(\text{EI})}) \leq D_{KL}(\hat{\mathbb{P}}||\hat{\mathbb{Q}}^{(\text{BMA})})$.

Now suppose the inequality is true for n observations, s.t.

$$\sum_{k=1}^{n} \log\left(\frac{a_k}{b_k}\right) \leq \sum_{k=1}^{n} \frac{c_k(a_k - b_k)}{a_k b_k}.$$

Then it follows that for an $n + 1^{th}$ observation,

$$\begin{aligned}\sum_{k=1}^{n+1} \log\left(\frac{a_k}{b_k}\right) &\leq \sum_{k=1}^{n} \frac{c_k(a_k - b_k)}{a_k b_k} + \log\left(\frac{a_{n+1}}{b_{n+1}}\right) \\ &\leq \sum_{k=1}^{n+1} \frac{c_k(a_k - b_k)}{a_k b_k}\end{aligned}$$

by the arguments above, and so the result holds in general.

A.6 Fitted Dirichlet log-coefficients

	Model				
AY	Gamma	ODP	DGLM	Agg. GLM	Agg. GAMLSS
2	-0.46	-0.49	-0.73	-0.52	-2.70
3	-1.42	-1.34	-1.58	-1.20	-4.92
4	-1.11	-0.93	-1.16	-0.91	-2.39
5	0.81	1.33	1.35	1.44	-6.29
6	-1.76	-1.82	-2.45	-1.73	-4.51
7	-1.74	-1.62	-2.04	-1.50	-5.42
8	-1.75	-1.59	-1.99	-1.58	-4.12
9	-1.78	-2.08	-3.67	-2.13	-6.43
10	-2.16	-2.71	-3.59	-2.67	-5.92
11	-1.57	-1.91	-2.53	-1.93	-5.50
12	-1.72	-2.13	-2.56	-2.09	-5.40

Table A.1: Dirichlet log-coefficients without strong predictor

	Model				
AY	Gamma	ODP	DGLM	Agg. GLM	Agg. GAMLSS
1	-1.28	-1.38	-1.33	-1.42	-3.54
2	-2.19	-2.12	-2.09	-2.31	-5.60
3	-1.63	-1.59	-1.46	-1.66	-3.03
4	-0.09	0.14	0.27	0.06	-6.93
5	-2.34	-2.46	-2.23	-2.41	-5.00
6	-2.64	-2.66	-2.40	-3.19	-7.19
7	-2.74	-2.79	-2.47	-3.01	-6.03
8	-2.11	-2.69	-2.15	-2.72	-6.75
9	-2.67	-2.98	-2.54	-3.35	-6.57
10	-3.07	-3.24	-3.00	-3.59	-7.13
11	-3.56	-3.66	-3.47	-3.78	-6.86
12	-1.81	-1.28	-1.99	-0.98	-1.89

Table A.2: Dirichlet log-coefficients with strong predictor

A.7 The skew-normal distribution

The skew-normal distribution, first introduced by Azzalini (1985), is defined as

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right),$$

where ξ, ω, and α are respectively the location, scale, and shape parameters. ϕ and Φ are the standard normal density and cumulative distribution functions. The relationship between ξ and ω and the usual normal parameters μ and σ can be expressed through the mean and variance, such that

$$\mu = \xi + \omega\frac{\alpha}{\sqrt{1+\alpha^2}}\sqrt{\frac{2}{\pi}}$$
$$\sigma^2 = \omega^2\left(1 - \frac{2}{\pi}\frac{\alpha^2}{1+\alpha^2}\right).$$

In particular, if $\alpha = 0$, then $\xi = \mu$ and $\omega = \sigma$.

The equivalence with the normal distribution can easily be shown by setting $\alpha = 0$ as follows.

$$\begin{aligned} f(x) &= \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right) \\ &= \frac{1}{\omega}\phi\left(\frac{x-\xi}{\omega}\right) \\ &= \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

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

Antonio, K. and Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 2014(7):649–669.

Avanzi, B., Li, Y., Wong, B., and Xian, A. (2024). Ensemble distributional forecasting for insurance loss reserving. *Scandinavian Actuarial Journal*, pages 1–42.

Avanzi, B., Taylor, G., Wang, M., and Wong, B. (2021). SynthETIC: an individual insurance claim simulator with feature control. *Insurance: Mathematics and Economics*, 100:296–308.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 12(2):171–178.

Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Annals of Applied Statistics*, 12(1):310–329.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Benito, S., López-Martín, C., and Navarro, M. Á. (2023). Assessing the importance of the choice threshold in quantifying market risk under the POT approach (EVT). *Risk Management*,

25(1):6.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463.

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.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, 7(4):401–406.

Bladt, M. and Yslas, J. (2023). Robust claim frequency modeling through phase-type mixture-of-experts regression. *Insurance: Mathematics and Economics*, 111:1–22.

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, article 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.

Caeiro, F. and Gomes, M. I. (2015). Threshold selection in extreme value analysis. *Extreme value modeling and risk analysis: Methods and applications*, pages 69–87.

CatIQ (2024). Canadian Insured Losses From Catastrophic Events Exceed CAN $3 Billion In 2023. https://public.catiq.com/2024/01/08/canadian-insured-losses-from-catastrophic-events-exceed-can-3-billion-in-2023/.

Chavez-Demoulin, V., Embrechts, P., and Nešlehová, J. (2006). Quantitative models for operational risk: extremes, dependence and aggregation. *Journal of Banking & Finance*, 30(10):2635–2658.

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.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Climatic change*, 127:353–369.

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/.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Conflitti, C., De Mol, C., and Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4):1096–1103.

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.

Curceac, S., Atkinson, P. M., Milne, A., Wu, L., and Harris, P. (2020). An evaluation of automated gpd threshold selection methods for hydrological extremes across different scales. *Journal of Hydrology*, 585, article 124845.

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

De Jong, P., Heller, G. Z., et al. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

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.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Djeundje, V. B. (2022). On the integration of deterministic opinions into mortality smoothing and forecasting. *Annals of Actuarial Science*, 16(2):1–17.

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. *Proceedings of the 17th International Conference on Machine Learning*, 747:223–230.

Duval, F. and Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, 7(3):79.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

European Insurance and Occupational Pensions Authority (2023). Solvency II. https://www.eiopa.europa.eu/browse/regulation-and-policy/solvency-ii_en.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press.

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.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, 6:93–116.

Friedland, J. (2010). *Estimating unpaid claims using basic techniques*, volume 201. Casualty Actuarial Society.

Frigyik, B. A., Kapila, A., and Gupta, M. R. (2010). Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, 6:1–27.

Fu, C. and Sayed, T. (2023). Dynamic bayesian hierarchical peak over threshold modeling for real-time crash-risk estimation from conflict extremes. *Analytic methods in accident research*, 40,

article 100304.

Gabrielli, A., Richman, R., and Wüthrich, M. V. (2020). Neural network embedding of the over-dispersed poisson reserving model. *Scandinavian Actuarial Journal*, 2020(1):1–29.

Gabrielli, A. and Wüthrich, M. (2018). An individual claims history simulation machine. *Risks*, 6(2):29.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, pages 473–483.

Ghaddab, S., Kacem, M., de Peretti, C., and Belkacem, L. (2023). Extreme severity modeling using a glm-gpd combination: application to an excess of loss reinsurance treaty. *Empirical Economics*, 65(3):1105–1127.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

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

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia university press.

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.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, 44:461–465.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

He, Y., Peng, L., Zhang, D., and Zhao, Z. (2022). Risk analysis via generalized pareto distributions. *Journal of Business & Economic Statistics*, 40(2):852–867.

Hershey, J. R. and Olsen, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages 317–321.

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.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

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.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

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.

Jessup, S., Mailhot, M., and Pigeon, M. (2023a). Impact of combination methods on extreme precipitation projections. *Annals of Actuarial Science*, 17(3):459–478.

Jessup, S., Mailhot, M., and Pigeon, M. (2023b). Uncertainty in heteroscedastic bayesian model averaging. *Available at SSRN 4650798*.

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

Judge, G. G. and Mittelhammer, R. C. (2004). A semiparametric basis for combining estimation

problems under quadratic loss. *Journal of the American Statistical Association*, 99(466):479–487.

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

Kapetanios, G., Mitchell, J., Price, S., and Fawcett, N. (2015). Generalised density forecast combinations. *Journal of Econometrics*, 188(1):150–165.

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.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190.

Laudagé, C., Desmettre, S., and Wenzel, J. (2019). Severity modeling of extreme insurance claims for tariffication. *Insurance: Mathematics and Economics*, 88:77–92.

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.

Le, T. M. and Clarke, B. (2022). Model averaging is asymptotically better than model selection for prediction. *The Journal of Machine Learning Research*, 23(1):1463–1516.

Li, J. and Liu, J. (2023). Claims modelling with three-component composite models. *Risks*, 11(11):196–211.

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.

Li, Y., Tang, N., and Jiang, X. (2016). Bayesian approaches for analyzing earthquake catastrophic risk. *Insurance: Mathematics and Economics*, 68:110–119.

Liu, C. and Maheu, J. M. (2009). Forecasting realized volatility: a bayesian model-averaging approach. *Journal of Applied Econometrics*, 24(5):709–733.

Liu, M., Sun, X., Qiao, Y., and Wang, Y. (2023). Causal discovery with unobserved variables: A proxy variable approach. *arXiv preprint arXiv:2305.05281*.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible

extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23(2):213–225.

Masoudnia, S. and Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293.

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.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27(1):117–137.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

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

Mohandes, M., Deriche, M., and Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6:19626–19639.

Northrop, P. J., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(1):93–120.

Office of the Superintendent of Financial Institutions (2023). Own Risk and Solvency Assessment. https://www.osfi-bsif.gc.ca/Eng/fi-if/rg-ro/gdn-ort/gl-ld/Pages/e1918.aspx.

Ossberger, J. (2020). Package 'tea'.

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.

Pešta, M. and Okhrin, O. (2014). Conditional least squares and copulae in claims reserving for a single line of business. *Insurance: Mathematics and Economics*, 56:28–37.

Pickands, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131.

Pigeon, M. and Denuit, M. (2011). Composite lognormal–pareto model with random threshold. *Scandinavian Actuarial Journal*, 2011(3):177–192.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5):1155–1174.

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.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

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

Silva Lomba, J. and Fraga Alves, M. I. (2020). L-moments for automatic threshold selection in extreme value analysis. *Stochastic environmental research and risk assessment*, 34(3):465–491.

Smyth, G. K. and Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin*, 32(1):143–157.

Spedicato, G. A., Clemente, A. G. P., and Schewe, F. (2014). The use of GAMLSS in assessing the distribution of unpaid claims reserves. *Casualty Actuarial Society E-Forum*, vol. 2.

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

Taylor, G. (2012). *Loss reserving: an actuarial perspective*, volume 21. Springer Science & Business Media.

Taylor, G. and McGuire, G. (2016). Stochastic loss reserving using generalized linear models. *CAS Monograph*, 3:1–112.

Thiombiano, A. N., El Adlouni, S., St-Hilaire, A., Ouarda, T. B., and El-Jabi, N. (2017). Non-stationary frequency analysis of extreme daily precipitation amounts in Southeastern Canada using a peaks-over-threshold approach. *Theoretical and Applied Climatology*, 129:413–426.

Tseung, S. C., Badescu, A., Fung, T. C., and Lin, X. S. (2020). LRMoE: an R package for flexible actuarial loss modelling using mixture of experts regression model. *Available at SSRN 3740215*.

Tseung, S. C., Chan, I. W., Fung, T. C., Badescu, A. L., and Lin, X. S. (2022). A posteriori risk classification and ratemaking with random effects in the mixture-of-experts model. *arXiv preprint arXiv:2209.15212*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

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.

Webb, G. I. and Zheng, Z. (2004). Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):980–991.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6):465–480.

Wüthrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*. John Wiley & Sons.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32:1–34.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an

application to return level estimation for us wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Youngman, B. D. (2020). evgam: An r package for generalized additive extreme value models. *arXiv preprint arXiv:2003.04067*.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):507–564.

Zhao, X. and Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, 46(2):290–299.

Zhao, X. B., Zhou, X., and Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics*, 45(1):1–8.

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.