Flood frequency analysis at multiple durations

Danielle M. Barna

Supervisors:
Kolbjørn Engeland
Thordis Thorarinsdottir
Chong-Yu Xu

The Faculty of Mathematics and Natural Sciences

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2711

ISSN 1501-7710

Cover: UiO.
Print production: Graphic center, University of Oslo.

Summary

Flooding currently affects more people than almost any other natural hazard (Van Loenhout et al., 2020) and the proportion of the world's population that lives in flood-exposed areas is growing rapidly (Tellman et al., 2021). Part of the way society manages exposure to flood risk is through estimation of *flood design values*. These values give estimates of flood magnitude within a given return period and are essential to making adaptive decisions for a variety of hydrologic applications, e.g., infrastructure design, land use planning, and water resource management. For flood retention-specific applications—e.g. floodplain management and reservoir design—we often need design values at multiple durations. Here our focus is the retention capacity of a man-made or natural basin. We are therefore concerned with the total flow volume we can expect to see over a short duration like one hour vs a longer one like one day, regardless of whether that volume of water comes from a single event or multiple consecutive events.

This thesis explores and develops statistical methods for obtaining design values at different durations for flood retention-specific applications. Specifically, we propose an extension to an existing *flood-duration-frequency* model that allows for more realistic modeling of the relationship between design values of different duration at individual locations. A Bayesian inference framework for these local models is also proposed, allowing for accessible uncertainty estimation and estimation of a mixture model that helps establish the importance of the model extension. We also assess the suitability of regression-based regional flood frequency analysis models for estimating design values at multiple durations at out-of-sample locations, and offer recommendations for regional model structure if design value estimation at multiple durations is the goal.

Sammendrag

Flom påvirker flere mennesker enn nesten noen annen natur fare (Van Loenhout et al., 2020), og andelen av verdens befolkning som bor i områder utsatt for flom øker raskt (Tellman et al., 2021). En av måtene vi håndterer eksponering for flomrisiko på, er å treffe beslutninger for hvordan samfunnet kan tilpasses til flommer. Dette krever beregninger av dimensjonerende verdier for flom som brukes for dimensjonering av infrastruktur, arealplanlegging og vannressursforvaltning. En dimensjonrende flom er beskrevet ved flomstørrelser for et gitt gjentaksintervall. For anvendelser der kapasiteten til å takle flomvolum er viktig, for eksempel flomsoneforvaltning og reservoarutforming, trenger vi ofte dimensjonerende verdier for flomvolum over flere varigheter. Da trenger vi beregninger av totalt volum vi kan forvente å se over en kort varighet som en time vs. en lengre som en dag, uavhengig av om det vannvolumet kommer fra en enkelt hendelse eller flere påfølgende hendelser.

Denne avhandlingen utforsker og utvikler statistiske metoder for å beregne dimensjonerende verdier for flommer med ulike varigheter. Spesifikt foreslår vi en utvidelse av en eksisterende flom-varighet-frekvens modell som tillater en mer realistisk modellering av forholdet mellom dimensjonerende verdier med ulike varigheter på steder med vannføringsobservasjoner. Det er utviklet et Bayesiansk rammeverk for estimering av slike lokale modeller. Dette rammeverket gjør det mulig å estimere usikkerhet i dimensjonerende verdier samt å estimere en blandingsmodell som brukes for å fastslå viktigheten av modellutvidelsen. Vi vurderer også egnetheten til regresjonsbaserte regionale flom-varighet-frekvensmodeller for å estimere dimensjonerende verdier med ulik varighet på steder uten vannføringsmålinger Basert på resultatene gir vi anbefalinger for regional modellstruktur hvis målet er å estimere dimensjonerende verdier med flere varigheter.

Contents

List of Figures

List of Tables

Acknowledgements

Thank you to all my friends and colleagues that have supported and motivated me throughout this PhD and my move to Norway. I have been lucky enough to work with some truly fantastic people over these last three years. Those involved with this thesis deserve some special thanks:

To Kolbjørn, Thordis and Chong-Yu, I deeply appreciate your mentorship and unwavering academic support, especially during the early stages of this project. Thank you for all the Teams meetings, manuscript editing assistance, and good chats. I had heard of the high scientific expertise within this research group, and I can now confirm that it is excellent. Thank you too to Thomas, for welcoming me to the research group in Göttingen and providing such engaged feedback on some of the thornier statistical problems. The opportunity to work with you all has taken my statistics and hydrology knowledge to new levels. I feel so lucky to have had a team of supportive supervisors I could depend on for thoughtful guidance throughout the PhD.

To my colleagues at NVE, to participants in the ClimDesign project, and to those at the Norwegian Computing Center, thank you for making the workplace such a fun and engaging place to be. I also want to thank my fellow students and researchers in Europe and the US for being a source of motivation and lending a helping ear when needed.

And to my family: to Mike and Kristen and associated crew, thanks for providing a home and a place for me to crash every time I showed up super jetlagged from a trans-Atlantic flight. To my parents, for making the trip over here and bringing three days of sun to Bergen. And to Nick and Ryan. I'm sorry you had to settle for mopeds instead of motorcycles on your Italian vacation, but at least now we know what happens when you try to bring a plant through customs.

Chapter 1

Introduction

1.1 Motivation

Floods are one of the most widespread and costly natural hazards worldwide, and their destructive capacity is expected to rise in the near future due to climate change-induced increases in flood prevalence and growing economic value in vulnerable areas (Alfieri et al., 2017; Field, 2012). Estimation of flood design values will be essential for societal adaptation in flood prone areas. Flood frequency analysis provides a well-established statistical framework for this estimation. Many hydrologic applications where flood retention is important, e.g. floodplain management and reservoir design, need design values at several *durations*, where the duration d represents the total flow volume for a time span of d hours. In Norway, the standard approach to obtaining design values at different durations is to use daily data (data averaged over a calendar day) in a flood frequency analysis and subsequently estimate a constant scaling ratio to scale between estimates of different duration (Midtømme, 2011). This thesis investigates whether the relationship between design values of different duration can be modeled more concretely and directly. In order to do this, we develop new models and inference approaches that allow us to assess the relationship between design values of different duration.

Different modeling scenarios arise depending on whether we are building a model for a single site with sufficient data for local frequency analysis or extending estimates to ungauged or data-deficient sites, as well as if we have observed data for a specific duration or if we are extrapolating to an unobserved duration.

Existing models that can extrapolate to unobserved durations depend on a parametric relationship between return levels of different durations. These models often assume that the ratio between return levels of different duration is constant and not dependent on return period. This is a limitation of these models since we might expect floods of short durations to be more heavy-tailed than floods of longer durations. There is therefore a need for models that allow for more flexible tail behavior, as well as robust estimation frameworks for such models.

Regional models that extend design values to ungauged or data-deficient sites often base this extension on regression. We often aim to regionalize design values beyond the range of the observed data. Two approaches can be used: developing

regression models for flood quantiles or developing regression models for extreme value distribution parameters, where in both cases response values come from local frequency analysis. The choice of regression model, whether parametric or non- or semi-parametric, can significantly impact the regionalization process. There is limited existing literature comparing regression-based regional flood frequency analysis models when multiple durations are required. A systematic evaluation of the described modeling scenario is needed.

1.2 Objectives

This thesis focuses on two objectives, one for each of the modelling scenarios summarized above. For each objective, we develop statistical solutions. The statistics is always motivated by the operational potential of the approach and we develop methodology only to the extent required to address our practical issues.

The first objective is to develop local models for situations where we have sufficient data available at a single gauged site and wish to extend the flood frequency estimates to unobserved durations. For this objective the following research questions were identified:

1. Do models that allow for the ratio between return levels to change with return period improve our ability to predict unobserved sub-daily durations? (Paper I)

2. How sensitive are local models to the selected input durations? (Paper I)

The second objective is to development of regional models for situations where we have observed data at the duration of interest at a sufficient number of sites and wish to extend the flood frequency estimates at that duration to other, potentially ungauged, sites. For this objective the following research questions were identified:

3. Can a semi-parametric (i.e. "data-driven") regression model achieve comparable or improved performance to two benchmark models (one parametric and one non-parametric) on the 1 hour and/or the 24 hour duration? (Paper II)

4. Within a regional regression model, can we identify and describe duration-specific differences in how catchment covariates influence the median flood? How impactful are these differences? (Paper II)

5. How does developing regression models for flood quantiles compare to developing regression models for extreme value distribution parameters in terms of predictive performance and consistency between durations? (Paper III)

6. If our regional models produce estimates that are duration inconsistent, at what return period do we observe the inconsistent estimate? Is the return period within the range of the observed data? (Paper III)

1.3 Study design and outline

This thesis is comprised of three papers and an introduction. We employ a variety of Bayesian, frequentist, and machine learning approaches to fulfill the objectives in the thesis. Paper I addresses objective (i) and develops parametric models– and an associated Bayesian estimation framework–that allow for fully consistent estimation of return levels across durations and return periods for an at-site location with sufficient data to support flood frequency analysis. Additionally, we take an existing parametrization of the generalized extreme value distribution and introduce it in a hydrological context. Papers II and III address objective (ii) and develop models using existing non- and semi-parametric modeling approaches to investigate specific aspects of regional flood frequency analysis at multiple durations. We also introduce a preliminary machine-learning-based variable selection algorithm for regression-based regional flood frequency analysis. All papers use an aggregation-based approach to obtaining annual maxima of different duration, where annual maxima are sampled from discharge series averaged over different durations. Thus the duration d represents the total flow volume for a time span of d hours, not flood events that lasted precisely d hours. The models in paper I can be extended to any duration, although the current analysis focuses on the sub-daily durations of 1 and 12 hours. The models and analyses in papers II and III focus on the 1 hour and 24 hour durations. Papers II and III use a set of 232 gauging stations in Norway, which includes the 12 stations used in paper I as a subset.

The remainder of the thesis is structured as follows: chapter 2 provides a review of flood frequency analysis, existing approaches for obtaining flood frequency estimates at different durations, and the statistical methodology needed to estimate and evaluate the models in this thesis. Chapter 3 describes the dataset and study area, and details the data processing approach used to obtain data at different durations. Chapter 4 gives a short summary of each the three papers in this thesis with an emphasis on describing study novelty and main findings. Chapter 5 provides an overarching discussion of the work across all three papers, and chapter 6 presents a summary of the answers to the research questions and concluding statements.

Chapter 2

Scientific and methodological background

2.1 Flood frequency analysis

Design values estimate the relationship between a flood's *return level* (magnitude) and *return period* (frequency). Methods for estimating design values typically fall into one of three general categories, e.g. Filipova et al. (2019): (1) statistical flood frequency analysis (FFA), which uses flood data to estimate the magnitude of floods with specific return periods, (2) event-based hydrological modeling for a single design event, which uses design rainfall or other single realizations of initial conditions and precipitation as input to a hydrological model to simulate the desired flood event, and (3) derived flood frequency methods, which combine weather generators and hydrologic models to simulate synthetic discharge series for statistical estimation of return periods. The first approach–statistical FFA–is the focus of this thesis.

2.1.1 Extreme value theory

Flood frequency analysis usually requires us to estimate the probability of events that are more extreme than any that have been observed. The challenge is then how to statistically estimate a flood with a return period of, for example, 100 years given only a few decades of observed data. Extreme value theory provides a framework for extrapolations of this type.

Two main types of observed flood data are typically used in extreme value analyses: annual maximum series and peak-over-threshold series (Robson et al., 1999). This thesis focuses on annual maxima. Annual maxima are block maxima; that is, they are the maximum of a process over N time units of observation, where N is the number of observations in a year (Coles, 2001). Let $X_1, \ldots, X_N$ be a set of continuous, univariate random variables that are assumed to be independent and identically distributed. If the normalized distribution of the maximum $\max\{X_1, \ldots, X_N\}$ converges as $N \to \infty$ then it converges to a generalized extreme value (GEV) distribution (Fisher et al., 1928; Jenkinson, 1955). The GEV is the only possible limiting distribution of the properly normalized annual maxima and

as such it is widely applied to model annual maxima (Castro-Camilo et al., 2022). In many national guidelines for flood frequency estimation other distributions than the GEV are used (see, e.g. England Jr et al. (2019), Castellarin et al. (2012)). For example, the Log Pearson type III distribution is used in the USA, Poland, Lithuania and Slovenia, while the generalized logistic distribution is used in the UK and a mixture of two Gumbel distributions is used in Italy and Spain (Castellarin et al., 2012).

2.1.2 The generalized extreme value distribution

Extreme value theory defines three distinct representations of extreme value behavior for block maxima, each corresponding to varying tail behaviors in the distribution function of the X_i, $i \in \{1, ..., N\}$ (Coles, 2001). Classically, these are called the Gumbel, Frechet and Weibull families. The GEV distribution combines these three classes of extreme value distributions into a single family of models. In its standard location-scale parameterization, the cumulative distribution function is given as

$$G(y) = \exp\left\{-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{2.1}$$

which is defined on $\{y : 1+\xi(z-\mu)/\sigma > 0\}$ with parameter bounds $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$ and where y would be the observed annual maximum streamflow for a specific year. The shape parameter, ξ, controls the three classes of tail behavior for the GEV (Figure 2.1) (Coles, 2001). If $\xi > 0$, the limiting distribution is heavy-tailed with an infinite upper endpoint and finite, parameter-dependent lower endpoint. If $\xi < 0$, the GEV has a parameter-dependent upper endpoint. The case where $\xi = 0$ is interpreted as the limit when $\xi \to 0$ and leads to unbounded (parameter-free) support.

These endpoints may impose artificial bounds on the model when the GEV is a poor approximation to the data at hand, e.g. when the sample size is too small (Stein, 2017; Castro-Camilo et al., 2022). For this reason guidelines for FFA typically recommend assuming the shape parameter is equal to zero when the data series is short. An overview of country specific applications of the GEV for European countries can be found in Castellarin et al. (2012). Previous research (Castellarin et al., 2012; Midtømme, 2011; Kobierska et al., 2018) recommends the three-parameter GEV distribution for FFA on individual Norwegian stations with long data series.

2.1.3 The reparameterized GEV distribution

The GEV in its standard parameterization (i.e. Equation 2.1) resembles well-known location-scale families. However, while most location-scale parameterizations associate the location and scale parameters with the mean and variance, the location and scale parameters of the GEV are not easily interpreted in terms of these descriptors (Coles et al., 1996). Furthermore, the mean and variance are unsuitable for the skewed nature of the distribution and are undefined for large enough values of the shape parameter (Coles, 2001).

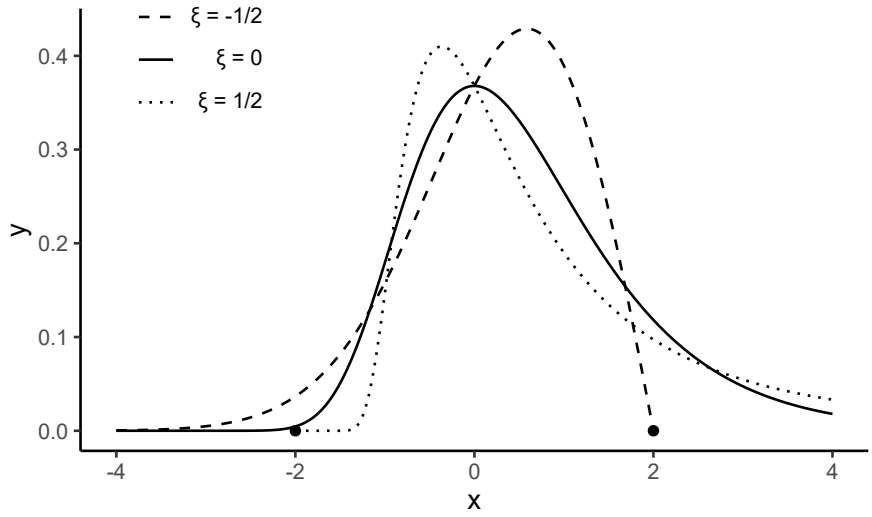

Figure 2.1: Support of GEV densities. Support endpoints are marked. All densities have $\mu = 0, \sigma = 1$.

Given these constraints, the parameters of a GEV model are more easily interpreted through quantile expressions. Castro-Camilo et al. (2022) suggest a reparameterization such that the new location parameter is the quantile defined by probability p ($0 < p < 1$). If $p = 0.5$, the relationship between the location parameter, μ, and the location parameter under the reparameterization, η, is given as

$$\eta = \begin{cases} \mu + \sigma \frac{\log(2)^{-\xi} - 1}{\xi} & \text{if } \xi \neq 0 \\ \mu - \log\{\log(2)\} & \text{if } \xi = 0. \end{cases} \tag{2.2}$$

In context of FFA, the new location parameter, η, has a reasonable interpretation as the median annual maximum flood, with units that match the original streamflow time series.

2.1.4 Flood frequency curves

Estimates of the quantiles of the GEV distribution give us the desired estimates of low-probability, large-magnitude floods. Using the GEV model under the reparameterization, quantiles are obtained by substituting η from Equation 2.2 for μ in Equation 2.1 and inverting the result:

$$z_p = \begin{cases} \eta - \frac{\sigma}{\xi}\left[(-\log(1-p))^{-\xi} - \log(2)^{-\xi}\right] & \text{if } \xi \neq 0 \\ \eta - \sigma \log\{-\log(1-p)\} + \log\{\log(2)\} & \text{if } \xi = 0. \end{cases} \tag{2.3}$$

where $G(z_p) = 1 - p$ and z_p is the return level associated with the return period T such that $T = 1/p$. That is, the expression in Equation 2.3 defines the *flood frequency curve* relating flood size to flood rarity (Robson et al., 1999).

If we define $\ell_p = -\log(1-p)$, then the plot of z_p against $\log \ell_p$ forms the *return level plot* (Coles, 2001). The logarithmic scale in the return level plot condenses the distribution's tail, emphasizing the impact of extrapolation and visually revealing

the type of tail behavior exhibited by the underlying GEV distribution. The return level plot is linear in the case $\xi = 0$. If $\xi < 0$ the flood frequency curve is bounded above (i.e. has a maximum possible value) and the return level plot is convex. If $\xi > 0$ the flood frequency curve is unbounded above and the return level plot is concave.

It is often useful to scale the flood frequency curve such that the two-year return period has a return level of 1 (that is, scale the flood frequency curve by dividing by the median flood) (Robson et al., 1999). Scaled versions of the flood frequency curve are typically called *growth curves*. The scaling factor–the median flood in this case–is typically called the *index flood*. This scaling is beneficial for hydrologic analyses because separating the order of magnitude of a flood from the shape and slope of the growth curve allows for cross-catchment comparisons (Robson et al., 1999; Dalrymple, 1960). Under the reparameterization in Equation 2.2, this scaling is straightforward as the median is explicitly included as a parameter in the flood frequency curve (Equation 2.3).

2.1.5 Approaches for frequency analysis at ungauged catchments

Flood frequency analysis is often required at ungauged sites or sites with insufficient data. It is then necessary to use data from nearby or similar gauged stations to estimate flood quantiles at the site of interest. This is termed regional flood frequency analysis (RFFA). The process of transferring information from hydrologically similar catchments to a particular catchment of interest is called *regionalization* (Blöschl et al., 1995). There are many different approaches to regionalization; see Blöschl (2013) or Odry et al. (2017) for a review. The focus in this thesis is on regionalization via regression models that define a functional relationship between a specific flood quantile and appropriate predictors, i.e. catchment or climatic characteristics.

RFFA typically takes one of two paths (Fischer et al., 2021): (i) regionalize flood quantiles of given return periods (in terms of Equation 2.3, this would mean regionalizing z_p directly) or (ii) regionalize parameters of the extreme value distribution (in terms of Equation 2.3, regionalizing the GEV parameters η, σ and ξ). The well-known index flood method (Dalrymple, 1960) is a special case of (ii), where a regression model is established for the index flood and the other distributional parameters are assumed constant for all sites in a region.

Index flood methods

The classic approach to RFFA is the *index flood method* (Dalrymple, 1960; Hosking et al., 1988). The index flood method assumes that the flood frequency curve follows the same distribution up to a site-specific scaling factor for all locations in a region. The regions are defined to have similar flood generating processes. This site-specific scaling factor is the index flood, i.e. a "typical"–usually the mean or median–flood for a particular catchment. There are three separate steps to the index flood method: (1) identification of homogenous regions or hydrologically

similar stations (2) estimation of the index flood (3) derivation of a growth curve that gives the relationship by which we can scale an index flood to a desired return level.

If we have no data at a specified site, the index flood is derived from climatic and catchment properties or based on appropriately scaled nearby measurement stations. Then a regional growth curve is applied to the estimated index flood. The regional growth curve is typically the average of all the individual growth curves from sites in the identified region (Dalrymple, 1960). A variety of methods can be used to derive the index flood from climatic and catchment properties (Blöschl, 2013; Farquharson et al., 1992). Regression models are commonly used but geostatistical or process-based methods can also be applied (Bocchiola et al., 2003).

Although regression can be used to estimate the index flood, the index flood method itself is distinct from regression-based RFFA. Regression-based RFFA models attempt to capture the continuous variability of flood frequency curve characteristics across space and/or climatic and catchment characteristics, while the index flood method groups catchments that are assumed to share the same distributional assumptions.

Regression methods

Regression models for RFFA are based on the assumption that spatial variations in flood statistics are closely linked with regional catchment and climate characteristics (Robson et al., 1999). The relationship between flood statistics and catchment descriptors is likely nonlinear (Pandey et al., 1999; Tarquis et al., 2011), mainly due to the inherent nonlinearity in hydrological processes (Durocher et al., 2015). This nonlinearity stems from various factors, such as the non-linear response of runoff to rainfall and snowmelt (Gioia et al., 2012), non-linear snowmelt processes, and the influence of both events and catchment characteristics on flood generation. Traditionally, to handle this nonlinearity, the predictors are transformed such that they have a linear relationship with flood statistics and a linear or log-linear regression model is fit. This is the approach used in the current regional median flood model for Norway (Engeland et al., 2020).

Regression models in RFFA can be either parametric or non- or semi-parametric. Parametric models include linear, log-linear, nonlinear, and generalized linear models (Cunnane, 1988; Griffis et al., 2007; Pandey et al., 1999; Thorarinsdottir et al., 2018; Clarke, 2001). Non- or semi-parametric models include generalized additive models, artificial neural networks, random forests, boosted or bagged tree ensembles, and support vector machines (Chebana et al., 2014; Shu et al., 2008; Aziz et al., 2014; Desai et al., 2021; Esmaeili-Gisavandani et al., 2023; Wang et al., 2015; Allahbakhshian-Farsani et al., 2020; Laimighofer et al., 2022a; Gizaw et al., 2016). Parametric models rely on a parametric description of what is called the *functional form* between predictors and response. For example, we assume the relationship between the median flood and the mean temperature in February is completely described by the functional form x^2. In this situation, we would need to estimate a regression coefficient—that is, we would need to estimate the magnitude and direction of this functional form—but the underlying

relationship will always be described by the square function. Parametric models are easy to interpret and estimate and therefore widely used in RFFA (Blöschl, 2013). However, choosing the wrong functional form can introduce significant bias to the results. Consequently, a large part of the work required when using a parametric model revolves around identifying appropriate polynomial terms and predictor transformations (Guisan et al., 2002).

Non- and semi-parametric regression models are termed "data-driven" because the data determine the functional form relating the response to catchment descriptors, rather than imposing a predefined parametric relationship through model structure (Yee et al., 1991). Typically, we use a data-driven model when a complex, nonlinear relationship between the predictors and response needs to be established; however, if the true relationship is linear, a data-driven model will recover the linear relationship (Härdle, 1990; Hastie et al., 1987). In this sense data-driven models offers a potential simplification of the modeling process as we no longer need to identify appropriate predictor transformations, although other aspects of the modeling process (e.g. model selection and diagnostics) must still be treated carefully (Härdle, 1990).

Here, we define a data-driven model as semi-parametric if we must specify the probability distribution for the response. For instance, a generalized additive model (GAM) modeling flood event distribution as log-normal falls into this category (see, e.g. Barna et al. (2023b) or Chebana et al. (2014)). In contrast, a non-parametric model, like a boosted tree ensemble (Laimighofer et al., 2022a; Jarajapu et al., 2022) or artificial neural net (Aziz et al., 2014), does not make explicit assumptions about the underlying distribution of flood events. Choosing between a semi- and non-parametric model is guided by application. In the semi-parametric framework, interpretation and inference are generally more straightforward thanks to the underlying statistical assumptions (Hastie et al., 1987). We can perform reliability analyses and calculate performance metrics that might not be feasible with a non-parametric model. On the other hand, non-parametric models can have higher predictive accuracy (Mosavi et al., 2018). Additionally, they can handle predictor sets that might be problematic for semi-parametric models. For instance, boosted tree ensembles can effectively manage correlated or irrelevant predictors within a large set of potentially important predictors (Elith et al., 2008), a common scenario in hydrologic applications (Galelli et al., 2013).

2.2 Treatment of different flood durations

When considering multiple durations, we typically want estimates that are *consistent* across durations. Here, consistent means we want the estimated values to be consistent with physical realities: we do not want, for example, an estimate that says more water will arrive in a 1-hour window than in a 24-hour window if the latter time period encompasses the former.

Enforcing this consistency between durations in flood frequency analysis is challenging from a statistical perspective. Flood frequency models are extreme value models that are, by necessity, based on extrapolation. Independently

estimating flood frequency curves for each duration of interest is not guaranteed to give consistent results. Even minor deviations in model specifications or slight variations in the parameters of the extreme value distribution across different durations may be magnified on extrapolation and result in inconsistent estimates.

This section reviews the statistical methodologies developed to address the problem of consistency across durations for applications where the total volume of water is of interest. The focus on flood-retention specific applications means that a duration d represents the total flow volume for a time span of d hours, not individual flood events that lasted precisely d hours. Consistency across durations is generally enforced in one of two ways:

(i) Models that simultaneously estimate several durations and quantiles at once under consistency constraints.

(ii) Post-processing of model output that has been independently estimated at several durations.

The current NVE guidelines (Engeland et al., 2020) are a special case of (i) that focus on two specific durations (the instantaneous duration and a duration corresponding to the total flow volume observed over a calendar day). The specification of a total flow volume over a calendar day distinguish these guidelines from the methods presented here in this chapter.

Within option (i) efforts are often focused on ways of making the consistency constraints more realistic. One way of doing this is to attempt to explicitly model the dependence structure between annual maxima at different durations. Modeling this dependency has been an active area of research within precipitation applications that also assess volume over specific durations, i.e. intensity-duration-frequency (IDF) models. Examples of this dependency modeling would be, for example, Jurado et al. (2020), Tyralis et al. (2019), and Muller et al. (2008). An important distinction needs to be made here between the approaches listed here and bivariate frequency analyses, which rely on identification of individual flood events and explicitly model the dependence structure between peak discharge and event duration (see, for example, the copula models of Gräler et al. (2013)). Generally, the assessment of the total flow volume over a specific time window requires aggregation of flood events. Event-based methodology for bivariate frequency analysis is typically not appropriate here.

Another way of making the consistency constraints in option (i) more realistic is to structure the constraints such that they produce a certain type of tail behavior. This does not require explicit modeling of dependencies between durations. Most recent work in this area has centered around development of so-called "multiscaling" models, which build the consistency constraints such that the slope and the intercept of the flood frequency curve can scale independently– i.e., the model allows for the ratio between growth curves of different durations to be dependent on return period. Multiscaling has been implemented for IDF models (Van de Vyver, 2018; Courty et al., 2019; Fauer et al., 2021) and Barna et al. (2023a) adapts this to QDF models. Typically, these multiscaling models are "empirical multiscaling" models, i.e. they do not attempt to place strict

mathematical assumptions on how the variance or other higher-order moments change with increasing duration. Strictly theoretical multiscaling models would be, for example, those presented in Gupta et al. (1990) or Van de Vyver (2018).

2.2.1 Flood-Duration-Frequency (QDF) models

Flood–duration–frequency (QDF) models are a type of extreme value model that simultaneously estimate return levels for several durations under consistency constraints. Typically the underlying extreme value distribution is assumed to be GEV. QDF models are a type of *dependent GEV*, or d-GEV model, and are analogous to intensity–duration–frequency (IDF) models for precipitation.

The foundations of QDF modeling were developed in the 1990s through analyses of n-day flood volumes as explored in Balocki et al. (1994) and Sherwood (1994). The original QDF model is attributed to Javelle et al. (1999). Historically, application of QDF models has been concentrated in France, Canada, and Britain in the early 2000s (Javelle et al., 2002; Javelle et al., 2003; Zaidman et al., 2003; Cunderlik et al., 2006; Crochet, 2012; Onyutha et al., 2015). More recent applications of QDF models can be found in Renima et al. (2018), Markiewicz (2021), and Breinl et al. (2021). An extended QDF model that relaxes some of the underlying assumptions in the original model can be found in Barna et al. (2023a).

Local QDF model

Let $\mathbf{y}_d$ be a vector of annual maxima at duration d with index $i \in [1, ..., n]$ referring to the ith element. Then the $\mathbf{y}_d$ under the original QDF model proposed in Javelle et al. (2002) are independently distributed

$$y_{d,i} \sim \text{GEV}\left(\eta_d, \beta, \xi\right) \tag{2.4}$$

where

$$\eta_d = \eta\left(1 + d/\Delta\right)^{-1} \tag{2.5}$$

and η is the location parameter of the GEV under the quantile based reparameterization proposed in Castro-Camilo et al. (2022). The parameter β is given as

$$\beta = \log\left(\frac{\sigma}{\eta}\right). \tag{2.6}$$

Then the parametric relationship between quantiles of different duration is given as

$$z_{d,p} = \frac{\eta}{1 + d/\Delta}\left[1 + e^{\beta}\left\{\frac{\left(-\log(1-p)\right)^{-\xi} - \log(2)^{-\xi}}{\xi}\right\}\right] \tag{2.7}$$

where $\Delta > 0$. A value close to zero for Δ indicates the total flow volume arrives quickly, analogous to a flashy/peaked hydrograph with a pronounced duration dependency for the median flood, whereas a high value of Δ indicates a slower

arrival of the total flow volume, analogous to a wide hydrograph with minor duration dependency for the floods. The traditional flood frequency curve–that is, a GEV distribution fit to an instantaneous time series–is recovered in the limit of the aggregation window as $d \to 0$.

In the original QDF model only η is dependent on d and Δ. Since only the magnitude of the median flood (η) is duration-dependent in the model in Equation 2.7, the underlying assumption of the original QDF model is that the slope and shape of the growth curve does not change with duration. That is, the original QDF model assumes the ratio between floods of different duration is independent of return period.

Regional QDF model

The regional QDF model as presented in Javelle et al. (2002) is an extension of the index flood approach. It includes an additional parameter that quantifies how the index flood changes with duration. This additional parameter is the characteristic duration parameter, Δ. The steps for the regional QDF model are: (1) identification of homogenous regions or hydrologically similar stations (2) estimation of the index flood (3) estimation of the characteristic duration parameter (4) derivation of a growth curve for the instantaneous duration that gives the relationship by which we can scale an index flood and characteristic duration parameter to a desired return level and specified duration. To estimate Δ in step (3), we fit local QDF models to each site in the region. These local QDF models are then extrapolated to the instantaneous duration (i.e., $d = 0$). The regional growth curve is then an aggregate of the at-site QDF estimates of the instantaneous duration.

To use the regional QDF model at a site with no data, we construct regional relationships for both Δ and the index flood such that they can be estimated by climatic and catchment properties. In Javelle et al. (2002), this is achieved with linear regression. Then, at a target site $s \in S$, the set of all stations in the data set, the regional instantaneous growth curve can be unscaled with the site-specific index flood and site-specific characteristic duration parameter:

$$z_{s,d,p} = \lambda(p) \cdot \eta_s \cdot (1 + d/\Delta_s)^{-1} \tag{2.8}$$

where $z_{s,d,p}$ is the return level at target site s, duration d and probability p corresponding to return period T such that $T = 1/p$. Here $\lambda(p)$ is the regional instantaneous growth curve. The site-specific index (median) flood η_s and site-specific characteristic duration parameter Δ_s are specified by linear models that take relevant climatic and catchment descriptors as predictors.

The regional QDF model carries assumptions from both the index flood method and the original local QDF model. Specifically, like the local QDF model, it assumes that the shape and slope of the growth curve do not change with duration (i.e., the growth curve follows the same distribution up to a duration-specific scaling factor for all durations of interest). Additionally, like the index flood method, it assumes that the growth curve follows the same distribution up to site-specific scaling factor for all sites within a homogeneous region. If a

parametric regression model is used to estimate η_s and Δ_s, the regional QDF model additionally assumes that, although the magnitude of the functional form may change when regression coefficients are scaled for different durations, the relationship between catchment descriptors and the site-specific parameter will always be described by the same functional form regardless of the duration being considered.

2.2.2 Post-processing of return levels

Post-processing allows us to independently estimate extreme value models at each duration of interest, and thereafter, in a separate step, adjust the resulting return levels such that they are duration consistent. Let $z_p^{d_i}$ be the return level at probability p from an extreme value model fit to data at observed duration d_i, $i \in \{1, \ldots, D\}$. Here the superscript distinguishes the independent fit at observed duration d_i from a QDF model evaluated at target duration d. If the return levels at a specific duration stem from a single extreme value distribution they are necessarily monotonically increasing for increasing return period T, where $T = 1/p$. Duration consistency at probability level p is then defined as

$$z_p^{d_i} \leq z_p^{d_j} \text{ for } d_i \geq d_j \tag{2.9}$$

for any $i, j \in \{1, \ldots, D\}$; that is, the return levels, when modeled with units of volume over time, should be monotonically increasing as the aggregation interval narrows.

Choice of a post-processing method that enforces duration consistency often depends on the particular model architecture and inference approach used to fit the extreme value model. If a Bayesian inference approach is used, the quantile selection algorithm proposed in Roksvåg et al. (2021) is an option; the algorithm searches for consistent return levels within the quantiles of the full posterior distributions of the return levels. The selection of posterior quantiles adjusts the flood frequency curves subject to consistency across durations and return levels. Alternatively, an option that is not dependent on inference method is adjustment via isotonic regression, also discussed in Roksvåg et al. (2021) and implemented in Meyer (2013). Here point estimates for return levels are adjusted according to the isotonic regression such that the resulting set of return levels is consistent across durations and return periods. While the isotonic regression approach can be used with any inference approach, it does not account for the uncertainty information inherent to the fit of the extreme value distribution and can yield a flood frequency curve that falls outside the support of the underlying distribution.

2.3 Model inference

Model inference is the process of estimating a statistical model and assessing uncertainty in the fit. Mathematically, there are different ways to structure the inference process. We could, for example, adopt a frequentist approach to inference. This requires us to treat probability as the long-run relative

frequency of an event. This approach gives us access to well-known estimation methodologies such as unpenalized maximum likelihood and special cases thereof, e.g. least squares and generalized least squares; these methodologies generate point estimates for the quantities of interest (Gelman et al., 2020). Alternatively, we could adopt a Bayesian approach to inference and treat probability as the relative plausibility of an event. This approach allows us to define a probabilistic expression of prior information and posterior uncertainty, from which it is possible to estimate probability distributions for the quantities of interest (Gelman et al., 2013).

Inference approaches in FFA have traditionally been frequentist (Hu et al., 2020). It is common to use method of moments, both ordinary moments and linear moments/probability weighted moments, or maximum likelihood estimations to obtain parameter estimates (Kobierska et al., 2018; Renard et al., 2013; Ball et al., 2019; England Jr et al., 2019; Castellarin et al., 2012; Robson et al., 1999). Inference approaches for QDF models have also traditionally been frequentist. Javelle et al. (2002) proposes a two-step estimation method, where the characteristic duration parameter is estimated as the value that minimizes the dispersion of the time scaled values of the growth curves of different duration, and then the remaining model parameters are estimated using the method of probability weighted moments (Hosking et al., 1985).

Choice of the particular mathematical structure for inference is driven by model architecture and practical application. For complex model architectures—e.g. the generalized additive models (GAMs) and extreme gradient boosted tree ensembles (XGBoost) used in paper II—it is often practical to use existing implementations (Wood, 2017; Chen et al., 2015) of inference approaches. In other cases, we have a predetermined preference for a specific inference approach. This is the case for paper I, which developed a Bayesian approach to QDF modeling. Bayesian inference for extreme value models is relatively new (Coles, 2001) and implementations are often less widely available than those for frequentist approaches. As such, implementation of the inference approach is often an important component of a Bayesian extreme value analysis. This section provides a review of the Bayesian inferential framework.

2.3.1 The Bayesian framework

The goal of Bayesian inference is to combine an initial set of beliefs about the data generating process (i.e. *prior information*) with a model of the data such that we can make conclusions that are consistent with both sources of information (Hoff, 2009). *Bayes' rule* provides a mathematical structure for this combination. Let $\mathcal{Y}$ be the set of all possible datasets, from which our observed dataset $\mathbf{y}$ will result. Let Θ be the set of possible parameter values, from which we hope to identify parameters $\boldsymbol{\theta}$ that best describe the conditions that generate our data. Then Bayes' rule expresses our joint beliefs about $\mathbf{y}$ and $\boldsymbol{\theta}$ in terms of probability distributions over $\mathcal{Y}$ and Θ. Specifically, the prior distribution $\pi(\boldsymbol{\theta})$ describes our beliefs about the behavior of $\boldsymbol{\theta}$ independent of the observed data. The sampling distribution, $\pi(\mathbf{y}|\boldsymbol{\theta})$, describes our belief that we would observe $\mathbf{y}$ if we knew $\boldsymbol{\theta}$

to be true. The prior distribution and the sampling distribution are combined to yield the posterior distribution:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.10}$$

which describes our belief that the true value is given by $\boldsymbol{\theta}$, having observed dataset $\mathbf{y}$. Here $\propto$ means "proportional to". To obtain inferences for any particular element $\theta_i \in \boldsymbol{\theta}$ we integrate out the remaining components to reach the posterior distribution $\pi(\theta_i|\mathbf{y})$.

Certain aspects of the Bayesian inferential framework have the potential to be advantageous for extreme value analyses (Coles et al., 1996; Coles, 2001). First and foremost, the natural scarcity of extreme value data means incorporating prior information about the parameters is appealing. Incorporating a prior has the potential to steer the model away from less-favored values of the parameters, leading to more stable estimation (Martins et al., 2000). Second, extreme value analyses commonly need predictions, i.e. we need to estimate the probability of future events reaching extreme levels. Predictive inference is practical in a Bayesian framework because we can express a prediction and its associated uncertainty through a full predictive distribution (Gelman et al., 2013).

2.3.2 Markov chain Monte Carlo methods for parameter estimation

Analytically computing posterior quantities of interest from Eqn. 2.10 can be challenging or impossible. Obtaining exact values involves integrating the joint distribution $\pi(\mathbf{y}, \boldsymbol{\theta})$ which can be infeasible if the dimension of $\boldsymbol{\theta}$ is large. However, we can approximate the posterior quantities of interest using simulation methods like *Markov chain Monte Carlo* (MCMC), which allows us to sample from complex probability distributions by constructing a Markov chain whose stationary distribution matches the desired target distribution (Robert et al., 1999). MCMC methods are a versatile class of approaches with general principles and procedures that are applicable to a wide range of problems, facilitating estimation of complex models that are challenging to tackle using alternative methods. For instance, using a reversible jump MCMC sampler (Richardson et al., 1997) we can sample mixture representations with an unknown and varying number of components, e.g. the mixture model in paper I.

MCMC methods can be computationally expensive, particularly when dealing with complex or high-dimensional target distributions (Robert et al., 1999). However, there are various techniques available to mitigate this computational burden. For example, adaptive proposal mechanisms can be employed to improve the efficiency of the sampling process. Or, for instance, gradient-based MCMC methods, such as Hamiltonian Monte Carlo (HMC), can be used to take advantage of gradient information from the target distribution to guide the sampling process more effectively (Gelman et al., 2015). Many of these techniques are implemented in user-friendly software packages (e.g. `BUGS` by Lunn et al. (2000), and `Rstan` by Stan Development Team (2023)), especially for common models and distributions.

Specific probability distributions like GEV or generalized Pareto used in extreme value analysis may require custom solutions.

2.4 Model evaluation

Typically we care about the predictive performance of our models. The way we evaluate this predictive performance depends partly on the type of value we are predicting. When we are predicting a single value (a point estimate), we employ a scoring function, such as the absolute error, to measure how close it is to the observed value. However, if we have access to the complete predictive distribution, our model evaluation can be probabilistic in nature and we can compare the probability distribution of model output to the corresponding empirical distribution of observed data. In most cases, the second option provides a more comprehensive assessment because it allows us to consider uncertainty in the prediction (Thorarinsdottir et al., 2013; Gneiting, 2008). Nevertheless, practical scenarios often necessitate point estimates, and a number of inference methods are better suited for generating such point estimates (Gneiting, 2011). This section provides a review of scoring functions, as well as one probabilistic evaluation method (the integrated quadratic distance, or IQD).

2.4.1 Scoring functions

A scoring function S depends both on the predicted value and observed data. Following the notation in Gneiting (2011), we average the scoring function over the observed cases to generate the performance criterion:

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} S(\hat{y}_i, y_i) \tag{2.11}$$

where there are k predicted cases with corresponding point estimates $\hat{y}_i, \ldots, \hat{y}_k$ and verifying observations $y_1, \ldots, y_k$. The scoring function is typically negatively oriented, i.e. a smaller value indicates better predictive performance.

Different scoring functions measure different aspects of the predictive distribution. When the observed data set is small–as is often the case in extreme value analyses–it is often prudent to assess the predictions on multiple scoring functions (Coles, 2001). For instance, when the goal is evaluation of predicted return levels, scoring functions expressed in terms of specific discharge units, such as root mean squared error or absolute error, may prioritize minimizing errors in regions with the highest discharge values (Engeland et al., 2020). On the other hand, scoring functions that consider relative differences, like mean absolute percent error or mean relative error, can mitigate the scaling issue but may be sensitive to extreme over- or under-estimations (Gneiting, 2011).

2.4.2 Integrated quadratic distance (IQD)

The IQD measures the similarity between two distributions by integrating over the squared distance between the distribution functions (Thorarinsdottir et al.,

2013). Let F_1 and F_2 be two univariate distribution functions. In practice we approximate F_1 and F_2 by the empirical CDF of a sample from the posterior. The distance between F_1 and F_2 as measured by the IQD is then given by

$$\mathrm{IQD} = \int_{-\infty}^{+\infty} \left(F_1(y) - F_2(y)\right)^2 \mathrm{d}y \tag{2.12}$$

where lower values of the IQD indicate better overall performance. The IQD is the score divergence associated with the well-known proper scoring rule the continuous ranked probability score (CRPS) (Gneiting et al., 2007; Hersbach, 2000); the main difference between IQD and CRPS is that CRPS calculates the integrated squared distance between a distribution and a scalar observation specified by a Heaviside step function whereas IQD calculates the integrated squared distance between two distributions (Thorarinsdottir et al., 2013).

2.4.3 Significance of scores

The permutation test, as defined in Thorarinsdottir et al. (2020), determines the difference in performance between two models, m^1 and m^2, by computing

$$c = \frac{1}{k} \sum_{i=1}^{k} \left(\phi(m_i^1) - \phi(m_i^2)\right) \tag{2.13}$$

Here, k represents the total number of predicted cases and $\phi(\cdot)$ is a divergence, or distance, function. When comparing a point estimate and verifying observation, $\phi(\cdot)$ can be a scoring function. When comparing probability distributions, $\phi(\cdot)$ can be the IQD (see, e.g., Thorarinsdottir et al. (2013)). If c is negative, it indicates that model m^1 performs better than model m^2 in terms of the distance function, and vice versa. The permutation test creates resampled copies of c with randomly swapped models m^1 and m^2. Under the null hypothesis that both models perform equally well, the set of permutations cannot be differentiated from c. The statistical test formalizes this concept by determining which quantile c occupies in the set of permutations; if the p-value is less than 0.05, then it suggests that the performance of m^1 is significantly better than m^2.

Chapter 3

Study area and data

3.1 Norwegian climatology and catchments

The diverse topography and wide range of latitudes in Norway create significant variation in temperature and precipitation patterns throughout the region. As a result, Norway exhibits a large variety of hydrological regimes (Gottschalk et al., 1979); see Figure 3.2. In this figure rain was computed using precipitation and temperature data from the SeNorge 2.0 dataset (Lussana et al., 2019), while snowmelt was derived from the SeNorge snow model (Saloranta, 2014).

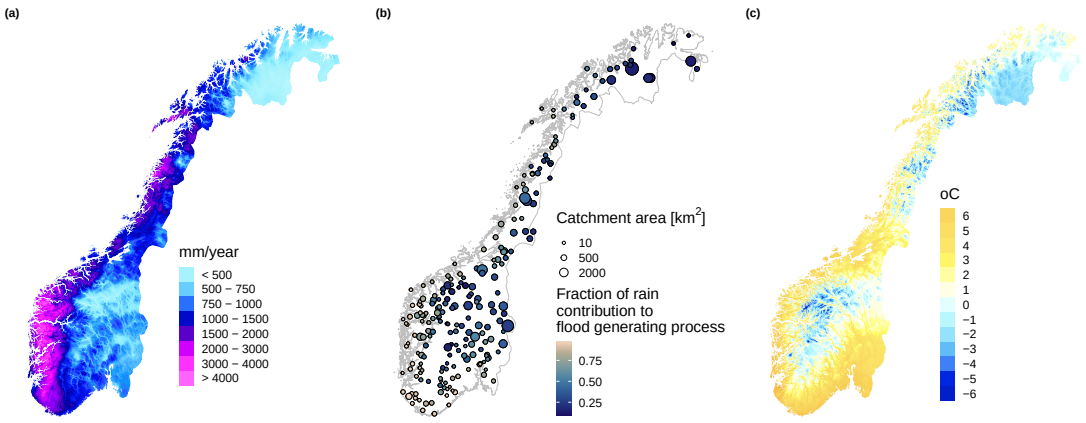

Figure 3.2: Panel (a) shows average precipitation totals (mm) for the entire year from the period 1991-2020. Panel (b) shows locations of the 232 gauging stations used in this thesis, where catchment area and average fraction of rain contribution to flood are indicated by size and color, respectively. Panel (c) shows average temperature (°C) for the entire year from the period 1991-2020.

The two major flood generating processes in Norway are snowmelt and rainfall (Gottschalk et al., 1979). The regional importance of snowmelt as a runoff generating process varies greatly across the country. The regions primarily driven by snowmelt are the inland and northern regions, which predominantly experience high flows during spring and summer. Western and coastal regions are primarily driven by rainfall and experience high flows during autumn and winter. However, local climate and mixed or transitional flood regimes introduce

substantial variability to these regional trends, and seasonal patterns are not very distinct in rainfall-driven catchments (Vormoor et al., 2016).

The regional models in this thesis use a set of 232 gauging stations deemed suitable for flood frequency analysis. See Engeland et al. (2020) for a discussion of Norwegian gauging stations suitable for frequency analyses. The selected stations exhibit a diversity of hydro-climatic regimes relative to Nordic catchments. The 12 stations used the at-site QDF analyses in paper I are a subset of the 232 stations used for regional analysis in papers II and III.

3.2 Catchment descriptors

Geographical and hydro-climatic catchment descriptors available for the selected stations are listed in Table 3.1. Note that all hydro-climatic descriptors were based on interpolated observations given by the SeNorge 2.0 dataset (Lussana et al., 2019). Rain was defined as precipitation when the temperature is above 0 °C. Snow melt was extracted from the SeNorge snow model (Saloranta, 2014).

Table 3.1: Descriptions of all catchment descriptors used for regional modeling, grouped into geographical and hydro-climatic descriptors. Abbreviations are further used in the text and figures.

Variable	Description	Unit
A	Logarithm of catchment area	km^2
O	Catchment circumference	m
A_P	Catchment area / circumference * 1000	km
D, D_{net}	Drainage density (total river length / area), (total river length excluding lakes / area)	-
C_L	Logarithm of catchment length	km
R_L	Length of main river	km
R_{TL}, $R_{TL,net}$	Total river length, and total river length excluding lakes	km
R_G, R_{G1085}	Gradient of main river, and gradient of main river excluding the 10 % lowest and 15 % highest reaches	m/km
H_{10}, H_{50}, H_{90}, H_{MAX}, H_{MIN}	The 10th, 50th, and 90th percentile of the hypsographic curve, maximum elevation, minimum elevation	m.a.s.l.
H_F	Catchment relief (maximum elevation - minimum elevation)	m
C_S	Mean slope	°
A_{Glac}, A_{Agr}, A_{Bog}, A_U, A_L, A_{LE}, A_{For}, A_{Mount}	Percentage of catchment covered by glaciers, agriculture, bogs, urban areas, lakes, effective lake percentage, forests, mountains	%
Q_N	Mean annual runoff 1961-1990	$l/s/km^2$
P_{Jan}, P_{Feb}, P_{Mar}, P_{Apr}, P_{Mai}, P_{Jun}, P_{Jul}, P_{Aug}, P_{Sep}, P_{Oct}, P_{Nov}, P_{Dec}	Mean precipitation from 1961-1990 in January, February, March, April, May, June, July, August, September, October, November, December	mm/month
P_N	Mean annual precipitation 1961-1990	mm/year
$P_{Med1Max}$, $P_{Med2Max}$, $P_{Med3Max}$, $P_{Med4Max}$, $P_{Med5Max}$	Median of annual 1-, 2-, 3-, 4-, and 5-day precipitation	mm/day
T_{Jan}, T_{Feb}, T_{Mar}, T_{Apr}, T_{Mai}, T_{Jun}, T_{Jul}, T_{Aug}, T_{Sep}, T_{Oct}, T_{Nov}, T_{Dec}	Mean temperature from 1961-1990 in January, February, March, April, May, June, July, August, September, October, November, December	°C
T_N	Mean annual temperature 1961-1990	°C
W_{Jan}, W_{Feb}, W_{Mar}, W_{Apr}, W_{Mai}, W_{Jun}, W_{Jul}, W_{Aug}, W_{Sep}, W_{Oct}, W_{Nov}, W_{Dec}	Mean sum of rainfall and snowmelt from 1961-1990 in January, February, March, April, May, June, July, August, September, October, November, December	mm/month
W_N	Mean annual sum of rainfall and snowmelt 1961-1990	mm/year
$W_{Med1Max}$, $W_{Med2Max}$, $W_{Med3Max}$, $W_{Med4Max}$, $W_{Med5Max}$	Median of annual 1-, 2-, 3-, 4-, and 5-day rainfall and snowmelt	mm/day

Catchment areas range from 0.52 km^2 to 6182 km^2, with a median of 124 km^2. About 53% of the catchments have over 1% of their area covered by lakes, with a median effective lake coverage of 2.8%. Mean annual precipitation varies from 390 mm to 3196 mm, showing an east-west gradient, with higher levels along the west coast. Mean annual temperature spans from -4.0 °C to 7.2 °C, with a median of 0.15 °C. Temperature is influenced by elevation and latitude, decreasing as

elevation and latitude increase. Catchment altitudes range from sea level to 1104 m.a.s.l., and relief varies from 54 m to 2019 m. Catchments with greater relief are typically located in the rugged mountain ranges along the west coast, contrasting with flatter regions in the east.

Latitude and longitude are excluded as catchment descriptors. Paper II focuses on identifying and describing hydrologically significant relationships at various durations. Given this explanatory approach, we prefer to explain spatial variations in flood sizes using geographical and hydro-climatic descriptors other than latitude and longitude. Catchments that are close in space as the crow flies may be very different in character, or quite distinct topographically (for example, if the catchments are divided by a high mountain, as is often the case in the western region of Norway) (Sælthun et al., 1997). Selecting a single latitude longitude point to represent the entire area of a catchment is also non-trivial. Given these considerations, we chose to address the question of spatial dependence through the regionalization study in Paper II, where we examine if splitting the area of study into pre-determined regions improves modelling performance. We found splitting by regions did not meaningfully improve model performance, which does not provide strong evidence for adding latitude / longitude terms to the regional models. Other types of models that explicitly borrow information from the spatial structure of the observations—e.g. geostatistical spatial models (such as those in Merz et al. (2008))—would be an option for further investigating spatial or regional dependence but are distinct from the marginal models for extremes developed in this thesis.

3.3 Streamflow data

The observed streamflow time series were obtained from the national hydrological database Hydra II hosted by the Norwegian Water Resources and Energy Directorate (NVE). The streamflow records have at least 20 years of quality controlled data for periods with minimal influence from river regulations and a sufficient quality for high streamflows; see Engeland et al. (2016) for details.

The data have undergone rigorous quality control conducted by the hydrometric section at NVE. Ice jams are an issue at numerous Norwegian stations and can impact the accuracy of rating curves used for estimating streamflows from water level measurements. In such instances, NVE's internal quality assurance protocols were applied to ensure precise discharge values. Furthermore, years with less than 300 days of data were excluded from the analysis.

The streamflow records are comprised of various collection methods, resulting in data of different frequencies. Generally, earlier parts of the records offer daily time resolution, while later segments feature higher (subdaily) measurement frequencies, particularly after the transition to digitized limnigraph records and digital measurements around 1980. The frequency of the subdaily measurements varies from catchment to catchment. The measurement frequency at each catchment was chosen by NVE to accurately represent flood peaks at individual stations. This choice follows internal quality assurance protocols and is dictated

by catchment properties: a higher frequency of measurements is needed to capture the behavior of quicker, "flashier" floods vs slower, smoother floods.

The regularity of subdaily measurements varies based on the source: the digital measuring stations are set up to have a regular measuring frequency, with approximately half of the digital stations in our study measuring every 60 minutes and the rest at 30 minutes or less. In contrast, the digitized limnigraph records do not have a consistent frequency. The digitization process captures a higher resolution of data around flood peaks and a coarser resolution of data in periods with minimal flood activity. For stations with both limnigraph and digital measurement periods, we assessed the resolution of the digitized limnigraph records at flood peaks and found it was comparable to the digital measurement frequency.

Given the focus on sub-daily floods in the thesis, it is necessary to ensure that the sampling frequency of the data adequately represents peak flood magnitudes. In Paper I, which centers on a case study of twelve stations, this is achieved through manual verification of the station-specific annual maxima. Papers II and III rely on a larger dataset of 232 stations and construct regression relationships for the median annual maximum flood (as opposed to individual annual maxima as used in Paper I). The median annual maximum flood at both durations studied (1 and 24 hours) is computed over the total number of years of data available at each station. This means that for certain stations, especially those with longer record lengths, the median is constructed from annual maxima derived from streamflow time series at a combination of different resolutions. In the case of the 1 hour median annual maximum flood, this means the 1 hour median may partly rely on data at a coarser time resolution interpolated to 1 hour spacing.

Thus the focus on the 1-hour median annual maximum flood in Papers II and III prompts an examination of the percentage of subdaily data within the records. We calculate the number of years of subdaily data for each station as all years that have at least 200 days of subdaily data, and Figure 3.3 illustrates the distribution of subdaily record percentages in our dataset. Approximately 100 stations have subdaily data percentages exceeding 90%, while others range from 20% to 90%. Stations with less than half their record as subdaily data undergo manual validation to ensure that the sampling frequency adequately captured flood peaks at those locations. Stations with lower subdaily percentages are often larger catchments that have extensive overall record lengths, not below-average amounts of subdaily data. On average, selected stations have around 27 years of high-frequency data, with total record lengths in the dataset ranging from 20 to 129 years at station 62.5 (Bulken), as shown in Figure 3.3.

How to best make use of the available data will always be a question when estimating summary statistics for short durations. Improved data collection methods offering higher resolution data are crucial to analysis of short durations, but this does not mean that older (and often much longer) portions of the data record are useless in the context of short durations. Catchment dynamics ultimately dictate what sort of data resolution is necessary for capture of the flood peak. Inappropriately including coarser resolution data can introduce bias, while excluding it can result in high uncertainty in short-duration statistics. To assess the impact of sampling frequency on model performance in Paper II, we computed

correlations between model performance and both total record length and the percentage of subdaily data at the 1-hour duration. There was no correlation between model performance and either total record length or percentage of subdaily data for the evaluated metrics in the paper.

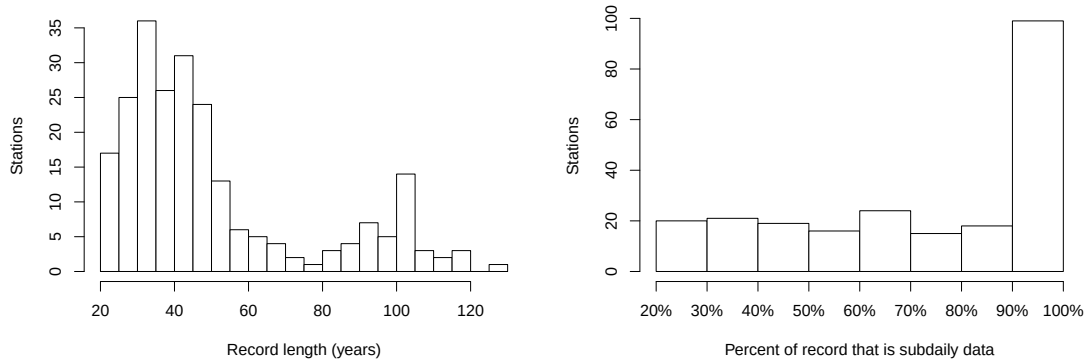

Figure 3.3: Histograms for record length (years) and percent of the record that is subdaily data. Only years that had at least 200 days of subdaily data count towards the subdaily data total when calculating the record percentage. Stations with less than 50 % of the record comprised of subdaily data were manually validated to make sure the sampling frequency of the data was high enough to represent flood peaks at that location.

3.4 Data processing for durations

The focus in this thesis is on flood-retention specific applications. This means a duration d represents the total flow volume for a time span of d hours, not individual flood events that lasted precisely d hours. This assessment of the total flow volume over a specific time window requires an aggregation-based approach as used in Breinl et al. (2021) and Javelle et al. (2002). Specifically, let $x_0(\tau)$ be a time series at the reference duration, where the reference duration is the finest time resolution of interest. Even spacing in the reference duration is enforced via regular sampling of a linear interpolation of the observed data. A moving average of window length d is applied to $x_0(\tau)$ to manufacture a new time series, $x_d(t)$:

$$x_d(t) = \frac{1}{d} \int_{t-d/2}^{t+d/2} x_0(\tau)\, d\tau \tag{3.1}$$

Block maxima or peak over threshold values can then be extracted from $x_d(t)$ to form sets of maxima given as:

$$\{y_{d,1}, y_{d,2}, \ldots, y_{d,n}\} \tag{3.2}$$

where, in the case of annual maxima, n is the number of years of data. The width d used as the length of the averaging window corresponds to the duration of interest and the average in Equation (3.1) can be repeatedly applied under different d to manufacture new sets of maxima that correspond to different durations of interest.

These sets of maxima produced under different d are dependent; that is, since longer duration series are always aggregated from series of shorter duration, the values in one set of maxima depend on the values in the other sets. This dependency structure that is neither predictable nor directly relatable to catchment properties. Figure 3.4 illustrates this. In some cases, annual maxima for different durations can originate from the same flood event, creating strong dependency due to temporal overlap and serial correlation. In other cases, maxima at different durations stem from different flood events with potentially distinct flood generation processes, resulting in weak dependency. This change in cross-duration correlation is not inherently related to catchment properties. Additionally, annual maxima may not consistently decrease as the aggregation interval increases. The factors causing this inconsistent behavior (e.g., two flood events of similar volume occurring closely or a wide and flat-topped flood event) are also not directly tied to catchment properties.

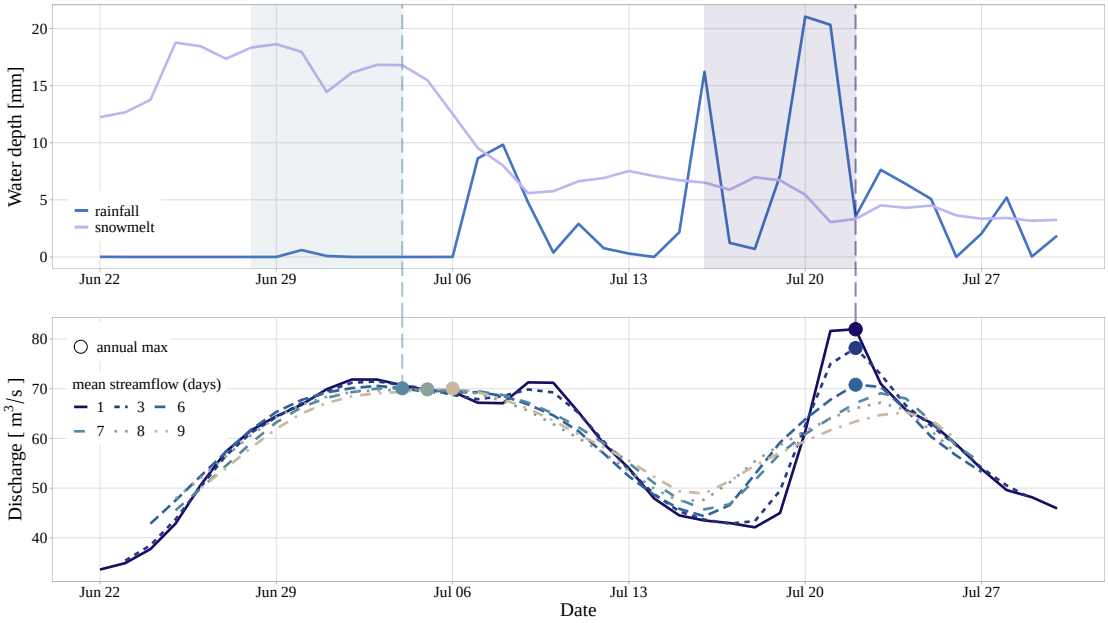

Figure 3.4: Figure showing two reasons why the dependency structure introduced by aggregation-based treatment of durations is not easily modeled: (i) annual maxima for each duration are not always primarily issued from the same flood event. In some cases, these flood events can have completely different generating processes (top panel; the shaded areas show the window of time from which the flood generating process is calculated) and (ii) annual maxima are not guaranteed to decrease as the duration of the averaging window is increased (see annual maxima at 7 days or greater). Data is from Sjodalsvatn gauging station, for the year 2009. Figure is obtained from Barna et al. (2023a).

Chapter 4

Contributions of the thesis

4.1 Extended QDF model

Existing QDF models usually assume that only the index flood changes with duration while the growth curve is held constant, e.g. Javelle et al. (2002), Cunderlik et al. (2006), and Breinl et al. (2021). This is termed "simple scaling" and is illustrated in the middle panel of Figure 4.5.

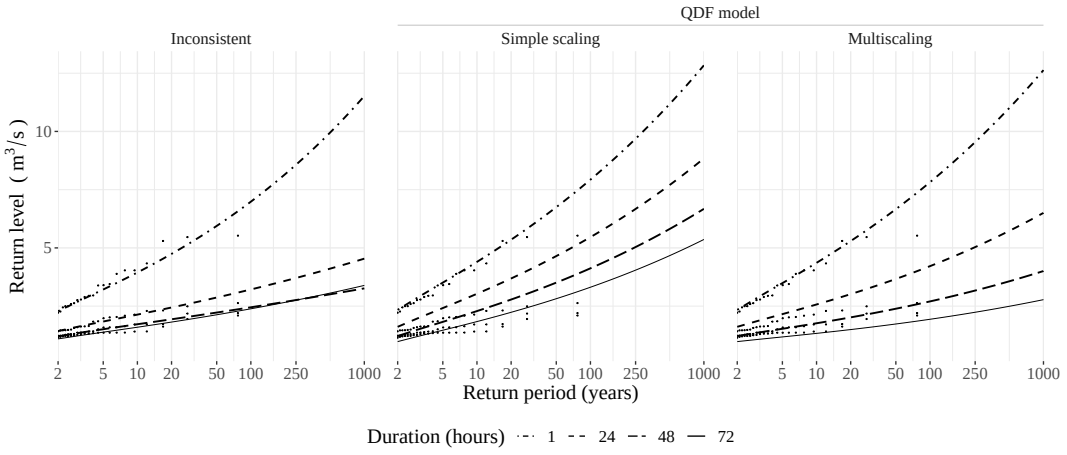

Figure 4.5: Return level plots from a synthetic data set showing (i) flood frequency curves estimated independently for four durations (left panel), (ii) output from a simple scaling QDF model (middle panel), and (iii) output from a multiscaling QDF model (right panel). The independent fits do not account for duration dependency. The simple scaling model accounts for duration dependence in the magnitude of the index flood but not the growth curve. The multiscaling model accounts for duration dependence in both the magnitude of the index flood and the slope of the growth curve. Figure is obtained from Barna et al. (2023a).

However, this assumption of constant growth curve across durations contradicts empirical analyses of runoff scaling properties in Norway which show the *culmination factor*, i.e. the ratio between peak and daily floods, may vary depending on return period (Engeland et al., 2016; Sælthun et al., 1997). In

these situations, application of simple scaling–which ignores the effect of duration dependency on the growth curve–can lead to poor estimation in the tails of the distribution.

"Multiscaling" models (see right panel, Figure 4.5) that allow for the ratio between growth curves of different durations to be dependent on return period already exist in the IDF literature (Van de Vyver, 2018; Courty et al., 2019; Fauer et al., 2021). However, in all existing models the different scaling components are placed on the location and scale parameters of the GEV in its standard location-scale parameterization. It is therefore challenging to directly interpret how the different scaling components influence the index flood and the growth curve, which are best interpreted in terms of quantile expressions.

Paper I introduces a multiscaling extension of the QDF model originally proposed by Javelle et al. (2002). This extension allows for the magnitude of the index flood and the slope of the growth curve to scale independently with duration. Additionally, paper I presents a Bayesian inference framework applicable to both the original QDF model and its extended version. This framework allows for flexible propagation of uncertainty and simultaneous model selection and parameter estimation. This approach enables us to evaluate how sensitive QDF models are to input durations and to determine the significance of adopting the multiscaling extension versus the original QDF model. Current QDF models are typically estimated in a two-step procedure where the characteristic duration parameter is estimated first, followed by an estimation of the remaining parameters (Javelle et al., 2002; Cunderlik et al., 2006). However, such two-step estimation does not typically provide uncertainty information and is difficult to use with multiscaling models.

4.1.1 Defining the extended QDF model

Let the relationship between the location parameter of the GEV and the median flood be given as in Equation 2.2. Furthermore, let the scale parameter be decomposed as a product of the median flood and a remainder term expressed as an exponential function, e^{β}, such that the new scale parameter β is given as

$$\beta = \log\left(\frac{\sigma}{\eta}\right). \tag{4.1}$$

Then the multiscaling extension of the QDF model allows η and β to depend on the aggregation interval d and additional parameters Δ_1 and Δ_2, respectively. The ξ parameter is kept duration-invariant due to the difficulties in estimating the ξ parameter stemming from the involved parametric form of the CDF (Equation 2.1). The annual maxima are

$$y_{d,i} \sim \text{GEV}\left(\eta_d, \beta_d, \xi\right) \tag{4.2}$$

where

$$\eta_d = \eta\,(1 + d\Delta_1)^{-1} \tag{4.3}$$

$$\beta_d = \log\left(\frac{\sigma}{\eta_d(1 + d\Delta_2)}\right) \tag{4.4}$$

and the distribution's quantiles for a duration d corresponding to exceedance probability p are given by

$$z_{d,p} = \frac{\eta}{1 + d\Delta_1}\left[1 + \frac{e^{\beta}}{1 + d\Delta_2}\left\{\frac{(-\log(1-p))^{-\xi} - \log(2)^{-\xi}}{\xi}\right\}\right] \tag{4.5}$$

with constraint

$$0 < \Delta_2 < \Delta_1. \tag{4.6}$$

The constraint on the Delta parameters reflects the fact that the data aggregation performed in QDF modeling is more likely to have a larger effect on the flood magnitude than on the decomposed scale parameter. Note the inverse of the characteristic duration parameter Δ from Javelle's original QDF model is used here for numerical stability during estimation. The value of the Δ_1 parameter reflects the "flashiness" of the floods measured; a narrow hydrograph will be associated with larger values of Δ_1. The Δ_2 parameter does not have an equally accessible hydrologic interpretation but can be interpreted as a measure of difference in growth curve slope across aggregation intervals; that is, if the ratio between peak and daily floods is heavily dependent on return period we would expect to see larger values of Δ_2.

As the aggregation window shrinks to zero, that is, as $d \to 0$, the extended model is equivalent to the standard GEV model that creates the traditional flood frequency curve. Similarly, as $\Delta_2 \to 0$, the extended model approaches Javelle's QDF model. The extended QDF model can thus be considered an extension of the original in the same way the original is an extension of the traditional flood frequency curve.

4.1.2 Defining the mixture model

We define also a mixture model that combines elements of the original and extended QDF models in an attempt to access the flexibility of the extended model without adding unnecessary complexity. The model is a weighted average of the original and extended models such that the density of the annual maximuma is given by

$$\sum_{j=1}^{2} m_j\, g(\cdot|\boldsymbol{\theta}_j) \tag{4.7}$$

where m_j is the weight on the component model, g is the density of the GEV distribution, $\boldsymbol{\theta}_1 = \{\eta_d^{\text{DD}}, \beta_d^{\text{DD}}, \xi^{\text{DD}}\}$ and $\boldsymbol{\theta}_2 = \{\eta_d^{\text{J}}, \beta^{\text{J}}, \xi^{\text{J}}\}$. Here the superscripts on the parameter sets denote the extended and original models, respectively (in Barna et al. (2023a) the extended model is referred to as the "Double-Delta" model and

the original as "Javelle"). Using the Bayesian estimation framework, parameter estimation and selection can be carried out simultaneously and the Δ_2 parameter is only added if merited.

Thus Equation 4.7 is a representation of a non-standard density from which it is possible to obtain quantile estimates that are an average over the distributions given by the extended model in Equation 4.2 and the original model in Equation 2.4, if the original model is reparameterized to take the inverse of the characteristic duration parameter.

4.1.3 A Bayesian framework for QDF

For the Javelle and Double-Delta models, Bayesian inference is performed using a Metropolis-Within-Gibbs algorithm (Robert et al., 1999). That is, samples from the conditional distribution of the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, are obtained by iterative sampling from the full conditional distributions of the individual parameters so that each component of the model is updated in turn. Prior distributions for the individual parameters assume independence. The prior on η (units: m^3/s) is a diffuse truncated normal distribution (truncNormal(40,100)) with a lower bound at zero. The prior on β is a diffuse Normal(0,100). For ξ, we adopt the methodology from Martins et al. (2000) and employ a shifted Beta(6,9) distribution within the interval $[-0.5, 0.5]$. In the Double-Delta model, the prior for Δ_1, equivalent to the prior for Δ in the Javelle model, is a Lognormal(0,5). The same values are applied to the prior for Δ_2, which uses a truncated Lognormal distribution with the lower bound determined by Δ_1.

The conditional distribution of the mixture model is given by

$$\pi\left(m, \boldsymbol{\theta}|\mathbf{y}\right) \propto \pi(m)\pi\left(\boldsymbol{\theta}|m\right) g(\mathbf{y}|\boldsymbol{\theta}, m) \tag{4.8}$$

where $\pi(\cdot|\cdot)$ is the generic conditional distribution consistent with this joint specification and $m \in \{\mathrm{DD}, \mathrm{J}\}$, $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, and $\mathbf{y} = (y_{d,i})_{i=1,d=1}^{i=n,d=D}$, where n is the number of years of data and D is the number of durations. The models have equal prior probability, with $\pi(m = J) = \pi(m = DD) = 0.5$. Simplification of Equation 4.8, considering the model without the model specification and separate parameter sets, gives the conditional distributions of Double-Delta and Javelle.

Moving between models changes the dimension of $\boldsymbol{\theta}$. To account for this, we employ a reversible jump MCMC algorithm, similar to the reversible jump methodology for normal mixtures described in Richardson et al. (1997). The reversible jump MCMC proceeds as follows:

1. updating $\boldsymbol{\theta}$:
   (a) if $m = \mathrm{DD}$ update η^{DD}, else update η^{J};
   (b) if $m = \mathrm{DD}$ update β^{DD}, else update β^{J};
   (c) if $m = \mathrm{DD}$ update ξ^{DD}, else update ξ^{J};
   (d) if $m = \mathrm{DD}$ update Δ_1 and Δ_2 parameters in sequence, else update Δ;
2. splitting one Delta into two, or combining two Deltas into one.

Step 1 is repeated 10 times under the same model before Step 2 (proposal to jump between models) is taken. Repeating Step 1 for either the Javelle or Double-Delta model details the MCMC algorithm used to fit the respective model. To move from Double-Delta to Javelle we need to merge Δ_1 and Δ_2 into one Δ. The combine proposal is deterministic and given by

$$\Delta = \Delta_1 + \Delta_2. \tag{4.9}$$

The reverse split proposal, going from Javelle to Double-Delta, involves one degree of freedom, so we generate a random variable u such that

$$u \sim Beta(5, 1) \tag{4.10}$$

which is then used to set

$$\begin{aligned} \Delta_1 &= u\Delta \\ \Delta_2 &= (1-u)\Delta. \end{aligned} \tag{4.11}$$

For this split move the acceptance probability is $\min\{1, A\}$ where

$$A = \frac{\pi(m', \boldsymbol{\theta}'|\mathbf{y})}{\pi(m, \boldsymbol{\theta}|\mathbf{y})q(u)} |J| \tag{4.12}$$

where $q(u)$ is the density function of u and J is the Jacobian of the transformation described in Equation 4.11. The acceptance probability for the corresponding combine move is $\min\{1, A^{-1}\}$ but with substitutions that adhere to the proposal in Equation 4.9.

4.1.4 Main findings for paper I

The multiscaling extension of the QDF model allows for a better approximation of tail behavior at multiple durations, as evidenced by improved modeling of both short-duration events and events with long return periods. In a case study comprising 12 study locations in Norway, the extended QDF model performs better than the original QDF model in 83% of the out of sample subdaily durations studied when performance is measured by the integrated quadratic distance. The mixture model shows the importance of the multiscaling extension: selectively adding the second characteristic duration parameter (Δ_2 in Equation 4.5) is not advantageous at the shortest durations as these durations tend to need maximum flexibility from the QDF models.

In general, QDF models are generally able to predict out-of-sample durations with a relatively moderate loss in accuracy when compared to in-sample estimates for the same durations. However, we found the QDF framework to be highly sensitive to the choice of durations used to fit the models. In particular, care should be taken to fit the QDF models with the minimum number of durations needed for the inference algorithm to converge. Generating too many sets of dependent data to fit the model can produce results that are both biased and overconfident (Figure 4.6). Moreover, the assumption of a constant shape parameter may be too restrictive to model a wide range of durations (Figure 4.7).

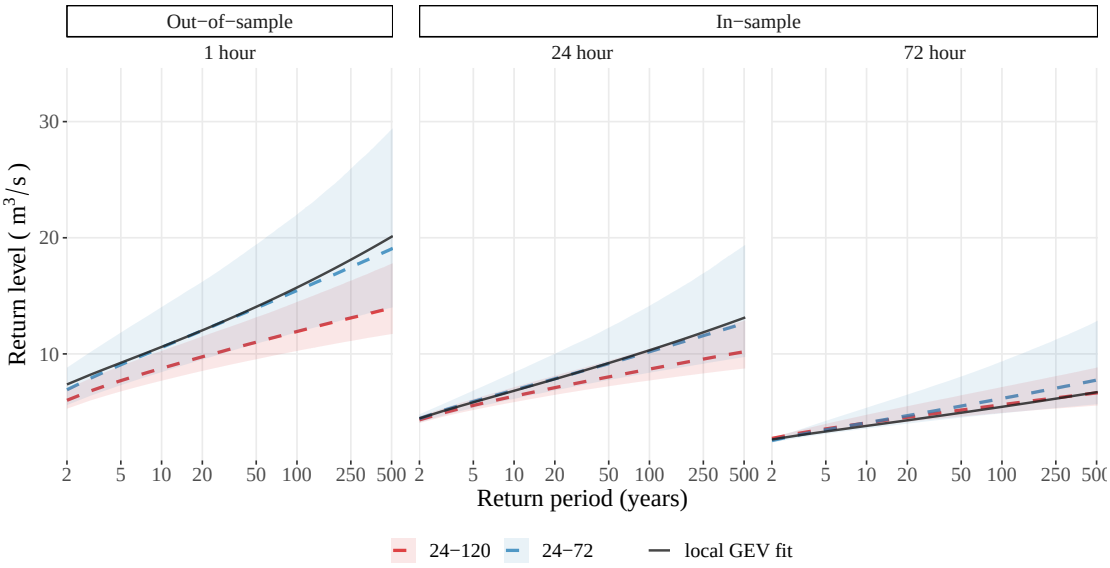

Figure 4.6: Return level plots generated from QDF models fit to two different data sets: one set with six durations [24, 36, 48, 72, 96, 120 hours] and one set with four durations [24, 36, 48, 72 hours]. The model fit to the six duration set is both overconfident and biased at shorter durations; the posterior mean return level estimates are consistently underestimated when compared to locally fit GEV models (solid black lines) and the 90% credible interval is artificially narrow and fails to capture the locally fit model for the 24 and 1 hour durations.

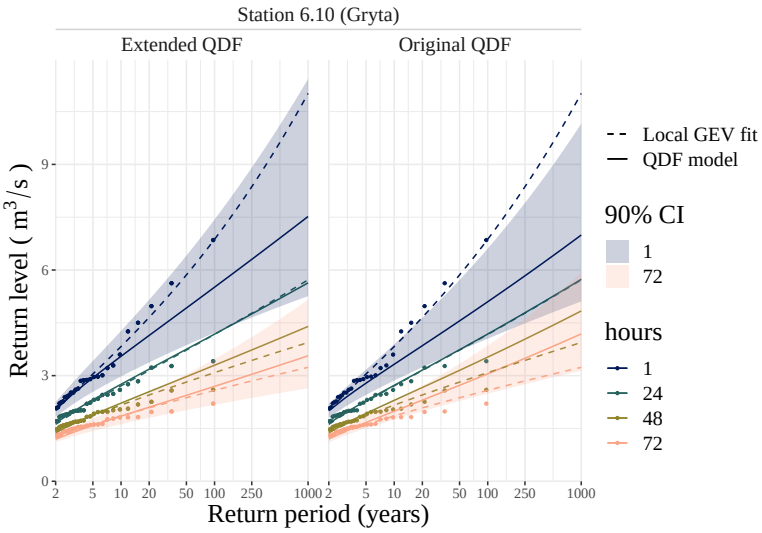

Figure 4.7: Return level plots showing a selected station where QDF models differ substantially from the reference model on in-sample durations. The reference models show a change in shape parameter with increasing duration. Figure is obtained from Barna et al. (2023a).

4.2 Index flood estimation at multiple durations

Regression models for RFFA are based on the assumption that spatial variations in flood quantiles are closely related to regional covariates (i.e., catchment and climatic characteristics) and that a regression model can describe this relationship. If the regression model is parametric (e.g., a linear, log-linear, nonlinear or generalized linear model) we additionally assume we know the functional form that describes the relationship between flood quantiles and regional covariates. A key question here is whether the functional forms established for flood quantiles at one duration are suitable for flood quantiles at a different duration. That is: are there duration-specific differences in how regional covariates influence flood quantiles?

We identify a gap in RFFA when it comes to assessing regression models at multiple durations. This gap exists both when (i) assessing predictive performance of flood estimates at different durations and (ii) explaining regional models at different durations, i.e. identifying and describing the underlying predictor-response relationships the regression models rely on.

In paper II we apply an existing semi-parametric model architecture (a generalized additive model, GAM) to address (i) and (ii) for a single robust flood quantile: the median (index) flood. We test the predictive performance of the GAM compared to the existing log-linear model for median flood estimation in Norway and a full machine learning model (an extreme gradient boosting tree ensemble, XGBoost). We establish the adequacy of the GAM as an explainable model through predictive performance. Predictive performance is measured as both predictive accuracy and reliability. We then use the GAM to detect and describe the functional relationships between the median flood and climatic and catchment descriptors at multiple durations.

While this study has relevance outside of QDF models, it also checks an often neglected assumption for regional QDF and regional regression-based IDF models: that differences in duration can be entirely captured by duration-specific scaling applied equally to all coefficients in a parametric regression model.

4.2.1 XGBoost based predictor pre-selection for GAMs

Predictor selection for data-driven models (e.g. GAMs and XGBoost) is in general challenging and a major roadblock to setting up analyses such as this one (Galelli et al., 2013). A secondary contribution of this study is development of a workflow that relies on a machine-learning based "tool", that, when used in combination with expert judgement, enhances the practicality of constructing GAMs.

The workflow for predictor selection has three steps. First, in Step A, machine learning-based algorithm for predictor selection is applied to the full regional covariate set to generate the *pre-selection set*. Then, in Step B, expert judgement uses the pre-selection set to inform selection of the *potential predictor set*. Finally, in Step C, the *final predictor set* is selected from the potential predictor set using the routines for shrinkage selection within the model architecture of the GAM as presented in Marra et al. (2011). The idea here is that Step A has the potential to

uncover predictor information that was previously unclear, while at the same time keeping Step A separate from Step C allows for a formal treatment of predictor selection uncertainty and reliance on a proven selection technique that has been rigorously compared to existing methodologies.

However, using a data-driven model for predictor selection implies in most cases a model-based preselection, which is not guaranteed to generate a predictor set that will work within other model architectures (i.e., Step A has the potential to generate a set of predictors that is good for the chosen machine learning model but not for a GAM). Therefore, part of the development of the workflow was careful selection of an algorithm and machine learning model architecture that could complement GAM development.

The chosen machine learning model is a gradient boosted tree ensemble restricted to have tree depth of one. The implementation of the gradient boosted tree ensemble is provided by XGBoost, and the tree depth is limited to one as we do not consider variable interactions in the GAMs.

The idea here is that depth-restricted tree-based ensembles can capture similar nonlinear predictor-response relationships to GAMs. This is a purely practical conjecture: both methods are insensitive to monotone transformations of predictors and adept at identifying nonlinear predictor-response relationships. However, unlike GAMs, boosted tree ensembles can infer the relative importance of predictors even when inputs are irrelevant or highly correlated. This makes them an appealing option for pre-selection based on the entire catchment descriptor set. To ensure nonredundant predictor selection, we incorporate the boosted tree ensemble within an appropriate predictor selection algorithm.

The chosen algorithm is an existing algorithm (the *Iterative Independent Selection*, or IIS, algorithm of Galelli et al. (2013), developed for application to hydrology (Prasad et al., 2017; He et al., 2022; Pesantez et al., 2020)) that has been adapted to take a particular machine learning model architecture (XGBoost). This adaptation of IIS to XGBoost was suggested by Alsahaf et al. (2022) for parameter selection for classification problems; we extend the adaptation to regression problems. Additionally, given our particular use case, we made several minor additions to the algorithm to increase its redundancy and allow for practitioners to assess the consistency of the algorithm output. These were (i) averaging an internal variable ranking over 25 bootstrap samples as in Laimighofer et al. (2022b) and (ii) running the IIS algorithm within a resampling method and assessing the consistency of the algorithm output over the resampled data.

4.2.2 Main findings for paper II

We find that the predictive accuracy and reliability of the GAM developed (*floodGAM*) matched or exceeded that of the benchmark models at both durations studied (1 and 24 hours). Specifically, on the 1 hour duration, floodGAM was both more reliable and more accurate than the existing log-linear model for median flood estimation in Norway. This is illustrated in Figure 4.8, which shows that application of the log-linear model (RFFA_2018) outside of the duration it was developed for results in a systematic underestimation of the median flood in large,

snowmelt driven catchments. Predictive accuracy between floodGAM and the existing model was measured by five different scoring measures (root mean squared error, continuous ranked probability score, mean absolute error, mean relative error, mean absolute percent error) and the differences were statistically significant for all measures at the 1 hour duration. On the 24 hour duration, which was the duration used to develop the existing log-linear model, there were no statistically significant differences between floodGAM and the existing log-linear model, or between floodGAM and XGBoost on either the 1 hour or 24 hour mean absolute error (this was the only scoring measure computed for XGBoost due to constraints introduced by distributional assumptions in the analysis).

Within the predictor set selected for this study, we observe duration-specific differences in the form of the functional relationship between the median flood and the two catchment descriptors effective lake percentage and catchment shape. Significant duration-specific differences were not observed on the remaining five predictors. Ignoring the differences observed on the two predictors results in a statistically significant decline in predictive performance.

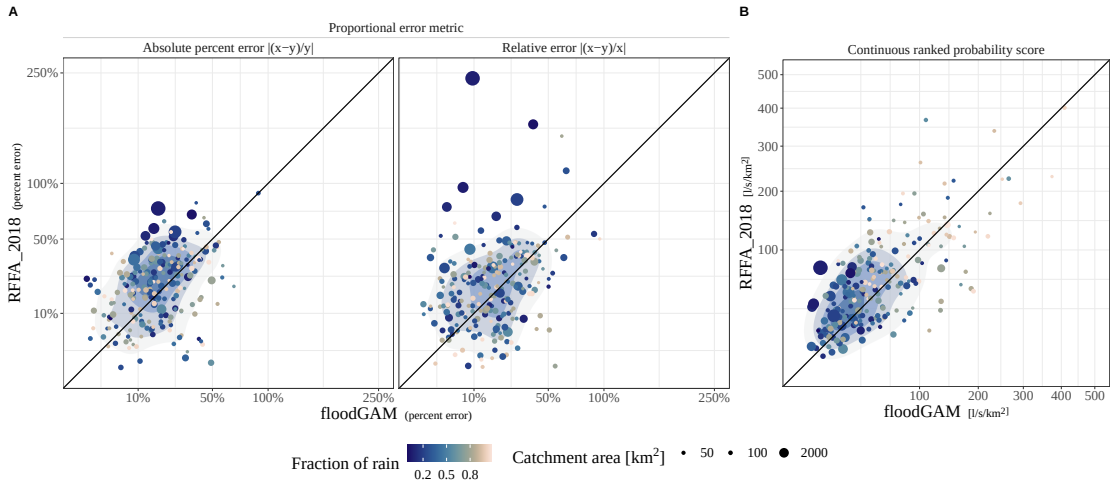

Figure 4.8: Model to model comparison on absolute percent error, relative error, and the continuous ranked probability score for the log-linear benchmark model (RFFA_2018) and floodGAM on the 1 hour duration. In the panel headers, x represents the predicted value and y the observed value. Points falling above the diagonal line indicate stations where RFFA_2018 performed worse than floodGAM. Points falling below the diagonal line indicate stations where floodGAM performed worse than RFFA_2018. The 2D kernel density estimation of point density is underlaid to aid visual interpretation. Point size shows catchment area, point color indicates the fraction of rain contribution to flood.

4.3 Regional flood frequency analysis at multiple durations

RFFA commonly takes one of two paths: (i) the quantile regression technique (QRT), where a regression model is developed for a predetermined flood quantile of interest (Haddad et al., 2012; Ahn et al., 2016; Rahman et al., 2020; Ouali et al.,

2016; Chebana et al., 2014; Zaman et al., 2012) and (ii) the parameter regression technique (PRT), where regression models are developed for the parameters of an extreme value distribution (Haddad et al., 2012; Ahn et al., 2016; Rahman et al., 2020; Thorarinsdottir et al., 2018; Lima et al., 2016). Since RFFA typically focuses on regionalization of high quantiles, implementations of the two approaches often rely on distributional assumptions that allow for extrapolation beyond the range of the observed data. In practice this means that the response variables for both the PRT and the QRT are obtained from at-site frequency analysis. In the case of the PRT, the response variables are the estimates of the parameters of the extreme value distribution obtained from at-site frequency analysis. In the case of the QRT, the target quantiles used as the response variables are estimated flood quantiles obtained from at-site frequency analysis. Using estimated quantiles to construct a series of (independent) regressions distinguishes the QRT from the more classic *quantile regression* (see, e.g. Fasiolo et al. (2021)).

In data-sparse situations model structure plays a large role. We identify a gap when it comes to assessing the impact of model structure on duration consistency for estimates at out-of-sample locations. Paper III compares both the predictive performance of the QRT and the PRT and their ability to provide estimates at out-of-sample locations that match the observed consistency between durations in the data. We consider two durations, the 1 hour and 24 hour durations. The durations are treated independently this analysis: we derive return level estimates using the PRT and the QRT once for the 1 hour duration and once for the 24 hour duration. Return level estimates from these independent fits are then assessed for consistency. Predictive accuracy is assessed for each duration within the range of the observed data using the quantile score. Duration consistency is assessed on the full range of return periods from 2 to 1000 years.

To ensure a direct comparison of the approaches, we maintain a constant predictor set. The predictor set is chosen based on prior work in paper II. Given the diverse range of durations, quantiles, and distributional parameters in our study, a practical approach is to use data-driven regression models, specifically a generalized additive modeling (GAM) approach. This method has been effectively employed in previous flood quantile modeling studies (Chebana et al., 2014; Msilini et al., 2022; Rahman et al., 2018; Barna et al., 2023b). Additionally, to further facilitate a direct comparison between the PRT and QRT, we reparameterize the generalized extreme value (GEV) distribution using the median as the location parameter, following recent work by Castro-Camilo et al. (2022). Since the median is also a target quantile for the QRT, this means we have one regression model that is common to both methods. They therefore model the center of the distribution equivalently. Any differences will be concentrated in how the spread and form of the distribution is modeled.

While regional QDF models can be used to obtain design values at ungauged locations, they introduce strong assumptions on the tail behavior of the model and can be challenging to estimate due to the properties of the GEV distribution. We therefore chose to establish regional models for each duration independently.

4.3.1 Practical at-site GEV estimation with the probabilistic programming language Stan

Both the PRT and QRT rely on at-site frequency analyses to obtain response variables. This at-site estimation can be time consuming, particularly when we need estimates at several durations. A secondary contribution of this study is implementation of a Bayesian inference method for at-site GEV estimation using the probabilistic programming language Stan. Stan provides full Bayesian statistical inference using a powerful Hamiltonian Monte Carlo (HMC) algorithm to fit models, provided the parameter bounds in the model are properly defined (Gelman et al., 2015). The quantile-based reparameterization used in paper III allows for us to define support-enforcing bounds for the GEV distribution; see, for example, the discussions at Barna (2021) and Barna (2022). Stan allows for very fast sampling (Neal et al., 2011), but, most importantly, provides a range of useful diagnostics: using HMC to explore the target distribution means failures in geometric ergodicity manifest in distinct behaviors that can be developed into diagnostic tools. Gelman et al. (2015) provides an overview of the model diagnostics within Stan. Given that we need accurate estimates at hundreds of individual stations and durations, and that some of the stations are being fit with as few as 20 data points, obvious indicators when something goes wrong in the at-site estimation present a simplification of the modeling process.

4.3.2 Main findings for paper III

The PRT is more effective at preserving duration consistency compared to the QRT. We observed five stations where the QRT produced duration inconsistent estimates at out-of-sample locations whereas the PRT (and the local fit for those locations) kept duration consistency. This QRT-specific inconsistency was characterized by deviations in the out-of-sample estimated return levels at high (greater than 2 year) return periods. This behavior was specific to QRT and was not easily attributable to either record length or specific catchment properties; see Figure 4.9.

Duration inconsistencies that are not characterized by the QRT-specific behavior described above tend to arise in data-rich areas of the distribution. We found that estimating the distribution center (median) out-of-sample produced duration inconsistencies at 46 out of 232 stations. That is, at 46 stations the estimated 24 hour median flood was higher than the estimated 1 hour median flood. Here estimates are given by the posterior mean. The estimates for the parameters controlling the spread and shape of the distribution, on the other hand, tend to decrease with increasing duration (Figure 4.10). Analysis of uncertainty shows that, for all distributional parameters, predictions at out-of-sample locations show considerable overlap in the range of possible values between the 1 and 24 hour durations.

Finally, we find that the predictive accuracy of the PRT and the QRT, as measured by the quantile score, is very similar at each return period tested (10-, 20-, and 50-year return periods). This aligns with the results shown in Ahn et al.

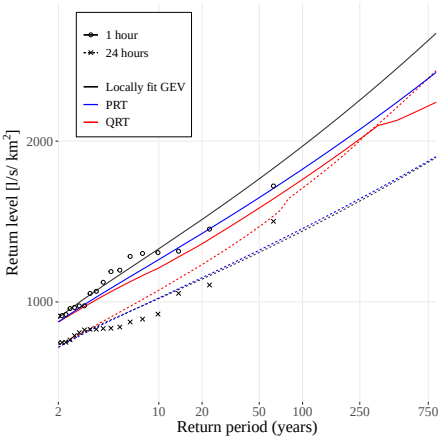

Figure 4.9: Return level plot for one of the five instances the QRT produces a duration inconsistent out-of-sample return level estimate but the PRT does not. Observed data points are indicated with circles (1 hour annual maxima) or crosses (24 hour annual maxima).

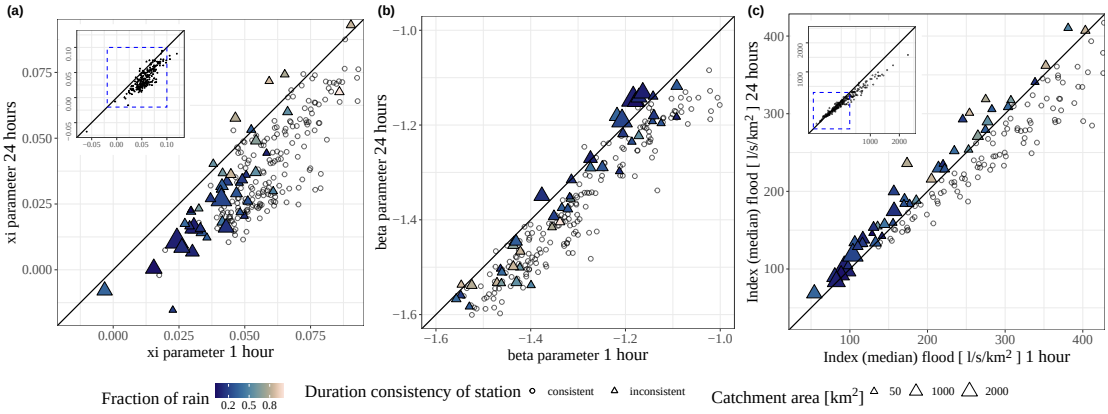

Figure 4.10: Duration-to-duration comparison of the out-of-sample parameter estimates for the three parameters of the GEV distribution (under the parameterization used in paper III). Stations that have at least one duration inconsistent return level are indicated by triangular points and colored and sized accroding to catchment descriptors.

(2016), Rahman et al. (2020), and Haddad et al. (2012), all of which tested the predictive performance at return periods within the range of the observed data and found only modest differences between the PRT and QRT. Beyond the range of the data, the PRT and QRT show similar tail behavior compared to local fits. When we evaluate return level predictions relative to the locally estimated shape parameter (a proxy for the tail behavior of the underlying flood frequency curve), both methods consistently result in higher return level estimates at stations with smaller shape parameter values and lower return level estimates at stations with larger shape parameter values.

Chapter 5

Discussion and future considerations

The end goal of frequency analyses at multiple durations is typically estimation of return levels that are consistent between durations. When we have sufficient data at a site, it is possible to define a fully consistent model. Paper I develops models for this at-site case, with a particular focus on a specific type of model: the QDF model.

Regional modeling of duration-consistent extremes is much more involved. QDF models, and their analogues for precipitation modeling (IDF models), have been extended to a regional context before (Javelle et al., 2002; Cunderlik et al., 2006). However, the assumptions required place very strict constraints on the model components. These constraints are typically not supported by empirical analyses (papers I and II). The regional models in papers II and III extend estimates to out-of-sample locations without explicit consistency constraints and investigate the nature of the duration inconsistencies that arise.

5.1 QDF models

QDF models establish a parametric relationship between quantiles of different durations, making them a less flexible option than independently estimating—and post-processing—return levels. Nevertheless, there are several potential benefits to this parametric formulation. Most notably, it permits extrapolation to durations that lack observed data. This unobserved duration of interest would typically be the shortest duration that captures peak discharge. Estimation of the culmination factor (the ratio between peak and daily floods), and establishing if this culmination factor increases with increasing return period, is also often relevant.

Paper I demonstrated that QDF models can effectively predict out-of-sample sub-daily durations (1 and 12 hours) using coarser time resolution data, showing only a modest decrease in accuracy compared to in-sample estimates. The multiscaling extension proves valuable in most cases (83% of the out-of-sample sub-daily durations), and allows for the culmination factor to increase with increasing return period. Careful selection of input durations is important to avoid biasing results and artificially reducing predictive uncertainty, but overall, the models

demonstrate a strong capability to replicate in-sample behavior at out-of-sample (i.e. unobserved) durations.

However, the results also indicate the assumption of a constant shape parameter is limiting, particularly when modeling a wide range of durations (Figure 4.7). Assuming a constant shape parameter while simultaneously fitting to multiple durations means the estimate of the shape parameter will always, by necessity, be a compromise between the input durations. In this case, it does not matter if we accurately replicate in-sample behavior if that behavior is not characteristic of the shortest target durations. This assumption of a constant shape parameter is both common in these types of models (Fauer et al., 2021) and not easily solved (Martins et al., 2000). In practice, we may be able to mitigate this issue at individual locations by carefully selecting input durations that are close enough to our out-of-sample duration of interest, if data for such input durations exist.

The assumption of a constant shape parameter is likely to be even more limiting in a regional context, particularly because the shape parameter is notoriously difficult to relate to catchment and climatic characteristics. Working around this issue with careful selection of input durations is unlikely to be a practical option in a regional model seeking to fit a variety of catchments at once. Another challenge in a regional context is implementation of an inference approach once the parameters of the QDF model depend on covariates. Maximum likelihood-based estimation is often unstable (Ulrich et al., 2020) and Bayesian approaches in this scenario are very computationally expensive: the more efficient (e.g. gradient-based) approaches are unsuitable for the particular parametric form of the QDF model, where the parameter dependent support is itself dependent on the input duration.

Given these considerations, we identify the most promising direction for QDF models is as a volume-based method for scaling to peak discharge at individual sites. Estimating peak discharge is, in general, challenging. Scaling between daily and unobserved peak discharge in Norway is currently performed by establishing a relationship between daily and peak discharge for a selection of the largest floods from a hydrologically similar catchment (Wilson et al., 2011). In this context, QDF models offer a "volume-based" scaling, in contrast to "event-based" scaling. The end goal for both of these approaches is some unobserved peak discharge. The question is then: is it better to try and reach this peak discharge by examining how flood volumes change (i.e. QDF), or picking out a few events we think are representative of the catchment and creating ratios from those?

A potential advantage of QDF models as an alternative to event-based scaling to peak discharge is that their structure accommodates incorporating data with varying record lengths for different durations. This potentially allows for the inclusion of information on short durations–even when data for these periods is limited–although future work would need to establish how varying the record length impacts the results, as it could, for example, place different weight on different durations. Nevertheless, from a practical standpoint a framework that allows for easy inclusion of all the data and approach that eliminates the need to make subjective choices about specific events or data to include has much appeal.

5.2 Regional models at multiple durations

QDF models enable us to extrapolate to unobserved durations. On the other hand, the regional models in this thesis enable us to extend *observed* durations to ungauged locations. That is: when we have observed data at a particular duration from an adequate number of stations, we can develop an RFFA model that enables us to extend this particular duration to an ungauged location. This extension does not take place under explicit duration consistency constraints. Our analyses show that there are special considerations when building RFFA models that perform well when applied to a variety of durations. In particular, both the type of regression model and the structure of the frequency analysis matter once we start considering more than one duration.

Paper II showed that data-driven regression models are beneficial—with regards to both predictive accuracy and reliability—when we require estimates at multiple durations. While it is possible to create a parametric regression model that performs well at a specific duration, achieving optimal performance at multiple durations would necessitate reformulating the parametric relationship for each one. As a result, data-driven models present a potential streamlining of the modeling process when we require estimates at multiple durations. In a subsequent study, paper III used the developed data-driven regression model in a regional frequency analysis and found that duration consistency is better preserved when the frequency analysis is structured such that the regression model operates on the parameters of the extreme value distribution, rather than directly on flood quantiles.

The parameters of an extreme value distribution describe its center, spread and shape. Paper III additionally showed duration inconsistencies often arise in conjunction with out-of-sample estimation of the distribution center. In the reparameterization of the GEV used in this thesis, the center of the distribution is described by the median. The median annual maximum flood is the index flood for a catchment, which in flood frequency analysis has the useful interpretation as the scaling factor separating the order of magnitude of a flood from the shape and slope of the growth curve (see Section 2.1.4). That is, the results in paper III demonstrate that difficulties distinguishing between the order of magnitude of floods at different durations lead to duration inconsistencies at a subset of out-of-sample locations. These difficulties are not observed to the same extent in the out-of-sample estimation of the parameters controlling the spread and shape of the distribution (Figure 4.10).

It is possible that out-of-sample estimation of the index flood is simply more challenging than estimating the spread and shape parameters; Haddad et al. (2012) found greater heterogeneity in the standard error of predicted values for the index (mean) flood compared to the standard deviation and skew of their extreme value distribution (Log-Pearson Type III). It is also possible that the duration-specific distinctions seen in the parameters controlling the shape and slope of the growth curve are a product of the aggregation-based approach we use to obtain different durations; i.e., that our data processing smooths the data sufficiently to influence the two parameters linked to variance and skewness in the

GEV distribution. Practically, these results suggest that the index flood, when it is used as a location parameter for a distribution, is the most promising candidate for consistency constraints if we aim to ensure duration consistency through explicit model constraints. The median, moreover, is a relatively interpretable parameter of the extreme value distribution, and its units align with the original streamflow time series. Analysis of uncertainty for each of the three distributional parameters shows another potential option for duration-consistent estimates: credible intervals for all three parameters show considerable overlap between different durations, making the post-processing methods in Roksvåg et al. (2021) a potential solution for extending RFFA model estimates if enforcing explicit consistency constraints is not desirable.

The regional models in this thesis use annual maxima since flood guidelines in Norway use annual maxima. However, an area of future work could involve exploring the impact of seasonal flood regimes on RFFA models at different durations. Paper II conducted a preliminary investigation in this area and found season-specific changes in the partial response curves for climatic variables, which were not evident in the geographical catchment descriptors or mean annual runoff (see Appendix F of paper II). A potential explanation for this behavior is that when using annual maxima, relationships between climatic predictors and annual maxima may represent a compromise between various flood generating processes. This aligns with findings in previous studies (e.g., Ouarda et al. (2006), McCuen et al. (2003), Fischer et al. (2021)). Mixture models that explicitly account for flood generating processes exist, such as those in Fischer (2018). However, these models are not explicitly suited to the aggregation-based approach to obtaining different durations. A potential avenue forward could be definition of seasonal blocks as in Ulrich et al. (2021), who developed IDF curves with monthly varying parameters.

Chapter 6

Conclusions

Objective (i): Development of local models for situations where we have sufficient data available at a single gauged site and wish to extend the flood frequency estimates to unobserved durations.

Summary: Extending to unobserved durations involves creating a parametric relationship between durations, like in QDF models. To make this parametric relationship more realistic, we can allow the index flood magnitude and growth curve slope to scale independently with duration. However, since QDF models are fit on several durations simultaneously the estimates will always, by necessity, represent a compromise between the input durations. This can pose a problem, if, for example, the shape parameter changes significantly from duration to duration.

Research questions:

1. **Do models that allow for the ratio between return levels to change with return period improve our ability to predict unobserved sub-daily durations? (Paper I)**

   Allowing the ratio between return levels of different to increase with increasing return period (allowing the index flood magnitude and growth curve slope to scale independently) improves our ability to model several durations simultaneously. This advantage is particularly pronounced when modeling events with long return periods and/or short (sub-daily) durations. This is in line with findings from precipitation (IDF) models (e.g. Fauer et al. (2021) and Van de Vyver (2018)). Our extension relies on a new use of an existing parametrization of the GEV distribution, allowing for the magnitude of the index flood and the slope of the growth curve to scale independently with duration. This increases the interpretability of the model as compared to existing approaches. The multiscaling extension for QDF models proposed in paper I is therefore recommended if the goal is extension to unobserved sub-daily durations.

2. **How sensitive are local models to the selected input durations? (Paper I)**

The QDF models are sensitive to the input durations used to fit them. The models should be fit with the minimum number of durations needed for the inference algorithm to converge; generating too many sets of dependent data to fit the model can produce results that are both biased and overconfident. This is an often overlooked aspect of models that simultaneously estimate durations under consistency constraints (i.e. QDF and IDF models). Our investigation reveals a new finding with practical implications.

Moreover, care should be taken to select an appropriate range of input durations. In particular, attempting to model a wide range of durations with significant changes in the shape parameter may not work well with QDF models, as they assume the shape parameter to be constant across durations. This issue is also noted in, for example, Roksvåg et al. (2021) and Fauer et al. (2021).

Objective (ii): Development of regional models for situations where we have observed data at the duration of interest at a sufficient number of sites and wish to extend the flood frequency estimates at that duration to other, potentially ungauged, sites.

Summary: The type of regression model and the structure of the frequency analysis matter once we start constructing RFFA models for more than one duration. First, the flexibility afforded by a semi-parametric regression model is beneficial because we do not have to reformulate the functional form of the predictor-response relationship at each duration. Previous research has established that it is advantageous to relax parametric assumptions in RFFA (Chebana et al., 2014; Msilini et al., 2022; Rahman et al., 2018); we show it is particularly beneficial when multiple durations are considered. Second, RFFA models that regress on the parameters of the extreme value distribution, rather than the quantiles, better preserve duration consistency. When predicting distribution parameters at out-of-sample locations it is often estimates of the distribution center (median) that are duration inconsistent. While other RFFA research compares regression on parameters vs quantiles of an extreme value distribution, our study is the first to examine the impact on duration consistency.

Research questions:

3. **Can a semi-parametric (i.e. "data-driven") regression model achieve comparable or improved performance to two benchmark models (one parametric and one non-parametric) on the 1 hour and/or the 24 hour duration? (Paper II)**

   Yes. The predictive accuracy and reliability of GAM developed for median flood estimation matches or exceeds that of the benchmark models at both durations studied. GAMs therefore present a modeling benefit and a potential streamlining of the modeling process when we require estimates at multiple durations. We note the fact that the performance of the log-linear model matches that of the GAM on the specific duration for which

the log-linear model was developed and estimated somewhat contradicts the findings in Chebana et al. (2014), Msilini et al. (2022), and Rahman et al. (2018), which consistently report that GAMs outperform log-linear models. This difference may be attributed to the treatment of statistical significance; existing research does not test for statistically significant differences. In agreement with existing research, we also noted slightly lower point estimates for the GAM compared to the log-linear model across all metrics and durations, but we found this difference was not statistically significant at the duration the log-linear model was developed for.

4. **Within a regional regression model, can we identify and describe duration-specific differences in how catchment covariates influence the median flood? How impactful are these differences? (Paper II)**

   We observe duration-specific differences in the form of the functional relationship between the median flood and some of the catchment descriptors. Ignoring these differences results in a statistically significant decline in predictive performance. This suggests that it may be difficult to obtain optimal performance on all durations when assuming a fixed or parametric form between predictors and response.

5. **How does developing regression models for flood quantiles compare to developing regression models for extreme value distribution parameters in terms of predictive performance and consistency between durations? (Paper III)**

   The predictive accuracy of the approaches is very similar, confirming the findings of Haddad et al. (2012), Ahn et al. (2016), and Rahman et al. (2020). However, the parameter regression technique is more effective at preserving duration consistency than the quantile regression technique.

6. **If our regional models produce estimates that are duration inconsistent, at what return period do we observe the inconsistent estimate? Is the return period within the range of the observed data? (Paper III)**

   Duration inconsistencies tend to occur in data-rich areas of the distribution; we observed the out-of-sample estimation of the index (median) flood to be duration inconsistent at about 20% of stations used in paper III. This aligns with the results of Haddad et al. (2012), which found greater heterogeneity in the standard error of predicted values for the index (mean) flood compared to regression models for the standard deviation and skew of the Log-Pearson Type III distribution.

Bibliography

Ahn, Kuk-Hyun and Richard Palmer (2016). "Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique." In: *Journal of Hydrology* 540, pp. 515–526.

Alfieri, Lorenzo et al. (2017). "Global projections of river flood risk in a warmer world." In: *Earth's Future* 5.2, pp. 171–182.

Allahbakhshian-Farsani, Pezhman et al. (2020). "Regional flood frequency analysis through some machine learning models in semi-arid regions." In: *Water Resources Management* 34, pp. 2887–2909.

Alsahaf, Ahmad et al. (2022). "A framework for feature selection through boosting." In: *Expert Systems with Applications* 187, p. 115895.

Aziz, Kashif et al. (2014). "Application of artificial neural networks in regional flood frequency analysis: a case study for Australia." In: *Stochastic environmental research and risk assessment* 28, pp. 541–554.

Ball, J et al., eds. (2019). *Australian Rainfall and Runoff: A Guide to Flood Estimation*. Commonwealth of Australia.

Balocki, James B and Stephen J Burges (1994). "Relationships between n-day flood volumes for infrequent large floods." In: *Journal of Water Resources Planning and Management* 120.6, pp. 794–818.

Barna, D. M. (2021). *Aberrant behavior in dynamically bounded parameters, reparameterized GEV distribution*. URL: https://discourse.mc-stan.org/t/aberrant-behavior-in-dynamically-bounded-parameters-reparameterized-gev-distribution/22726.

— (2022). *The support of the GEV distribution is not enforced with the currently implemented xi transform no. 1345*. URL: https://github.com/paul-buerkner/brms/issues/1345.

Barna, Danielle M et al. (2023a). "Flexible and consistent Flood–Duration–Frequency modeling: A Bayesian approach." In: *Journal of Hydrology* 620, p. 129448.

Barna, Danielle M et al. (2023b). "Regional index flood estimation at multiple durations with generalized additive models." In: *EGUsphere* 2023, pp. 1–43. DOI: 10.5194/egusphere-2023-2335. URL: https://egusphere.copernicus.org/preprints/2023/egusphere-2023-2335/.

Blöschl, Günter (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

Blöschl, Günter and Murugesu Sivapalan (1995). "Scale issues in hydrological modelling: a review." In: *Hydrological processes* 9.3-4, pp. 251–290.

Bocchiola, Daniele, Carlo De Michele, and Renzo Rosso (2003). "Review of recent advances in index flood estimation." In: *Hydrology and Earth System Sciences* 7.3, pp. 283–296.

Breinl, Korbinian et al. (2021). "Understanding the relationship between rainfall and flood probabilities through combined intensity-duration-frequency analysis." In: *Journal of Hydrology* 602, p. 126759.

Castellarin, A et al. (2012). "Review of applied statistical methods for flood frequency analysis in Europe." In.

Castro-Camilo, Daniela, Raphaël Huser, and Håvard Rue (2022). "Practical strategies for generalized extreme value-based regression models for extremes." In: *Environmetrics*, e2742.

Chebana, Fateh et al. (2014). "Regional frequency analysis at ungauged sites with the generalized additive model." In: *Journal of Hydrometeorology* 15.6, pp. 2418–2428.

Chen, Tianqi et al. (2015). "Xgboost: extreme gradient boosting." In: *R package version 0.4-2* 1.4, pp. 1–4.

Clarke, Robin T (2001). "Separation of year and site effects by generalized linear models in regionalization of annual floods." In: *Water resources research* 37.4, pp. 979–986.

Coles, Stuart (2001). *An introduction to statistical modeling of extreme values*. Springer.

Coles, Stuart G and Jonathan A Tawn (1996). "A Bayesian analysis of extreme rainfall data." In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 45.4, pp. 463–478.

Courty, Laurent G et al. (2019). "Intensity-duration-frequency curves at the global scale." In: *Environmental Research Letters* 14.8, p. 084045.

Crochet, Philippe (2012). "Flood-Duration-Frequency modeling application to ten catchments in Northern Iceland." In: *Report NVI* 6.

Cunderlik, Juraj M and Taha BMJ Ouarda (2006). "Regional flood-duration–frequency modeling in the changing environment." In: *Journal of Hydrology* 318.1-4, pp. 276–291.

Cunnane, Conleth (1988). "Methods and merits of regional flood frequency analysis." In: *Journal of Hydrology* 100.1-3, pp. 269–290.

Dalrymple, Tate (1960). *Flood-frequency analyses, manual of hydrology: Part 3*. Tech. rep. USGPO,

Desai, Shitanshu and Taha BMJ Ouarda (2021). "Regional hydrological frequency analysis at ungauged sites with random forest regression." In: *Journal of Hydrology* 594, p. 125861.

Durocher, Martin, Fateh Chebana, and Taha BMJ Ouarda (2015). "A nonlinear approach to regional flood frequency analysis using projection pursuit regression." In: *Journal of Hydrometeorology* 16.4, pp. 1561–1574.

Elith, Jane, John R Leathwick, and Trevor Hastie (2008). "A working guide to boosted regression trees." In: *Journal of animal ecology* 77.4, pp. 802–813.

Engeland, Kolbjørn et al. (2016). *Utvalg og kvalitetssikring av flomdata forflomfrekvensanalyser*. Tech. rep. NVE.

Engeland, Kolbjørn et al. (2020). *Lokal og regional flomfrekvensanalyse*. Tech. rep. NVE.

England Jr, John F et al. (2019). *Guidelines for determining flood flow frequency—Bulletin 17C*. Tech. rep. US Geological Survey.

Esmaeili-Gisavandani, Hassan, Heidar Zarei, and Mohammad Reza Fadaei Tehrani (2023). "Regional flood frequency analysis using data-driven models (M5, random forest, and ANFIS) and a multivariate regression method in ungauged catchments." In: *Applied Water Science* 13.6, p. 139.

Farquharson, FAK, JR Meigh, and JV Sutcliffe (1992). "Regional flood frequency analysis in arid and semi-arid areas." In: *Journal of Hydrology* 138.3-4, pp. 487–501.

Fasiolo, Matteo et al. (2021). "Fast calibrated additive quantile regression." In: *Journal of the American Statistical Association* 116.535, pp. 1402–1412.

Fauer, Felix S et al. (2021). "Flexible and consistent quantile estimation for intensity–duration–frequency curves." In: *Hydrology and Earth System Sciences* 25.12, pp. 6479–6494.

Field, Christopher B (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

Filipova, Valeriya, Deborah Lawrence, and Thomas Skaugen (2019). "A stochastic event-based approach for flood estimation in catchments with mixed rainfall and snowmelt flood regimes." In: *Natural Hazards and Earth System Sciences* 19.1, pp. 1–18.

Fischer, Svenja (2018). "A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons." In: *Journal of Applied Statistics* 45.15, pp. 2831–2847.

Fischer, Svenja and Andreas H Schumann (2021). "Regionalisation of flood frequencies based on flood type-specific mixture distributions." In: *Journal of Hydrology X* 13, p. 100107.

Fisher, Ronald Aylmer and Leonard Henry Caleb Tippett (1928). "Limiting forms of the frequency distribution of the largest or smallest member of a sample." In: 24.2, pp. 180–190.

Galelli, Stefano and A Castelletti (2013). "Tree-based iterative input variable selection for hydrological modeling." In: *Water Resources Research* 49.7, pp. 4295–4310.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020). *Regression and other stories*. Cambridge University Press.

Gelman, Andrew, Daniel Lee, and Jiqiang Guo (2015). "Stan: A probabilistic programming language for Bayesian inference and optimization." In: *Journal of Educational and Behavioral Statistics* 40.5, pp. 530–543.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. CRC press.

Gioia, A et al. (2012). "Influence of infiltration and soil storage capacity on the skewness of the annual maximum flood peaks in a theoretically derived distribution." In: *Hydrology and Earth System Sciences* 16.3, pp. 937–951.

Gizaw, Mesgana Seyoum and Thian Yew Gan (2016). "Regional flood frequency analysis using support vector regression under historical and future climate." In: *Journal of Hydrology* 538, pp. 387–398.

Gneiting, Tilmann (2008). *Probabilistic forecasting*.

— (2011). "Making and evaluating point forecasts." In: *Journal of the American Statistical Association* 106.494, pp. 746–762.

Gneiting, Tilmann and Adrian E Raftery (2007). "Strictly proper scoring rules, prediction, and estimation." In: *Journal of the American statistical Association* 102.477, pp. 359–378.

Gottschalk, Lars et al. (1979). "Hydrologic regions in the Nordic countries." In: *Hydrology Research* 10.5, pp. 273–286.

Gräler, Benedikt et al. (2013). "Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation." In: *Hydrology and Earth System Sciences* 17.4, pp. 1281–1296.

Griffis, Veronica W and Jery R Stedinger (2007). "Evolution of flood frequency analysis with Bulletin 17." In: *Journal of Hydrologic Engineering* 12.3, pp. 283–297.

Guisan, Antoine, Thomas C Edwards Jr, and Trevor Hastie (2002). "Generalized linear and generalized additive models in studies of species distributions: setting the scene." In: *Ecological modelling* 157.2-3, pp. 89–100.

Gupta, Vijay K and Ed Waymire (1990). "Multiscaling properties of spatial rainfall andlow distributions." In: *Journal of Geophysical Research: Atmospheres* 95.D3, pp. 1999–2009.

Haddad, Khaled, Ataur Rahman, and Jery R Stedinger (2012). "Regional flood frequency analysis using Bayesian generalized least squares: a comparison between quantile and parameter regression techniques." In: *Hydrological Processes* 26.7, pp. 1008–1021.

Härdle, Wolfgang (1990). *Applied nonparametric regression*. 19. Cambridge university press.

Hastie, Trevor and Robert Tibshirani (1987). "Generalized additive models: some applications." In: *Journal of the American Statistical Association* 82.398, pp. 371–386.

He, Shaokun et al. (2022). "Multi-objective operation of cascade reservoirs based on short-term ensemble streamflow prediction." In: *Journal of Hydrology* 610, p. 127936.

Hersbach, Hans (2000). "Decomposition of the continuous ranked probability score for ensemble prediction systems." In: *Weather and Forecasting* 15.5, pp. 559–570.

Hoff, Peter D (2009). *A first course in Bayesian statistical methods*. Vol. 580. Springer.

Hosking, Jonathan Richard Morley, James R Wallis, and Eric F Wood (1985). "Estimation of the generalized extreme-value distribution by the method of probability-weighted moments." In: *Technometrics* 27.3, pp. 251–261.

Hosking, JRM and JR Wallis (1988). "The effect of intersite dependence on regional flood frequency analysis." In: *Water Resources Research* 24.4, pp. 588–600.

Hu, Lanxin et al. (2020). "Sensitivity of flood frequency analysis to data record, statistical model, and parameter estimation methods: An evaluation over the contiguous United States." In: *Journal of Flood Risk Management* 13.1, e12580.

Jarajapu, Deva Charan et al. (2022). "Design flood estimation using extreme Gradient Boosting-based on Bayesian optimization." In: *Journal of Hydrology* 613, p. 128341.

Javelle, P, JM Gresillon, and G Galea (1999). "Flood hydrological regime characterization, using a discharge-duration-frequency model." In: *Sciences de laterre et des planet* 329, pp. 39–44.

Javelle, Pierre, Taha BMJ Ouarda, and Bernard Bobée (2003). "Spring flood analysis using the flood-duration–frequency approach: application to the provinces of Quebec and Ontario, Canada." In: *Hydrological Processes* 17.18, pp. 3717–3736.

Javelle, Pierre et al. (2002). "Development of regional flood-duration–frequency curves based on the index-flood method." In: *Journal of Hydrology* 258.1-4, pp. 249–259.

Jenkinson, Arthur F (1955). "The frequency distribution of the annual maximum (or minimum) values of meteorological elements." In: *Quarterly Journal of the Royal Meteorological Society* 81.348, pp. 158–171.

Jurado, Oscar E et al. (2020). "Evaluating the performance of a max-stable process for estimating intensity-duration-frequency curves." In: *Water* 12.12, p. 3314.

Kobierska, Florian, Kolbjørn Engeland, and Thordis Thorarinsdottir (2018). "Evaluation of design flood estimates–a case study for Norway." In: *Hydrology Research* 49.2, pp. 450–465.

Laimighofer, Johannes, Michael Melcher, and Gregor Laaha (2022a). "Low-flow estimation beyond the mean–expectile loss and extreme gradient boosting for spatiotemporal low-flow prediction in Austria." In: *Hydrology and Earth System Sciences* 26.17, pp. 4553–4574.

— (2022b). "Parsimonious statistical learning models for low-flow estimation." In: *Hydrology and Earth System Sciences* 26.1, pp. 129–148.

Lima, Carlos HR et al. (2016). "A hierarchical Bayesian GEV model for improving local and regional flood quantile estimates." In: *Journal of Hydrology* 541, pp. 816–823.

Lunn, David J et al. (2000). "WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility." In: *Statistics and computing* 10, pp. 325–337.

Lussana, Cristian et al. (2019). "seNorge_2018, daily precipitation, and temperature datasets over Norway." In: *Earth System Science Data* 11.4, pp. 1531–1551.

Markiewicz, Iwona (2021). "Depth–Duration–Frequency Relationship Model of Extreme Precipitation in Flood Risk Assessment in the Upper Vistula Basin." In: *Water* 13.23, p. 3439.

Marra, Giampiero and Simon N Wood (2011). "Practical variable selection for generalized additive models." In: *Computational Statistics & Data Analysis* 55.7, pp. 2372–2387.

Martins, Eduardo S and Jery R Stedinger (2000). "Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data." In: *Water Resources Research* 36.3, pp. 737–744.

McCuen, Richard H and R Edward Beighley (2003). "Seasonal flow frequency analysis." In: *Journal of Hydrology* 279.1-4, pp. 43–56.

Merz, Ralf, Günter Blöschl, and Günter Humer (2008). "National flood discharge mapping in Austria." In: *Natural Hazards* 46, pp. 53–72.

Meyer, Mary C (2013). "A simple new algorithm for quadratic programming with applications in statistics." In: *Communications in Statistics-Simulation and Computation* 42.5, pp. 1126–1139.

Midtømme, Grethe Holm (2011). *Retningslinjer for flomberegninger 2011*. Tech. rep. 4/2011, pp. 1–66.

Mosavi, Amir, Pinar Ozturk, and Kwok-wing Chau (2018). "Flood prediction using machine learning models: Literature review." In: *Water* 10.11, p. 1536.

Msilini, Amina et al. (2022). "Flood frequency analysis at ungauged catchments with the GAM and MARS approaches in the Montreal region, Canada." In: *Canadian Water Resources Journal/Revue canadienne des ressources hydriques* 47.2-3, pp. 111–121.

Muller, Aurélie, Jean-Noël Bacro, and Michel Lang (2008). "Bayesian comparison of different rainfall depth–duration–frequency relationships." In: *Stochastic Environmental Research and Risk Assessment* 22, pp. 33–46.

Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics." In: *Handbook of markov chain monte carlo* 2.11, p. 2.

Odry, Jean and Patrick Arnaud (2017). "Comparison of flood frequency analysis methods for ungauged catchments in France." In: *Geosciences* 7.3, p. 88.

Onyutha, Charles and Patrick Willems (2015). "Empirical statistical characterization and regionalization of amplitude–duration–frequency curves for extreme peak flows in the Lake Victoria Basin, East Africa." In: *Hydrological Sciences Journal* 60.6, pp. 997–1012.

Ouali, Dhouha, Fateh Chebana, and Taha BMJ Ouarda (2016). "Quantile regression in regional frequency analysis: a better exploitation of the available information." In: *Journal of Hydrometeorology* 17.6, pp. 1869–1883.

Ouarda, Taha BMJ et al. (2006). "Data-based comparison of seasonality-based regional flood frequency methods." In: *Journal of Hydrology* 330.1-2, pp. 329–339.

Pandey, Ganesh R and V-T-V Nguyen (1999). "A comparative study of regression based methods in regional flood frequency analysis." In: *Journal of Hydrology* 225.1-2, pp. 92–101.

Pesantez, Jorge E, Emily Zechman Berglund, and Nikhil Kaza (2020). "Smart meters data for modeling and forecasting water demand at the user-level." In: *Environmental Modelling & Software* 125, p. 104633.

Prasad, Ramendra et al. (2017). "Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm." In: *Atmospheric Research* 197, pp. 42–63.

Rahman, Ataur et al. (2018). "Development of regional flood frequency analysis techniques using generalized additive models for Australia." In: *Stochastic environmental research and risk assessment* 32, pp. 123–139.

Rahman, Ayesha S, Zaved Khan, and Ataur Rahman (2020). "Application of independent component analysis in regional flood frequency analysis: Comparison between quantile regression and parameter regression techniques." In: *Journal of Hydrology* 581, p. 124372.

Renard, Benjamin et al. (2013). "Data-based comparison of frequency analysis methods: A general framework." In: *Water Resources Research* 49.2, pp. 825–843.

Renima, Mohamed et al. (2018). "Regional modelling with flood-duration-frequency approach in the middle Cheliff watershed." In: *Journal of Water and Land Development* 36.

Richardson, Sylvia and Peter J Green (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 59.4, pp. 731–792.

Robert, Christian P, George Casella, and George Casella (1999). *Monte Carlo statistical methods*. Vol. 2. Springer.

Robson, A and D Reed (1999). *Flood Estimation Handbook. Vol. 3: Statistical Procedures for Flood Frequency Estimation*. Institute of Hydrology, p. 40.

Roksvåg, Thea et al. (2021). "Consistent intensity-duration-frequency curves by post-processing of estimated Bayesian posterior quantiles." In: *Journal of Hydrology* 603, p. 127000.

Sælthun, Nils Roar et al. (1997). *Regional flomfrekvensanalyse for norsk vassdrag*. Tech. rep. NVE.

Saloranta, Tuomo (2014). *New version (v.1.1.1) of the seNorge snow model and snow maps for Norway*. Tech. rep. Norges Vassdrags og Energidirektorat (NVE).

Sherwood, James M (1994). "Estimation of volume-duration-frequency relations of ungauged small urban streams in Ohio 1." In: *JAWRA Journal of the American Water Resources Association* 30.2, pp. 261–269.

Shu, Chang and Taha BMJ Ouarda (2008). "Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system." In: *Journal of Hydrology* 349.1-2, pp. 31–43.

Stan Development Team (2023). *RStan: the R interface to Stan*. R package version 2.32.3. URL: https://mc-stan.org/.

Stein, ML (2017). "Should annual maximum temperatures follow a generalized extreme value distribution?" In: *Biometrika* 104.1, pp. 1–16.

Tarquis, AM et al. (2011). "Preface" Nonlinear and scaling processes in Hydrology and Soil Science"." In: *Nonlinear processes in geophysics* 18.6, pp. 899–902.

Tellman, Beth et al. (2021). "Satellite imaging reveals increased proportion of population exposed to floods." In: *Nature* 596.7870, pp. 80–86.

Thorarinsdottir, Thordis L, Tilmann Gneiting, and Nadine Gissibl (2013). "Using proper divergence functions to evaluate climate models." In: *SIAM/ASA Journal on Uncertainty Quantification* 1.1, pp. 522–534.

Thorarinsdottir, Thordis L et al. (2018). "Bayesian regional flood frequency analysis for large catchments." In: *Water Resources Research* 54.9, pp. 6929–6947.

Thorarinsdottir, Thordis L et al. (2020). "Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods." In: *Environmental Research Letters* 15.12, p. 124041.

Tyralis, Hristos, Hellenic Air Force, and Andreas Langousis (2019). "Estimation of intensity-duration-frequency curves using max-stable processes." In: January. DOI: 10.1007/s00477-018-1577-2.

Ulrich, Jana, Felix S Fauer, and Henning W Rust (2021). "Modeling seasonal variations of extreme rainfall on different timescales in Germany." In: *Hydrology and Earth System Sciences* 25.12, pp. 6133–6149.

Ulrich, Jana et al. (2020). "Estimating IDF curves consistently over durations with spatial covariates." In: *Water* 12.11, p. 3119.

Van de Vyver, Hans (2018). "A multiscaling-based intensity–duration–frequency model for extreme precipitation." In: *Hydrological Processes* 32.11, pp. 1635–1647.

Van Loenhout, J, R Below, and D McClean (2020). *The human cost of disasters: an overview of the last 20 years (2000–2019)*. Tech. rep. Tech. rep., Centre for Research on the Epidemiology of Disasters (CRED) and ...

Vormoor, Klaus et al. (2016). "Evidence for changes in the magnitude and frequency of observed rainfall vs. snowmelt driven floods in Norway." In: *Journal of Hydrology* 538, pp. 33–48.

Wang, Zhaoli et al. (2015). "Flood hazard risk assessment model based on random forest." In: *Journal of Hydrology* 527, pp. 1130–1141.

Wilson, Donna et al. (2011). *A review of NVE's flood frequency estimation procedures*. Tech. rep. Norges vassdrags -og energidirektorat.

Wood, S.N (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman and Hall/CRC.

Yee, Thomas W and Neil D Mitchell (1991). "Generalized additive models in plant ecology." In: *Journal of vegetation science* 2.5, pp. 587–602.

Zaidman, Maxine D et al. (2003). "Flow-duration-frequency behaviour of British rivers based on annual minima data." In: *Journal of hydrology* 277.3-4, pp. 195–213.

Zaman, Mohammad A, Ataur Rahman, and Khaled Haddad (2012). "Regional flood frequency analysis in arid regions: A case study for Australia." In: *Journal of Hydrology* 475, pp. 74–83.

Paper I

Flexible and consistent Flood-Duration-Frequency modeling: A Bayesian approach

Journal of Hydrology 620 (2023) 129448

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Flexible and consistent Flood–Duration–Frequency modeling: A Bayesian approach

Danielle M. Barna [a,c,*], Kolbjørn Engeland [a], Thordis L. Thorarinsdottir [b], Chong-Yu Xu [c]

[a] Norwegian Water Resources and Energy Directorate, P.O. Box 5091 Majorstua, NO-0301 Oslo, Norway
[b] Norwegian Computing Centre, P.O. Box 114 Blindern, NO-314 Oslo, Norway
[c] Department of Geosciences, University of Oslo, P.O. Box 1047 Blindern, NO-0316 Oslo, Norway

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Dr-Ing, Editor-in-Chief, with the assistance of Zhenxing Zhang, Associate Editor.

Dataset link: https://doi.org/10.5281/zenodo.7085557

Keywords:
Flood frequency analysis
Design flood level
Flood–Duration–Frequency models
Generalized extreme value distribution
Bayesian statistics

ABSTRACT

Design flood values give estimates of flood magnitude within a given return period and are essential to making adaptive decisions around land use planning, infrastructure design, and disaster mitigation. Many hydrologic applications where flood retention is important, e.g. floodplain management and reservoir design, need design flood values for different durations. Flood–Duration–Frequency (QDF) models extend the standard statistical flood frequency analysis framework to multiple flood durations and are analogous to intensity–duration–frequency models for precipitation. Implementations of QDF models commonly assume simple scaling, where only the magnitude of the index flood is assumed to change with duration, despite empirical analyses showing a more complex dependence structure. We propose a multiscaling extension to existing QDF models where the magnitude of the index flood and the slope of the growth curve may scale independently with duration. In an application to 12 locations in Norway, we assess how three different QDF models capture relationships between floods of different duration. Incorporating duration dependency independently in both the index flood and the growth curve (extended QDF model) improves modeling of both short-duration events and events with long return periods. This model extension further expands the models' ability to simultaneously model a wide range of durations. As measured by the integrated quadratic distance, the extended QDF model performs better than the original QDF model in 83% of the out of sample subdaily durations studied. Additionally, we find that the choice of durations used to fit QDF models is a highly influential aspect of the modeling process.

1. Introduction

Floods are a widespread and costly threat to society worldwide. Their destructive capacity is likely to increase in the near future due to a rise in both the prevalence of floods under climate change and an increase in the economic value of flood-prone areas (Alfieri et al., 2017; Field et al., 2012). Estimation of design floods is an important aspect of societal adaptation to increased flood risk. Such estimation can be undertaken in one of three general ways, e.g. Filipova et al. (2019): (1) statistical flood frequency analysis (FFA), where observed historical flood events are used to estimate the magnitude of flood events with a certain return period, (2) event-based hydrological modeling for a single design event, where design rainfall or other single realizations of initial conditions and precipitation are used as input to a hydrological model that simulates the desired flood event and (3) derived flood frequency methods, which use weather generators coupled with hydrologic models to simulate long series of synthetic discharge that can be used to statistically estimate the desired return periods. The first approach—statistical FFA—is the focus of this paper.

Flood–Duration–Frequency (QDF) models extend the standard statistical FFA framework to multiple flood durations and are analogous to Intensity–Duration–Frequency (IDF) models for precipitation. Many hydrologic applications where flood retention is important, e.g. floodplain management and reservoir design, need flood estimates for different durations. Typically, the annual maxima used in QDF modeling are sampled from discharge series averaged over different durations (Javelle et al., 2002; Cunderlik and Ouarda, 2006). This means that the duration d represents the total flow volume for a time span of d hours, not flood events that lasted precisely d hours. This aggregation-based approach to obtaining annual maxima means QDF models provide an accessible way to get relationships between total flow volumes and durations for applications where the total volume of water is of interest.

In the QDF approach, an extreme value distribution (usually the generalized extreme value, or GEV, distribution) is fit to annual maxima from different durations. Then the relationship between the durations

* Corresponding author at: Norwegian Water Resources and Energy Directorate, P.O. Box 5091 Majorstua, NO-0301 Oslo, Norway.
E-mail address: daba@nve.no (D.M. Barna).

https://doi.org/10.1016/j.jhydrol.2023.129448
Received 14 December 2022; Received in revised form 17 March 2023; Accepted 23 March 2023
Available online 31 March 2023

and the fitted distributions is described by the QDF model. This allows for the quantiles of the distribution to be parametrically expressed as a continuous formulation of both return period and duration, where consistency between the quantiles of the distribution at different durations is enforced by the QDF model (Javelle et al., 2002). In practice this means that, for example, the T-year flood for the mean daily streamflow time series will never report a higher return level than the T-year flood for the instantaneous streamflow time series (where T describes the return period of the flood). Such consistency is not guaranteed when estimating extreme value distributions individually for several fixed durations and remains one of the main benefits of QDF modeling in situations where the return level at several durations is of interest. In addition, the parametric nature of the QDF model allows for extrapolation to unobserved durations and establishes the potential for prediction in ungauged basins (Javelle et al., 2002).

The foundations of QDF modeling were developed in the 1990s through analysis of the relationships between n-day flood volumes as explored in Balocki and Burges (1994) and Sherwood (1994). The original QDF model is generally attributed to Javelle et al. (1999). QDF modeling has found most of its application in France, Canada and Britain in the early 2000s (Javelle et al., 2002, 2003; Zaidman et al., 2003) although it has been applied a handful of times in the decades since (Cunderlik et al., 2007; Crochet, 2012; Onyutha and Willems, 2015). In a guide to hydrological practices, the (World Meteorological Organization, 2009) notes that QDF analysis remains under-utilized despite its strong potential.

In more recent years, the QDF model has been used to characterize flood events of different duration in Algeria (Renima et al., 2018), to inform development of a depth–duration–frequency relationship used to assess risk of rainfall-driven floods in Poland Markiewicz (2021) and as a comparison point to IDF models when assessing catchment behavior for runoff extremes in Austria (Breinl et al., 2021). As noted in Breinl et al. (2021), the relationship quantified by the QDF model is an analogue to the relationship quantified in IDF modeling for precipitation extremes: in the hypothetical situation where all rainfall becomes runoff and the time of concentration is instantaneous, the QDF and IDF models have identical relationships.

Available QDF models usually assume that only the index flood changes with duration, with the growth curve assumed constant across durations (e.g. Javelle et al., 2002; Cunderlik and Ouarda, 2006; Breinl et al., 2021). Here the index flood is the median annual maximum flood. The *growth curve* is a scaled version of the flood frequency curve created by taking the ratio of the flood of any frequency to the index flood (Robson and Reed, 1999). The multiplication of the growth curve and the index flood gives the flood frequency curve. We find it useful to discuss the flood frequency curve in terms of index floods and growth curves for a few reasons. First, it clarifies the discussion around an established problem with QDF models. Second, the concept of the flood frequency curve as an index flood and a growth curve fits with the reparameterization introduced in Section 3. Third, this language and reparameterization of the flood frequency curve aligns with regionalization methods; note that growth curves are presented in Dalrymple (1960) as "basic, dimensionless frequency curves" allowing for cross catchment comparisons.

This assumption of constant growth curve across durations contradicts empirical analyses of runoff scaling properties in Norway that show the ratio between peak and daily floods may be dependent on return period (Engeland et al., 2020; Sælthun et al., 1997). "Multiscaling" models that allow for this behavior—that is, models that allow for the ratio between growth curves of different durations to be dependent on return period—already exist in the IDF literature (Van de Vyver, 2018; Courty et al., 2019; Fauer et al., 2021). However, in all existing models the different scaling components are placed on the location and the scale parameter of the GEV distribution, respectively. This hinders a direct interpretation in terms of scaling of the index flood on the one hand and the growth curve on the other hand.

Here, we propose a multiscaling extension of the QDF model of Javelle et al. (2002), where the magnitude of the index flood and the slope of the growth curve may scale independently with duration.

The natural sparsity of available extreme value data means parameter estimation is, in general, challenging for extreme value models (Scarrott and MacDonald, 2012). The additional parameters introduced by multiscaling models compound these challenges (Fauer et al., 2021). We introduce an alternative parameterization of what we call the *characteristic duration* parameters that allows for more numerical stability. In addition, we adopt a Bayesian estimation approach that allows for all parameters to be estimated concurrently. Bayesian estimation of IDF models is well established and provides advantages such as accessible uncertainty assessments, scaling to regional models via hierarchical Bayesian approaches, and the ability to add information through prior distributions have been shown to be relevant (Cheng and AghaKouchak, 2014; Huard et al., 2010). Current QDF models are typically estimated in a two-step procedure where the characteristic duration parameter is estimated first, followed by an estimation of the remaining parameters (Javelle et al., 2002; Cunderlik et al., 2007). However, such two-step estimation does not typically provide uncertainty information, is difficult to use with multiscaling models, and, moreover, requires additional assumptions if the model is to be used in a regional context (Cunderlik and Ouarda, 2006).

Design flood estimation is often most concerned with estimation of peak discharge. In this case, a statistical estimation poses a challenge since flood series of length appropriate for statistical FFA often contain segments at a daily—or coarser—time resolution. This is dealt with in practice as a data quality issue; most national guidelines for FFA outline detailed data quality control steps and recommend application of FFA only when fine resolution time series of suitable length exist, or when catchment properties are such that daily data can be trusted to provide a representative profile of the flood peak (Ball et al., 2019; England et al., 2019; Castellarin et al., 2012). In the situation where we have neither fine resolution time series nor catchment properties that allow for construction of the flood peak from daily data, there exist methodologies for scaling daily data to approximate the instantaneous peak flow (Ding et al., 2015; Fill and Steiner, 2003).

In Norway, scaling between daily and instantaneous peak flows is performed by establishing a relationship between the daily flows and the instantaneous peak flows for the largest floods in the catchment. In the case where no data is available, the relationship can be provided by a hydrologically similar catchment. Wilson et al. (2011) notes the uncertainty in this method is likely to be large and difficult to reconcile with the uncertainty inherent to FFA. Therefore, it is of interest to investigate the skill of QDF models to predict floods at subdaily unobserved durations based on their parametric assumptions and available coarser-time-resolution data at the site of interest.

To summarize, the main objective of this study is to assess how different QDF models capture relationships between floods of different duration. In particular we want to answer the following questions: (i) is there one QDF model that best captures flood behavior at the shortest (sub-daily) durations? (ii) what are the models' abilities when estimating in sample and out of sample durations? and (iii) how sensitive are QDF models to input durations? To this aim, we evaluate three different models, one of which is the original QDF model as presented in Javelle et al. (2002). The other two models investigated are new QDF models that allow for differing degrees of duration dependency in the growth curve. For comparison, three-parameter GEV distributions are fit independently to each duration in line with the current guidelines (Midtømme, 2011; England et al., 2019).

The remainder of the paper is organized as follows: Section 2 introduces the data and describes several data artifacts unique to QDF modeling. Section 3 presents the three QDF models investigated in this study and details both the Bayesian framework and Markov chain Monte Carlo (MCMC) sampling. To facilitate both interpretation and inference, a quantile-based reparameterization of the GEV distribution is proposed. Section 4 describes QDF model behavior and assesses performance in relation to locally fit GEV distributions. The paper finishes with a discussion (Section 5) and conclusions (Section 6).

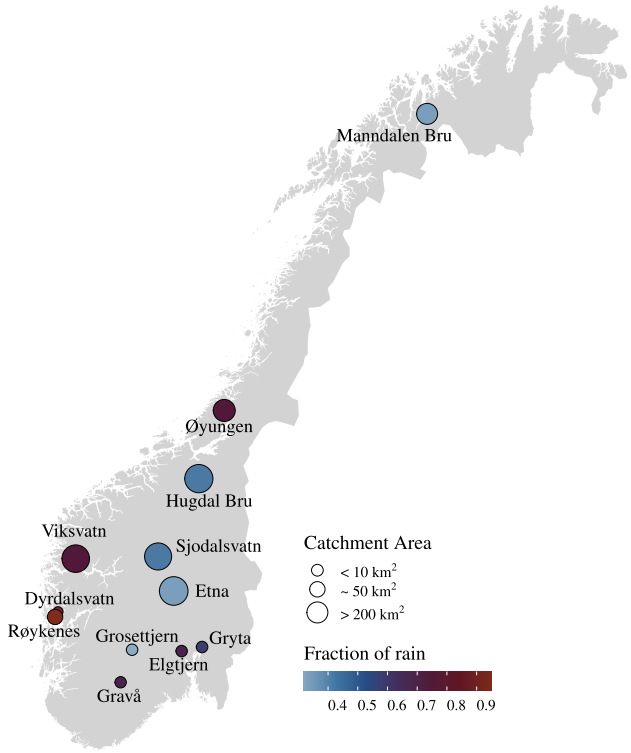

Fig. 1. Locations of twelve gauging stations used in study. Catchment area and fraction of rain contribution to flood are also indicated.

2. Data

The flood data came from 12 streamflow stations in Norway that have at least 28 years of quality-controlled data with minimal influence from reservoirs and other installations that might alter the natural streamflow. See Engeland et al. (2016) for details. All streamflow data were taken from the Norwegian hydrological database Hydra II hosted by the Norwegian Water Resources and Energy Directorate (NVE).

The locations of the gauging stations, as well as catchment areas and flood generating processes, are shown in Fig. 1. The selected stations are diverse relative to Nordic catchments, allowing us to evaluate the QDF models on a variety of flood behaviors. See Table D.7 for a listing of selected catchment properties. The catchment size ranges from 6.31 km^2 (Gravå) to 570 km^2 (Etna). In Norway the two major flood generating processes are snowmelt and rain. In Fig. 1 this is illustrated as the average fractional rain contribution to each flood event. The average rainfall contribution was estimated by calculating the ratio of rainfall to total water depth from both rainfall and snowmelt accumulated in a time window prior to each flood and then averaging these ratios over all flood events. For details see Engeland et al. (2020). A fraction of rain value close to one means the floods at this location are primarily driven by rain; a value closer to zero means snowmelt is the dominant flood-generating mechanism. Rain was calculated from the precipitation and temperature from SeNorge 2.0 dataset (Lussana et al., 2019). Snow melt was extracted from the SeNorge snow model (Saloranta, 2014). In our dataset the rain contribution varies from 0.32 at Grosettjern to 0.95 at Røykenes.

2.1. Data quality control

Each of the streamflow records encompasses a variety of collection methods. These differing collection methods provide data at different frequencies. Typically we find daily time resolution in the first part of a streamflow record and a higher frequency of measurements in the latter part of the streamflow record after adoption of digitized limnigraph records and/or digital measurements.

It is necessary to make sure that the sampling frequency of the data is high enough to represent peak flood magnitudes with sufficient quality. This is especially important at small catchments; a higher frequency of measurements is needed to capture the behavior of quicker, "flashier" floods vs slower, smoother floods. In the records for the smallest catchments, this constraint excludes substantial parts with a daily sampling frequency. For two large, primarily snowmelt driven catchments–Etna and Viskvatn–we used the daily data in addition to the more high-resolution data. The daily data was collected beginning in 1920 for Etna and in 1903 for Viksvatn. The high-resolution data was collected from 1983–2022 for Etna and from 1985–2022 for Viksvatn. For all the remaining stations we used data from approximately 1970 to 2022, which is collected via a combination of limnigraph and digital readings. Precise record lengths can be found in Table D.7. The time resolution of the digital measurements and the digitization of the limnigraph records were selected by NVE to be frequent enough to represent flood peaks at individual stations.

In addition to quality control on the sampling frequency, the data have already undergone a detailed quality control by the hydrometric section at NVE. Ice jams are an issue at many stations in Norway and may influence the validity of the rating curves used to calculate streamflows from measured water levels. When needed, specific correction procedures (as specified in internal quality assurance protocols at NVE) have been applied to get correct discharge. Any year with less than 300 days of data was discarded. The final data-set contains no extraordinary flood events as seen in Appendix E.

2.2. Data processing for QDF

The data set for the QDF analysis is constructed from an evenly spaced streamflow time series at the reference duration, where the reference duration is the finest time resolution of interest. Even spacing in the reference duration is enforced via regular sampling of a linear interpolation of the observed data.

Let $x_0(\tau)$ be this time series at the reference duration. A moving average of window length d was applied to $x_0(\tau)$ to manufacture a new time series, $x_d(t)$:

$$x_d(t) = \frac{1}{d}\int_{t-d/2}^{t+d/2} x_0(\tau)\, d\tau \tag{1}$$

Block maxima or peak over threshold values can then be extracted from $x_d(t)$ to form sets of maxima given as:

$$\{Q_{d,1}, Q_{d,2}, \ldots, Q_{d,k}\} \tag{2}$$

where, in the case of annual maxima, k is the number of years of data. The width d used as the length of the averaging window corresponds to the duration of interest and the average in Eqn (1) can be repeatedly applied under different d to manufacture new sets of maxima that correspond to different durations of interest. Under this aggregation approach, the durations d represent the total volume of water that arrives over a time span of d hours, not flood events that lasted precisely d hours.

These sets of maxima produced under different d are dependent; that is, since longer duration series are always aggregated from series of shorter duration, the values in one set of maxima depend on the values in the other sets. Recent advances in IDF have focused on use of multivariate extreme value theory models, which explicitly model this dependence structure between sets of maxima (Jurado et al., 2020). The QDF models presented in this study are simpler, so-called "univariate extreme value theory models" and do not account for this dependence structure. We use Fig. 2 to justify the choice of the simpler model.

The aggregation to total flow volume over a time span of duration d described in Eq. (1) introduces a dependency structure that is neither predictable nor directly relatable to catchment properties. Fig. 2

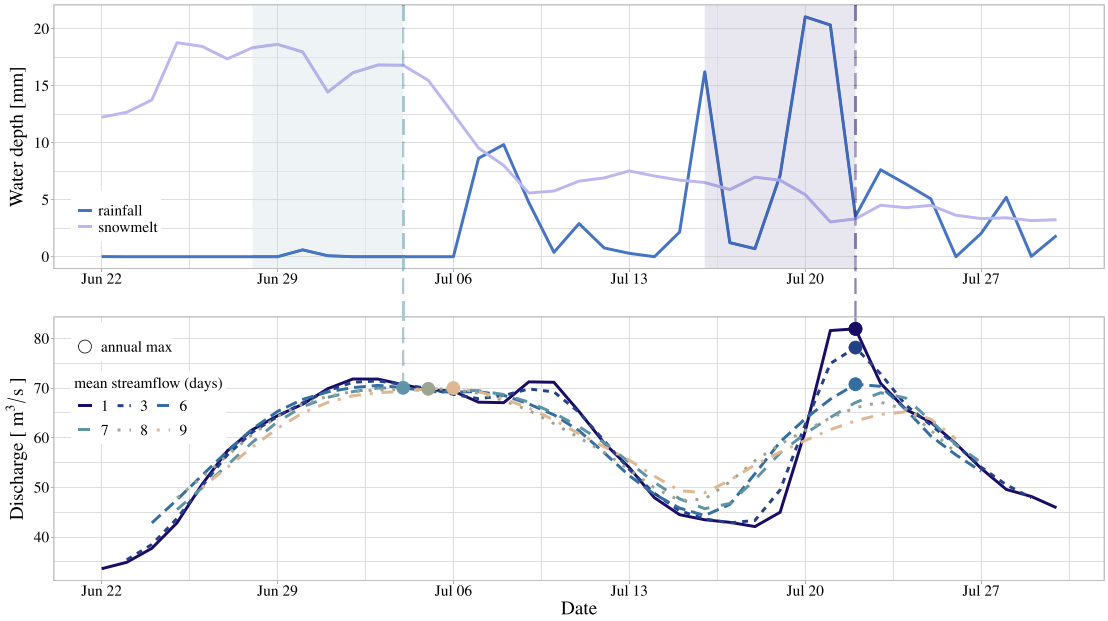

Fig. 2. In QDF modeling, the duration d represents the total flow volume for a time span of d hours, not flood events that lasted precisely d hours. This means longer duration series are always aggregated from series of shorter durations. This creates a dependency structure that is artificial yet not easily modeled. There are two reasons why this dependency structure is not easily modeled, both of which are illustrated in this figure: (i) annual maxima for each duration are not always primarily issued from the same flood event. In some cases, these flood events can have completely different generating processes (top panel; the shaded areas show the window of time from which the flood generating process is calculated) and (ii) annual maxima are not guaranteed to decrease as the duration of the averaging window is increased (see annual maxima at 7 days or greater). Data is from Sjodalsvatn gauging station, for the year 2009.

demonstrates this. First, annual maxima for different durations are in some cases primarily issued from the same flood event; however, in other cases the maxima at different durations are based on different flood events with potentially different flood generating processes. In the first scenario the annual maxima have a strong dependency due to overlapping temporal support and serial correlation. In the second there is weak dependency. This presence or absence of this change in across duration correlation is not directly relatable to catchment properties. Second, annual maxima are not guaranteed to decrease as the duration of the averaging window is increased and the circumstances that produce this inconsistent behavior in maxima (for example, two flood events of similar volume occurring within a short time period of each other, or a particularly wide and flat-topped flood event) are also not directly relatable to catchment properties.

3. Methods

Extreme value theory allows for the estimation of extreme events by providing a framework for modeling the tail of probability distributions where such extreme events would lie. Let $X_1, \ldots, X_n$ be a set of continuous, univariate random variables that are assumed to be independent and identically distributed. If the normalized distribution of the maximum $\max\{X_1, \ldots, X_n\}$ converges as $n \to \infty$ then it converges to a GEV distribution (Fisher and Tippett, 1928; Jenkinson, 1955). See Coles (2001) for further details.

In flood frequency analysis the set of values that is taken to be distributed GEV is typically the set of annual maxima. The GEV distribution is governed by a location, scale and shape parameter. The special case where the shape parameter is equal to zero is termed the Gumbel, or two-parameter, distribution. Both distributions are used in European FFA and an overview of country specific application can be found in Castellarin et al. (2012). Previous research (Castellarin et al., 2012; Midtømme, 2011; Kobierska et al., 2018) recommends the three-parameter GEV distribution for FFA on individual Norwegian stations. The following QDF models are thus based in the three-parameter form of the GEV, where the cumulative distribution function of the GEV is given as

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{3}$$

which is defined on $\{z : 1+\xi(z-\mu)/\sigma > 0\}$ with parameter bounds $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$ and where z would be the observed annual maximum streamflow for duration d for a specific year. The case where the shape parameter, ξ, is equal to zero is interpreted as the limit when $\xi \to 0$.

The remainder of this section is organized as follows: first, a quantile-based reparameterization of GEV distribution is adopted. Then three different QDF models–one established model and two new models–are introduced under this reparameterization. Finally, the fitting methodologies and model evaluation metrics are described.

3.1. Reparameterization of the GEV distribution

The parameters of a GEV model are most easily interpreted in terms of the quantile expressions; traditional descriptors such as the mean and variance are inappropriate for the skewed distribution of the GEV and, moreover, are undefined for certain values of the ξ parameter (Coles, 2001). We reparametrize the GEV distribution using the $\alpha = 0.5$ quantile in line with the recent work of Castro-Camilo et al. (2022). The relationship between the location parameter, μ, and the location parameter under the reparameterization, η (i.e. the median flood), is given as

$$\eta = \begin{cases} \mu + \sigma\frac{\log(2)^{-\xi}-1}{\xi} & \text{if } \xi \neq 0 \\ \mu - \log(\log(2)) & \text{if } \xi = 0. \end{cases} \tag{4}$$

Estimates of extreme quantiles are obtained by substituting η from Eq. (4) for μ in Eq. (3) and inverting the result, giving

$$z_p = \eta + \sigma\left\{\frac{(-\log(1-p))^{-\xi} - \log(2)^{-\xi}}{\xi}\right\}. \tag{5}$$

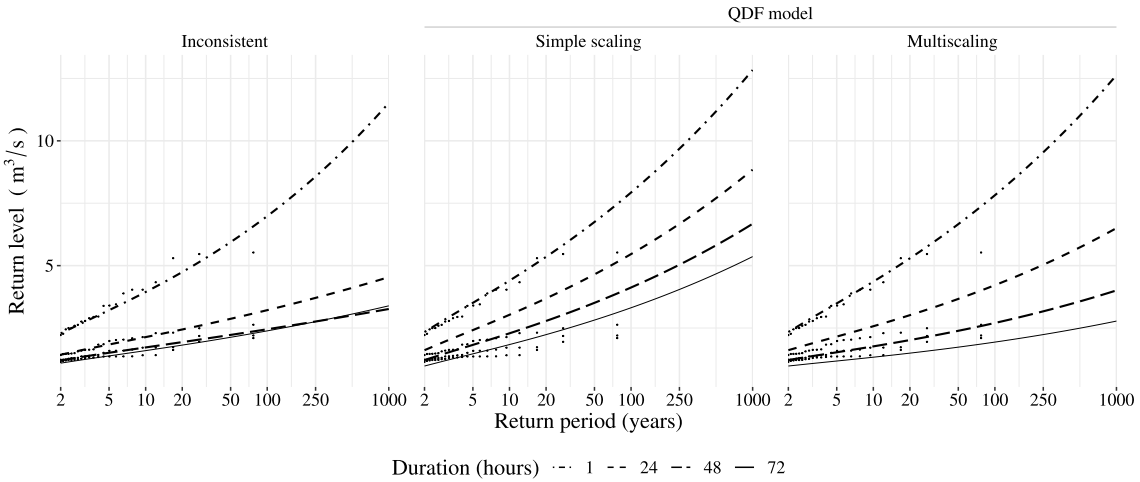

Fig. 3. Return level plots from a synthetic data set showing (i) flood frequency curves estimated independently for four durations (left panel), (ii) output from a simple scaling QDF model (middle panel), and (iii) output from a multiscaling QDF model (right panel). The independent fits do not account for duration dependency. The simple scaling model accounts for duration dependence in the magnitude of the index flood but not the growth curve. The multiscaling model accounts for duration dependence in both the magnitude of the index flood and the slope of the growth curve.

Here, $G(z_p) = 1 - p$ and z_p is the return level associated with the return period T such that $T = 1/p$. Finally, to reduce dependency between parameters, the scale parameter is decomposed as a product of the median flood and a remainder term expressed as an exponential function, e^{β}, such that the new scale parameter β is given as

$$\beta = \log\left(\frac{\sigma}{\eta}\right). \tag{6}$$

The location parameter η has a more reasonable interpretation under the reparameterization in Eq. (5): it is now the median of the GEV distribution, with units of m^3/s. Consequently, it is much easier to choose informative priors under the reparameterization—an important advantage in a Bayesian framework (Gelman et al., 2013).

In addition to providing interpretable parameters, this parameterization has the added benefit of aligning with the index flood approach popular in regional flood frequency modeling, where the median flood for a group of catchments is taken as a typical, or "index", flood (Dalrymple, 1960). Explicitly including the median as a parameter in the model means the order of magnitude of a flood can be separated from the shape and slope of the growth curve. This has potential to simplify the search for regressors in a regional QDF model (Castro-Camilo et al., 2022).

3.2. Models

This section discusses three competing models. First the original QDF model from (Javelle et al., 2002) is presented under the reparameterization in Section 3.1. Then the new extended QDF model is introduced. Finally, a mixture model taking components from both previous models is introduced. Each of these models introduces additional parameters to the classic GEV model. The models differ in the number of additional parameters added, but can all be classified as *duration-dependent GEV*, or d-GEV, models.

We motivate the development of the extended, multiscaling QDF model with Fig. 3.

The leftmost panel of the figure shows several flood frequency curves estimated independently for four durations. The curves for 48 and 72 h are inconsistent; that is, the 72 h frequency curve crosses the 48 h curve. Physically, there should not be a larger total volume of water during a 48 h interval than a 72 h interval. These inconsistencies can arise when we ignore duration dependence in both the index flood and growth curve—that is, when we estimate durations independently.

QDF models enforce consistency between flood frequency curves of different duration, as the middle and right panels of Fig. 3 show. Existing QDF models account for duration dependence in the index flood but not the growth curve. This is termed "simple scaling" and is illustrated in the middle panel of Fig. 3. However, ignoring the effect of duration dependency on the growth curve can lead to poor estimation in the tails of the distribution. Models that account for duration dependency in both the index flood and growth curve are called "multiscaling" models. The extended QDF model accounts for duration dependency in the growth curve by allowing the both the magnitude of the index flood and the slope of the growth curve to change with duration (right panel, Fig. 3).

3.2.1. Original QDF model

The annual maxima under the original QDF model proposed in Javelle et al. (2002) are independently distributed

$$Q_{d,i} \sim \text{GEV}\left(\eta_d, \beta, \xi\right) \tag{7}$$

where

$$\eta_d = \eta\,(1 + d\Delta)^{-1} \tag{8}$$

and the quantile function under the reparameterization in Section 3.1 is given as

$$z_{d,p} = \frac{\eta}{1 + d\Delta}\left[1 + e^{\beta}\left\{\frac{(-\log(1-p))^{-\xi} - \log(2)^{-\xi}}{\xi}\right\}\right] \tag{9}$$

where $\Delta > 0$. Note the inverse of the characteristic duration parameter Δ from Javelle's original QDF model is used here for numerical stability during estimation. A high value for Δ indicates the total flow volume arrives quickly, analogous to a flashy/peaked hydrograph with a pronounced duration dependency for the median flood, whereas a value close to zero indicates a slower timespan, analogous to a wide hydrograph with minor duration dependency for the floods. The traditional flood frequency curve–that is, a GEV distribution fit to an instantaneous time series–is recovered in the limit of the aggregation window as $d \to 0$.

In Javelle's model only η is dependent on d and Δ. This aligns with the literature base for IDF modeling in the sense that the model can be written as a separable function of d and p. Notice further that if the $1 + d\Delta$ quantity in Eq. (9) was replaced with a power relationship the model would match that of the IDF models summarized in Koutsoyiannis et al. (1998). The power relationship and separable functional dependence of the IDF model has its roots in stochastic process theory, although the model as typically applied does not rely on this theory base since

IDF models do not attempt to make explicit mathematical statements about how the higher order moments (e.g., variance) change with duration (Koutsoyiannis et al., 1998).

Since only the magnitude of the median flood (η) is duration-dependent in the model in Eq. (9), the underlying assumption of the original QDF model is that the slope of the growth curve does not change with duration.

3.2.2. Extended QDF model

The extended QDF model (referred to as the *Double-Delta* QDF model) is structured to be able to capture differences in slope of the growth curves coming from peak and daily values, or, indeed, values coming from any two different aggregation intervals. Changing the steepness of the growth curve dependent on duration requires extra flexibility in the tail behavior of the model, so the model allows η and β to depend on the aggregation interval d and additional parameters Δ_1 and Δ_2, respectively. The ξ parameter is kept duration-invariant due to the difficulties in estimating the ξ parameter stemming from the involved parametric form of the CDF (Eq. (3)). Under Double-Delta the annual maxima are independently distributed as

$$Q_{d,i} \sim \text{GEV}\left(\eta_d, \beta_d, \xi\right) \tag{10}$$

where

$$\eta_d = \eta\left(1 + d\Delta_1\right)^{-1} \tag{11}$$

$$\beta_d = \log\left(\frac{\sigma}{\eta_d(1 + d\Delta_2)}\right) \tag{12}$$

and the distribution's quantiles for a duration d corresponding to exceedance probability p are given by

$$z_{d,p} = \frac{\eta}{1 + d\Delta_1}\left[1 + \frac{e^{\beta}}{1 + d\Delta_2}\left\{\frac{(-\log(1-p))^{-\xi} - \log(2)^{-\xi}}{\xi}\right\}\right] \tag{13}$$

with constraint

$$0 < \Delta_2 < \Delta_1. \tag{14}$$

The constraint on the Delta parameters reflects the fact that the data aggregation performed in QDF modeling (see Section 2.2) is more likely to have a larger effect on the flood magnitude than on the decomposed scale parameter. Recall that the value of the Δ_1 parameter reflects the "flashiness" of the floods measured; a narrow hydrograph will be associated with larger values of Δ_1. The Δ_2 parameter does not have an equally accessible hydrologic interpretation but can be interpreted as a measure of difference in growth curve slope across aggregation intervals; that is, if the ratio between peak and daily floods is heavily dependent on return period we would expect to see larger values of Δ_2.

As the aggregation window shrinks to zero, that is, as $d \to 0$, the Double-Delta model is equivalent to the standard GEV model that creates the traditional flood frequency curve. Similarly, as $\Delta_2 \to 0$, the Double-Delta model approaches Javelle's QDF model. Double-Delta can thus be considered an extension of Javelle in the same way Javelle is an extension of the traditional flood frequency curve.

3.2.3. Mixture model

The mixture model is proposed in an attempt to access the flexibility of the Double-Delta model without adding unnecessary complexity. The model is a weighted average of the Double-Delta and Javelle models such that the density of the annual maxima is given by

$$\sum_{j=1}^{2} m_j\, g(\cdot|\theta_j) \tag{15}$$

where m_j is the weight on the component model, g is the density of the GEV distribution, $\theta_1 = \{\eta_d^{\text{DD}}, \beta_d^{\text{DD}}, \xi^{\text{DD}}\}$ and $\theta_2 = \{\eta_d^{\text{J}}, \beta^{\text{J}}, \xi^{\text{J}}\}$. Here the superscripts on the parameter sets denote the Double-Delta and Javelle models, respectively. Using Bayesian methodologies and the reversible-jump algorithm detailed in Section 3.3, parameter estimation and selection can be carried out simultaneously and the Δ_2 parameter is only added if merited.

Thus Eq. (15) is a representation of a non-standard density from which it is possible to obtain quantile estimates that are an average over the distributions given by the Double-Delta model in Eq. (10) and the Javelle model in Eq. (7).

3.3. Bayesian framework

For the Javelle and Double-Delta models, Bayesian inference is performed using a Metropolis-Within-Gibbs algorithm (Robert and Casella, 2004). That is, samples from the conditional distribution of the parameters θ_1 and θ_2, respectively, are obtained by iterative sampling from the full conditional distributions of the individual parameters so that each component of the model is updated in turn. Prior distributions for the individual parameters assume independence. The prior on η, which has units of m^3/s, is a diffuse truncated normal distribution truncNormal(40,100) with lower bound at zero. The prior on β is a diffuse Normal(0,100). For ξ, we follow the methodology in Martins and Stedinger (2000) and use a shifted Beta(6,9) distribution on the interval $[-0.5, 0.5]$. The prior for Δ_1 in the Double-Delta model, which is equivalent to the prior for Δ in the Javelle model, is a Lognormal(0,5). The same values are used in the prior for Δ_2, which uses a truncated Lognormal where the lower bound of the prior is given by Δ_1.

The conditional distribution of the mixture model is given by

$$p(m, \theta|\mathbf{Q}) \propto p(m)p(\theta|m)\, g(\mathbf{Q}|\theta, m) \tag{16}$$

where $p(\cdot|\cdot)$ is the generic conditional distribution consistent with this joint specification and $m \in \{\text{DD}, \text{J}\}$, $\theta \in \{\theta_1, \theta_2\}$, and $\mathbf{Q} = (Q_{d,i})_{i=1,d=1}^{i=k,d=n}$, where k is the number of years of data and n is the total number of durations. The models have equal prior probability, with $p(m = J) = p(m = DD) = 0.5$. Simplification of Eq. (16), considering the model without the model specification and separate parameter sets, gives the conditional distributions of Double-Delta and Javelle.

Moving between models changes the dimension of θ. To account for this, we employ a reversible jump MCMC algorithm, similar to the reversible jump methodology for normal mixtures described in Richardson and Green (1997). The reversible jump MCMC proceeds as follows:

1. updating θ:
   (a) if m = DD update η^{DD}, else update η^{J};
   (b) if m = DD update β^{DD}, else update β^{J};
   (c) if m = DD update ξ^{DD}, else update ξ^{J};
   (d) if m = DD update Δ_1 and Δ_2 parameters in sequence, else update Δ;
2. splitting one Delta into two, or combining two Deltas into one.

Step 1 is repeated 10 times under the same model before Step 2 (proposal to jump between models) is taken. Repeating Step 1 for either the Javelle or Double-Delta model details the MCMC algorithm used to fit the respective model. To move from Double-Delta to Javelle we need to merge Δ_1 and Δ_2 into one Δ. The combine proposal is deterministic and given by

$$\Delta = \Delta_1 + \Delta_2. \tag{17}$$

The reverse split proposal, going from Javelle to Double-Delta, involves one degree of freedom, so we generate a random variable u such that

$$u \sim Beta(5, 1) \tag{18}$$

which is then used to set

$$\begin{aligned} \Delta_1 &= u\Delta \\ \Delta_2 &= (1-u)\Delta. \end{aligned} \tag{19}$$

For this split move the acceptance probability is min $\{1, A\}$ where

$$A = \frac{p(m', \theta'|\mathbf{Q})}{p(m, \theta|\mathbf{Q})q(u)} |J| \tag{20}$$

where $q(u)$ is the density function of u and J is the Jacobian of the transformation described in Eq. (19). The acceptance probability for the corresponding combine move is min $\{1, A^{-1}\}$ but with substitutions that adhere to the proposal in Eq. (17).

3.3.1. Posterior return levels

The Markov chains detailed above return a collection of R samples

$$\theta^{[r]}, \quad r = 1, \ldots, R \tag{21}$$

where R is the total number of iterations in the MCMC with a suitable number of burn-in samples removed. Under the mixture model, θ can be either θ_1 or θ_2 dependent on iteration r, while posterior samples under Double-Delta or Javelle will return only θ_1 or θ_2, respectively. This Markov sample of the parameter set directly yields, by using the quantile function in either (9) or (13), a sample of quantiles

$$\left\{(z_{d,p})^{[1]}, \ldots, (z_{d,p})^{[R]}\right\}. \tag{22}$$

This sample approximates the posterior distribution of the pth return level at duration d. From this sample it is possible to derive approximations for the posterior mean and its credible intervals.

3.4. Evaluation methods

To assess the models we compare QDF model output to GEV distributions fit locally to each duration. Comparison is quantified first through the proper evaluation metric integrated quadratic distance (IQD) (Thorarinsdottir et al., 2013). Further, since the IQD is a measure of overall distributional similarity and is not always sensitive to small differences in tail behavior, we calculate the mean absolute percentage error (MAPE) for select high quantiles.

The IQD measures the similarity between two distributions by integrating over the squared distance between the distribution functions. Let G be the distribution function defined by the local GEV fit and G_{QDF} be the distribution function defined by the QDF model at the corresponding duration. In practice we approximate G and G_{QDF} by the empirical CDF of a sample from the posterior. The distance between G and G_{QDF} as measured by the IQD is then given by

$$\text{IQD} = \int_{-\infty}^{+\infty} \left(G(z) - G_{\text{QDF}}(z)\right)^2 \mathrm{d}z \tag{23}$$

where lower values of the IQD indicate better overall performance. The IQD is the score divergence associated with the well-known proper scoring rule the continuous ranked probability score (CRPS); the main difference between IQD and CRPS is that CRPS calculates the integrated squared distance between a distribution and a scalar observation specified by a Heaviside step function whereas IQD calculates the integrated squared distance between two distributions.

The MAPE provides a measure of similarity as the percent difference between the local GEV fit and the QDF model. Let $z_{d,p}^{\text{QDF}}$ be the return level at probability p for the QDF model evaluated at duration d, generated from the approximation to the posterior given in Eq. (22). Similarly, let $z_p^{\text{GEV,d}}$ be the return level at probability p for the local GEV fit to data at duration d. Then the MAPE is given by

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{z_p^{\text{GEV,d}} - z_{d,p}^{\text{QDF}}}{z_p^{\text{GEV,d}}} \right| * 100 \tag{24}$$

where n is the number of stations at which we wish to calculate the MAPE.

4. Results

We evaluate three models: the original QDF model (Javelle), the extended QDF model (Double-Delta), and the mixture model. We first assess how well the models capture flood behavior for in-sample durations at a variety of catchments. Then we evaluate which of the models is most effective at predicting out-of-sample durations, specifically short (less than 24 h) durations from long durations (greater than or equal to 24 h). Finally, we compare the models' estimation abilities at in- and out-of-sample durations.

Model evaluation is carried out by comparing the QDF models to a collection of GEV models fit individually to each duration. The IQD is used to assess model behavior across all quantiles; since it has low tail sensitivity it best captures model behavior where the bulk of our observations lie (i.e. return periods for which we have observed data). We turn to the MAPE to assess tail behavior, where both the QDF model and the reference model are extrapolated beyond the range of observed data.

4.1. Model sensitivity to input durations

The QDF models should be fit with the minimum number of durations needed to ensure converge of the MCMC sampler; feeding too many sets of dependent data into the model can bias return level estimates and artificially narrow the credible intervals. The bias is especially prevalent when the data is generated by aggregating over a longer time span and the goal is to predict short duration events.

To test this, the models were fit under three different sets of data: two durations (24 and 36 h); four durations (24, 36, 48, 72 h); and six durations (24, 36, 48, 72, 96, 120 h). For the two-duration set the MCMC sampler failed to converge. Results from the other two sets ("24–72" and "24–120") are displayed in Fig. 4. The 24–120 set provides a comparatively worse fit; the 90% credible interval for the this set fails to capture the locally fit GEV models (dashed gray lines) for the 24 and 1 h durations and the return levels are also underestimated to a greater extent than in the 24–72 set. This behavior is replicated across all three models and all twelve catchments (results not shown).

4.2. Model performance on in-sample durations

Here, we present results where the three QDF models are compared against locally fit GEV models at every in-sample duration, where the in-sample durations are 1, 24, 48, and 72 h. Such an in-sample comparison is useful for identifying specific scenarios where QDF models struggle to fit the data rather than strict model-to-model rankings: since models with more parameters have an in-sample advantage, Double-Delta is expected to perform better than either Javelle or the mixture model. Return level plots displaying the QDF model output and the reference model at these four in-sample durations are displayed in Figs. E.12–E.15.

4.2.1. Assessing model behavior using IQD

A comparison of in-sample IQD scores across stations, durations and methods is given in Fig. 5. The scores are relatively similar across models–most points fall on or along the diagonals in the two plots in Fig. 5. As expected, the scores exhibit a slight preference towards the Double-Delta model, which has the lowest average IQD score at 0.034 (highest distributional similarity to the reference model when all durations and stations are considered). The mixture model has the next lowest score at 0.037 and Javelle has the highest score at 0.040.

The analysis shows duration-specific preferences between models. The Double-Delta model has a better average IQD score than either Javelle or the mixture model at every in-sample duration where the average is taken over all 12 stations considered in the study. However, Double-Delta's advantage is strongest at the shortest durations. Table 1

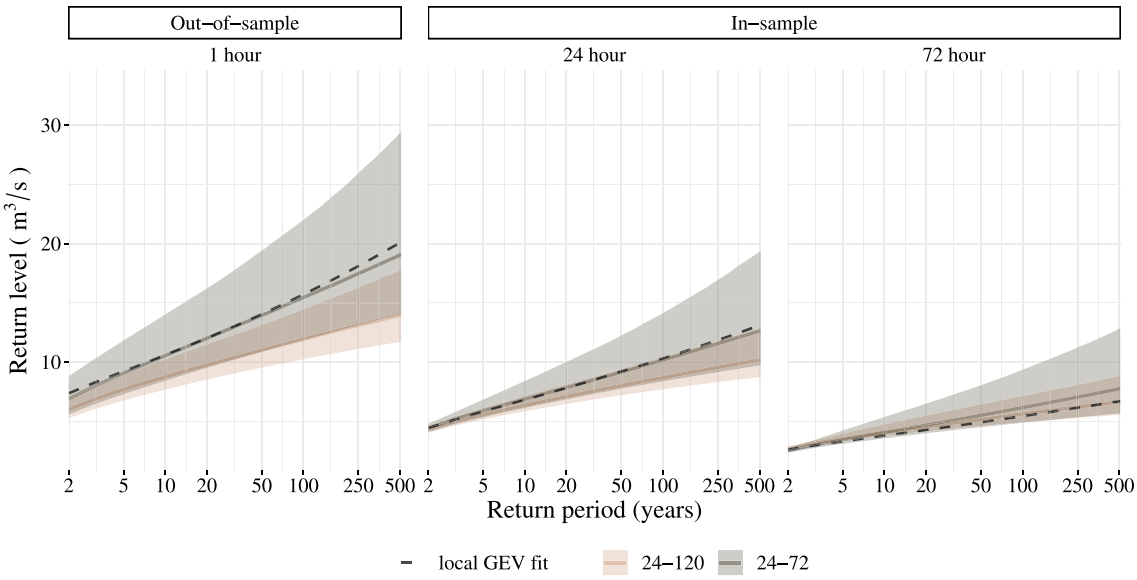

Fig. 4. Return level plots from the Dyrdalsvatn gauging station using the Double-Delta model fit to two different data sets: one set with six durations [24, 36, 48, 72, 96, 120 h] and one set with four durations [24, 36, 48, 72 h]. The model fit to the six duration set is both overconfident and biased at shorter durations; the posterior mean return level estimates are consistently underestimated when compared to locally fit GEV models (dashed gray lines) and the 90% credible interval is artificially narrow and fails to capture the locally fit model for the 24 and 1 h durations.

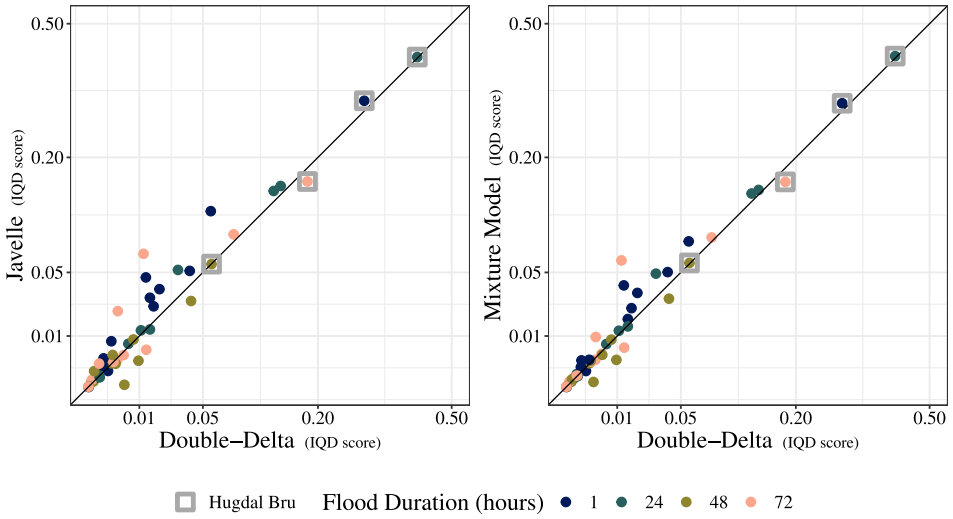

Fig. 5. Model-to-model comparison of interquantile distance (IQD) scores for each station and in-sample duration. Lower values of the IQD indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, left panel) and the mixture model (right panel). Notable values are indicated by gray squares, and are discussed in the main text.

Table 1
Number of stations at which the extended QDF model (Double-Delta) outperforms a comparison QDF model as measured by IQD. Here "MM" denotes the mixture model.

In-sample duration	Comparison model	
	Javelle	MM
1 h	10/12	10/12
24 h	9/12	9/12
48 h	7/12	7/12
72 h	7/12	8/12

reports the number of stations at which Double-Delta outperforms a comparison QDF model at each duration.

Despite QDF models showing an overall good performance, there are certain stations where each of the three QDF models differs substantially from the reference model. This behavior is particularly prevalent for the 1 and 24 h durations at Hugdal Bru, displayed in panel A of Fig. 6. We suspect the issues with the shorter durations at Hugdal Bru represent a conflict between the parameter constraints inherent in the QDF models and the runoff-generating processes for sub-daily streamflow at this particular station: Hugdal Bru is heavily snowmelt driven, with a strong diurnal melt pattern. The data averaging used in QDF modeling smooths out this sub-daily variation, but this relatively large reduction in variance is not reflected in the parameter constraints of the QDF model since the primary scaling occurs on the median flood (a constraint described in Eq. (14)). Thus the behavior of 1 h floods with return period under 5 years is difficult for the QDF models to fit. Floods with higher return periods tend to come from larger precipitation or melting events that supersede the diurnal cycle and

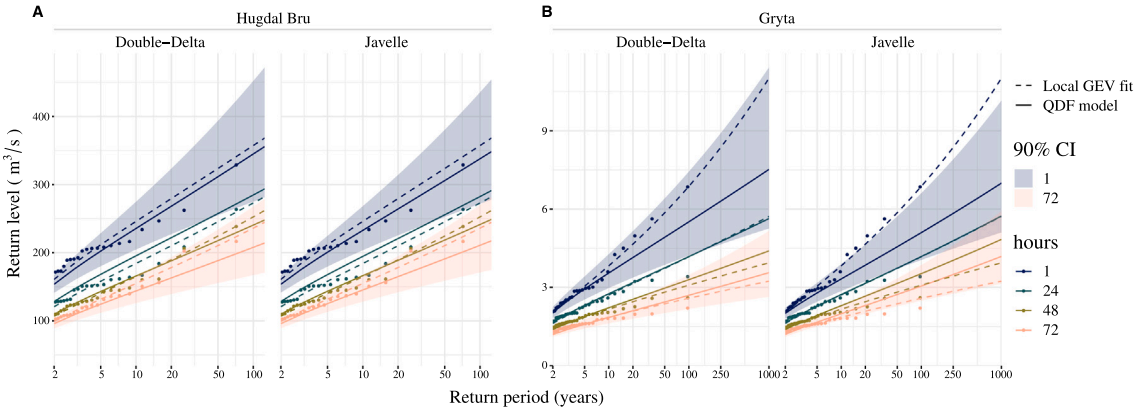

Fig. 6. Return level plots showing two selected stations where QDF models differ substantially from the reference model on in-sample durations. (**A**) Hugdal Bru: the 1 h floods with return period under 5 years are characterized by a diurnal melt-freeze cycle at this snowmelt-driven catchment; 1 h floods with longer return periods come from larger precipitation or warming events that supersede the diurnal cycle and as such have a more consistent relationship with longer durations and are more easily characterized by QDF models. (**B**) Gryta: the reference models show a change in shape parameter with increasing duration; QDF models cannot capture this behavior as the shape parameter is not duration dependent.

as such have a more regular relationship between durations. Durations above 24 h (without the diurnal cycle) also have a more regular relationship between durations.

The QDF models assume a constant shape parameter across all durations included in the analysis. As shown in panel B of Fig. 6, this assumption may lead to estimates that diverge from local duration-independent estimates where the latter analysis yields substantially varying shape parameter estimates across the durations. Here, the individually fit GEV models have shape parameters ranging from 0.140 for the 1 h duration to −0.037 for the 72 h duration. The QDF models do not have duration dependence built into the shape parameter and as such must choose one shape parameter for the entire set (in this case 0.018 for Double-Delta, 0.021 for the mixture model and 0.036 for Javelle). This inflexibility of the shape parameter is a known limitation of QDF models but is not easily solved as this parameter faces estimation difficulties due to the involved parametric form of the cumulative distribution function of the GEV. As a result, the QDF models tend to underestimate high quantiles for short durations and overestimate high quantiles for longer durations. Specifically for Gryta, using Javelle the 1 h duration is underestimated and the 48 and 72 h durations are both overestimated to a greater extent than we see in the Double-Delta model.

4.2.2. Assessing model behavior using MAPE

The within-sample MAPE was computed for the 100 year and 1000 year flood events (0.99 and 0.999 quantiles). These quantiles lie beyond the observed range of data for most of the stations and thus require extrapolation of both the QDF models and the reference model.

The Double-Delta model has the lowest MAPE at both return periods when all in-sample durations and stations are taken into account (5.9% error at the 100 year return period and 10.0% error at the 1000 year return period). The mixture model has the next lowest MAPE with 6.5% error at the 100 year return period and 12.1% error at the 1000 year return period. The Javelle model has the highest MAPE with 7.7% error at the 100 year return period and 12.1% error at the 1000 year return period. As with the IQD, the advantage of Double-Delta is strongest at the shortest durations; Table 2 reports the number of stations at which Double-Delta outperforms either Javelle or the mixture model.

The addition of the second delta parameter has the most impact when estimating events with long return periods. We see this in the differences in behavior of the model-to-model comparisons between the IQD and MAPE Figs. 5 and 7. Javelle and the mixture model appear more similar when evaluated by the IQD than they do under the MAPE;

Table 2
Number of stations at which the extended QDF model (Double-Delta) outperforms a comparison QDF model as measured by MAPE. Here "MM" refers to the mixture model.

In-sample duration	Comparison model		T
	Javelle	MM	
1 h	11/12	11/12	100
24 h	10/12	9/12	
48 h	4/12	4/12	
72 h	7/12	6/12	
1 h	11/12	11/12	1000
24 h	9/12	9/12	
48 h	4/12	4/12	
72 h	6/12	6/12	

that is, using the IQD score the two models have about the same amount of clustering around the diagonal when compared to Double-Delta. But using MAPE–which measures differences in tail behavior between the QDF models and reference model–we see a difference between Javelle and mixture model when compared to Double-Delta: the values for the mixture model are much more closely clustered around the diagonal in Fig. 7 than the values for Javelle. These stations that show an improvement in MAPE under the mixture model are those that have a high weight on the second delta parameter.

One of the stations that is most improved by the addition of the second delta is Gryta (marked by gray squares in Fig. 7). The return level plots in panel (B) of Fig. 6 show this station in particular benefits from the adjustment of growth curve slope afforded by the second delta. The second delta somewhat mitigates the effect of the assumption of a constant shape parameter across durations. However, even with this adjustment in growth curve slope both Double-Delta and the mixture model have high error values for the 1 h duration at Gryta-around 20%–30%.

4.3. Model performance on out-of-sample durations

Here, the models were fit with four durations (24, 36, 48 and 60 h) and the resulting parameter estimates were used to predict the 1 and 12 h durations. The QDF predictions were compared to locally fit GEV models using both the IQD and MAPE. Return level plots showing the reference and QDF models at both out of sample durations are displayed in Figs. F.16–F.19.

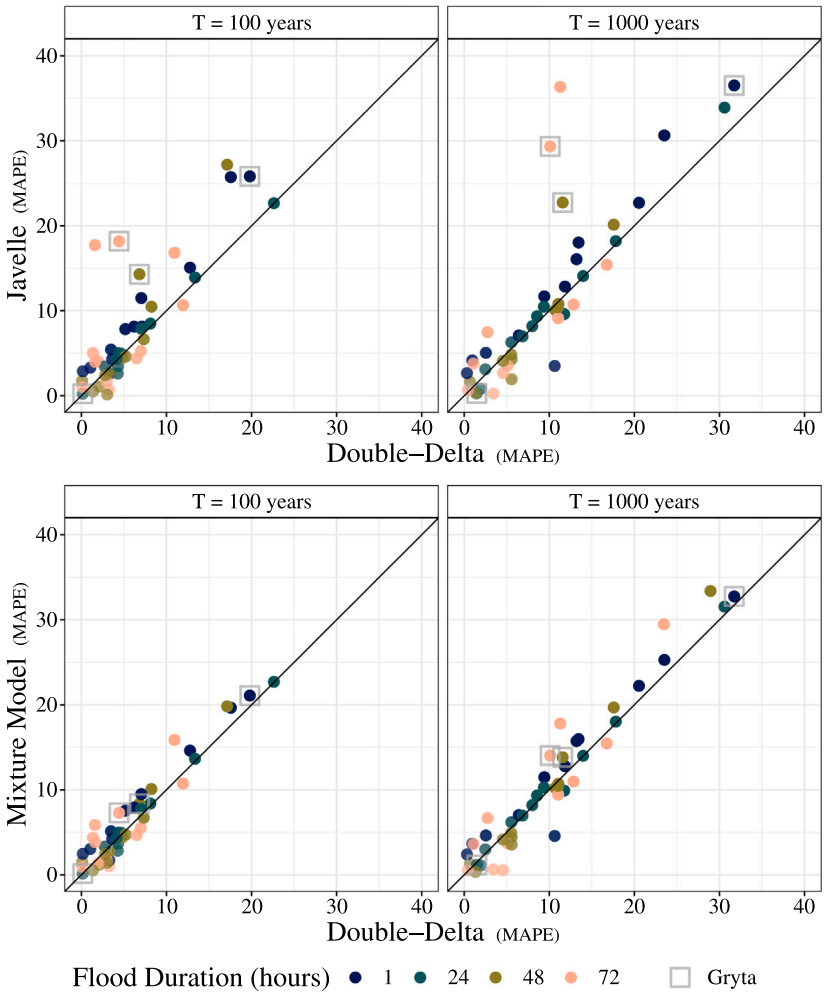

Fig. 7. Model-to-model comparison of the mean absolute percent error (MAPE) scores for each station and in-sample duration. Lower values of the MAPE indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, top panels) and the mixture model (bottom panels). Notable values are indicated by gray squares, and are discussed in the main text.

Double-Delta has the best average IQD score on the out of sample durations, reporting a score of 0.34 while the mixture model reports a score of 0.42 and Javelle reports 0.44. Fig. 8 shows a model-to-model comparison on the out of sample durations. There are only three station and duration combinations (both the 1 and 12 h durations at Sjodalsvatn and the 1 h duration at Dyrdalsvatn and Øyungen) where Double-Delta performs worse, as measured by the IQD, than the other two models. These stations are outlined in red in Fig. 8. At every other station and duration Double-Delta performs the same or better. All three QDF models provide a poor distributional fit for the sub-daily durations at Hugdal Bru and the 1 h duration at Røykenes. These stations are labeled by name in Fig. 8. Difficulties fitting the sub-daily durations of Hugdal Bru are discussed in Section 4.2.1. The 1 h duration at Røykenes exhibits a large change in shape parameter with an increase in duration like the station Gryta shown in panel B of Fig. 6.

Double-Delta has the best average MAPE score on the out of sample durations (11.1% error at the 100 year return period and 15.4% error at the 1000 year return period). The mixture model has the next lowest MAPE with 12.2% error at the 100 year return period and 16.9% error at the 1000 year return period. The Javelle model has the highest MAPE with 12.8% error at the 100 year return period and 17.4% error at the 1000 year return period. Double-Delta provides an equal or better fit at around 80% of the stations and durations at both return periods. Stations and durations where Double-Delta is outperformed by either Javelle or the mixture model are outlined in red in Fig. 9.

Several of the smallest catchments (Gravå, Gryta and Grosettjern) have high out-of-sample MAPE values. These three catchments have some of the highest variation in the shape and slope of the individually fit GEV models (see Tables A.3 and B.5, where the β parameter is taken as a proxy for slope).

A highly duration-dependent shape parameter is a known challenge for QDF models (see the scenario in panel B of Fig. 6) and we would expect the QDF models to struggle to find a shape parameter value that approximates both the longest and shortest durations even when these durations are in-sample. Furthermore, not only do we observe a large shape parameter range but this range crosses zero for both Gryta and Grosettjern, with the longer durations having a negative shape parameter while the shorter durations have a positive shape parameter. This is a substantial difference; a negative shape parameter corresponds to an entirely different distribution family (Weibull) than a positive shape parameter (Fréchet) within the GEV family.

Additionally, these three catchments experience the biggest change in growth curve slope between either the 1 and 24 h duration or the 12 and 24 h duration while the rate of change of growth curve slope is less for durations above 24 h; that is, there is a change in growth curve slope in the sub-daily durations that is not replicated in the longer durations. In summary, we observe high error for out of sample durations at Gravå, Gryta and Grosettjern because the relationship between the longer floods used to fit the model does not strongly inform the relationship between sub-daily floods for these catchments.

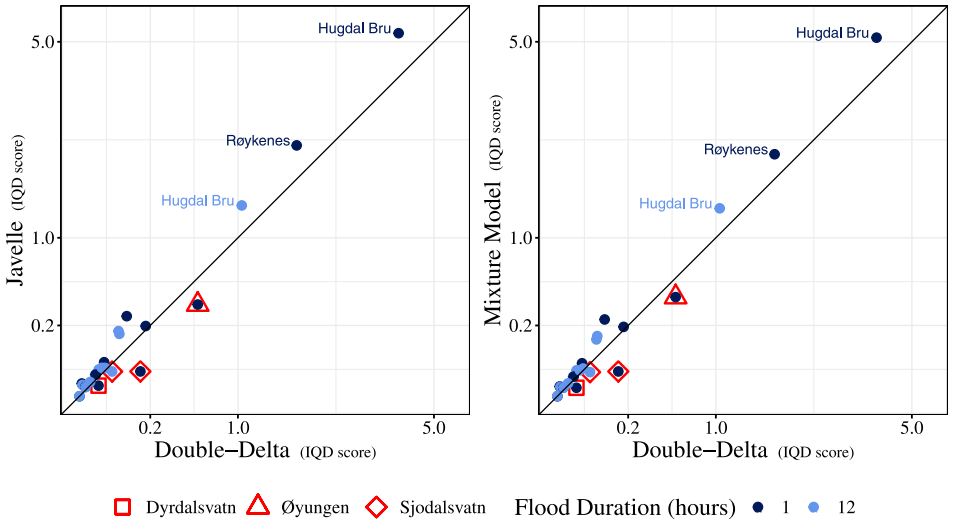

Fig. 8. Model-to-model comparison of interquantile distance (IQD) scores for each station and both out-of-sample durations. Lower values of the IQD indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, left panel) and the mixture model (right panel). Stations and durations where Double-Delta performs worse than the other two models are outlined in red. Stations and durations that are fit particularly poorly by all three QDF models are labeled by name. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

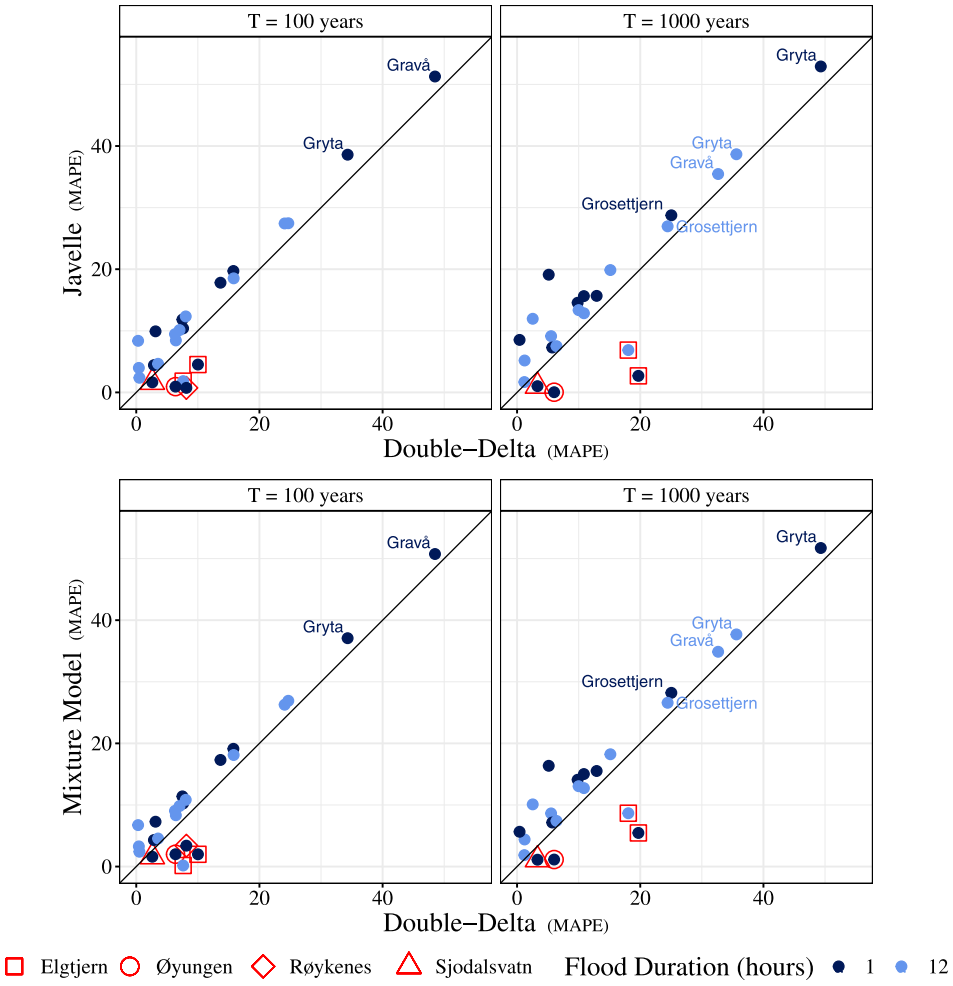

Fig. 9. Model-to-model comparison of mean absolute percent error (MAPE) scores for each station and both out-of-sample durations. Lower values of the MAPE indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, top panels) and the mixture model (bottom panels). Stations and durations where Double-Delta performs worse than the other two models are outlined in red. Stations and durations that are fit particularly poorly by all three QDF models are labeled by name. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

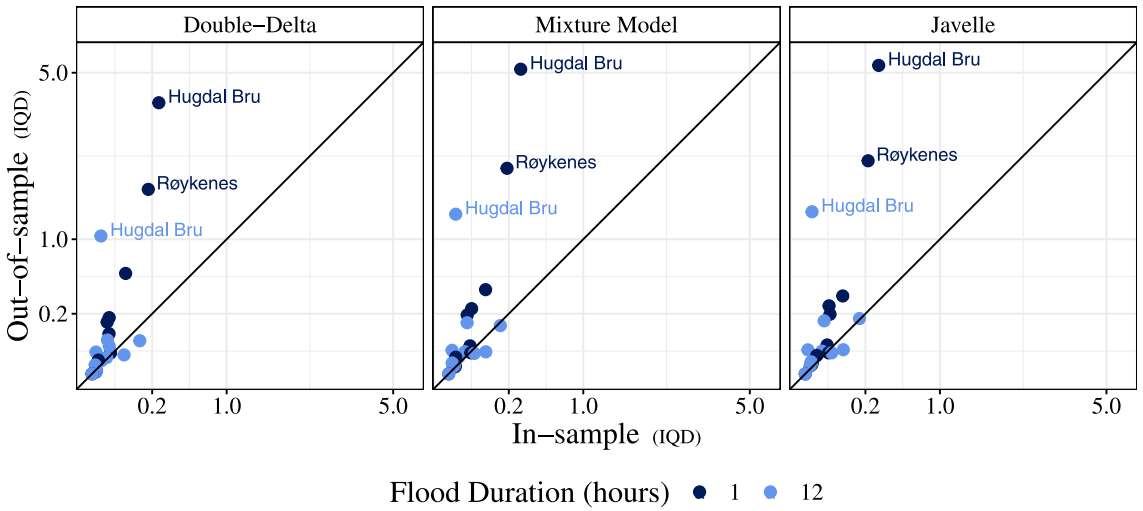

Fig. 10. Comparison of interquantile distance (IQD) score when durations are either predicted (out of sample) or included in the model fitting set (in sample). Lower values of the IQD indicate better performance. The out-of-sample set was fit with durations 24, 36, 48, 60 h and used to predict the 1 and 12 h durations. The in-sample set was fit with durations 1, 12, 24, 36, 48, 60 h. Stations and durations that are fit particularly poorly by all three QDF models are labeled by name.

4.4. Comparison of in- and out-of-sample sub-daily estimates

Here, the models were fit with six durations (1, 12, 24, 36, 48, 60 h) where the 1 and 12 h durations are evaluated as in-sample durations. The output from these models is then compared to the output from the previous section, where the models are fit on four durations (24, 36, 48, 60 h) that are used to predict the 1 and 12 h durations. The performance of each of these sets is evaluated at the 1 and 12 h durations using both the IQD, as shown in Fig. 10, and MAPE, as shown in Fig. 11.

The stations that have the greatest loss when going from in-sample to out-of-sample tend to be stations that already had high IQD or MAPE values. This means that if there is already a significant difference between the QDF and reference models this difference is likely to be amplified when predicting out of sample durations. Most stations and durations, however, have a relatively moderate loss when moving from in- to out-of-sample on both the IQD and MAPE (the exceptions to this are labeled in Figs. 10 and 11). For the MAPE, this difference is on the order of ±5%.

5. Discussion

We have, in accordance with our main objective, analyzed how different QDF models capture the relationship between floods of different duration at 12 locations in Norway. By examining differences in model fit between the three models studied, we identified reasoning to explain why the extended QDF model ("Double-Delta") outperforms the other two models on the particular stations and durations studied, and why this performance advantage is particularly pronounced for situations where the focus is on long return periods and/or short durations. Additionally, we tested the out-of-sample performance of QDF models on sub-daily durations by comparing to models fit with the sub-daily data included; we observed situations where the out-of-sample set returned evaluation scores that were in line with the in-sample set but also situations where the ability of QDF models to predict sub-daily, out-of-sample durations was severely limited. Finally, we assessed whether the choice of durations used to fit the QDF models impacts model estimation and concluded QDF models are sensitive to the durations used to fit them.

The Double-Delta model is what we term a "empirical multiscaling" model, where the main contribution of the proposed model is the ability to adjust to certain types of changes in dependence structure with respect to return period. Specifically, it can account for the situation where the ratio between growth curves increases with increasing return period. The original QDF model (Javelle), on the other hand, assumes this ratio to be constant. As evidenced by the return level plots in Figs. E.12–E.15, the assumption of a constant ratio will commonly not hold, in particular, if the shortest duration of 1 h is included in the comparison. The additional parameter in the Double-Delta model allows for a better approximation of the tail behavior, especially for short durations. Selectively adding the second delta–as the mixture model does–is not advantageous at the shortest durations as these durations tend to need maximum flexibility from the QDF models.

We make a distinction here between what we call *empirical multiscaling* and multiscaling in the strict theoretical sense. Strict theoretical multiscaling models would be, for example, those presented in Gupta and Waymire (1990) or the IDF models in Van de Vyver (2018). This distinction is often overlooked in the literature since the parameterization of empirical- and strict-multiscaling models are in most cases identical and the theoretical basis matters only for inference. However, we think it useful to note that the QDF models presented here are empirical and do not attempt to place strict mathematical assumptions on how the variance or other higher-order moments change with increasing duration.

A second important distinction needs to be made between QDF models and bivariate frequency analyses where the dependence structure between peak discharge and event duration is explicitly modeled. The aggregation-based approach to obtaining annual maxima means QDF models provide an accessible way to get relationships between peak volume and duration for applications where the total volume of water is of interest. If singular flood events are the focus—for example, if we need to know how long a road is closed following a particular flood event—a bivariate, event-based approach such as one of the copula models detailed in Gräler et al. (2013) is more appropriate.

QDF models are most useful when three considerations are kept in mind. Firstly, we found that the choice of durations used to fit the QDF model was a highly influential aspect of the modeling process. The particular durations chosen will impact what relationship between floods the QDF models can identify. In general, QDF models predict sub-daily unobserved durations just as well as when those durations are used to fit the model. However, as shown in Section 4.4, it is possible to select in-sample durations that do not inform the duration of interest. Avoiding this situation requires careful selection of appropriate in-sample durations. Such selection can be guided by design value application; for example, it is unlikely we would need the 60 or 72 h duration on the smallest catchments in this study and can

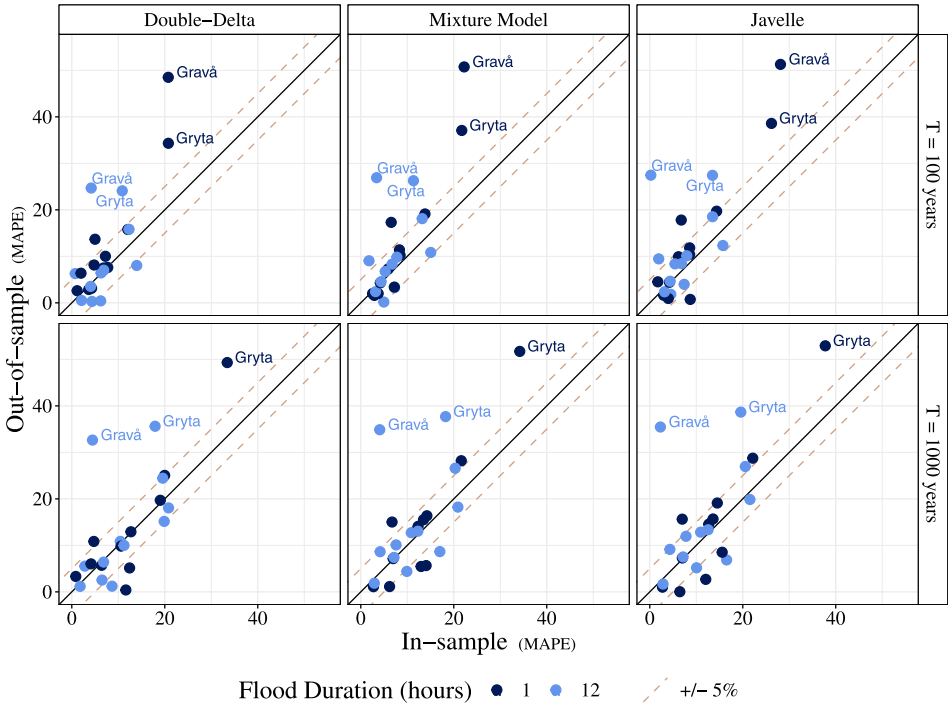

Fig. 11. Comparison of mean absolute percent error (MAPE) when durations are either predicted (out of sample) or included in the model fitting set (in sample). Lower values of the MAPE indicate better performance. The out-of-sample set was fit with durations 24, 36, 48, 60 h and used to predict the 1 and 12 h durations. The in-sample set was fit with durations 1, 12, 24, 36, 48, 60 h. Stations and durations that are fit particularly poorly by all three QDF models are labeled by name and dashed lines indicate ±5% difference from the diagonal.

therefore avoid the somewhat contrived scenarios where we use what are, for these catchments, only long-duration flood events to estimate the shortest durations.

Secondly, the range of the selected durations also influences the QDF model estimates. If the durations selected do not span a wide enough range the QDF models will struggle to converge (Section 4.1). However, too wide a range of durations can be challenging for QDF models if the statistical properties of the floods change significantly between durations (Section 4.2). We note that problems associated with the latter situation can be partially mitigated through the extra flexibility afforded by the extended QDF model (Double-Delta). Thirdly, we found that generating too many sets of dependent data to fit the model can produce results that are both biased and overconfident, particularly when the generated data is aggregated over a longer time span than the duration of interest (Fig. 4).

The QDF model assumes a constant shape parameter across all durations, as with nearly all duration-dependent extreme value models (Fauer et al., 2021). This situation is illustrated in panel B of Fig. 6. It would be technically possible to add duration dependence to the shape parameter of the models in Eqs. (9), (13), and (15). However, the observed difficulties in estimating the shape parameter in Section 4.3 and the issues documented in Martins and Stedinger (2000) indicate this approach may be very complex and pose severe estimation problems. Additionally, observation of the shape parameter values from individually fit GEV distributions demonstrate the shape parameter does not appear to change with duration in as structured a way as either the median flood (η) or the change in slope of the growth curves (where this change is described in part by β).

The Double-Delta model is a promising avenue for improved modeling of short-duration events and events with long return periods under a QDF modeling framework. We identify several areas of future research. Extending the analysis presented in this paper to include more gauging stations–including stations in diverse climate regions–is a priority; while the catchments used in this study are diverse for Nordic catchments, they are not diverse globally. Additionally, of particular interest is how this extended QDF model will function in a regional setting; many of the design flood values needed for operational use in Norway are at ungauged sites or at sites with incomplete or very short datasets.

Furthermore, it could be beneficial to include a more explicit consideration of flood generating processes within QDF methods; seasonal needs in reservoir management, for example, can mean that a varying flood storage capacity needs to be defined within a year. Methods exist to explicitly account for generating processes in FFA (see, for example, the mixture models in Fischer, 2018) but are not directly suited to the aggregation-based methodology underlying QDF. A potential avenue forward could be definition of seasonal blocks as in Ulrich et al. (2021), who developed IDF curves with monthly varying parameters.

Additionally, a potential area of improvement for predicting short durations when the majority of the data is at a daily (or longer) time resolution is to allow the QDF models to take data where the length of the data record varies by duration, such that some information on short durations can be included even if the data for these durations is relatively scant. Finally, non-stationarity due to climate change will be an important future area of research for QDF models. While this is outside the scope of this study, we identify a few references that could serve as an example for future research. Nonstationarity is addressed for regional QDF models in Cunderlik and Ouarda (2006) and for regional FFA models in a Bayesian framework in Guo et al. (2022).

6. Conclusions

This paper proposes a multiscaling extension of the QDF model of Javelle et al. (2002), where the magnitude of the index flood and the slope of the growth curve may scale independently with duration. In the original QDF model only the magnitude of the index flood scales across durations (Javelle et al., 2002). A Bayesian inference algorithm is developed where the original QDF model, the extended QDF model, or a mixture of the two may be estimated. In a case study comprising 12 study locations in Norway, we analyze how these

three different QDF models capture the relationship between floods of different duration. The results suggest it is advantageous to allow the index flood and growth curve slope to scale independently; that is, it is advantageous to let the ratio between growth curves of different duration be dependent on return period. This advantage is particularly pronounced for situations where the focus is on long return periods and/or short durations. Thus the extended QDF model is the most promising avenue for capturing flood behavior at the shortest (sub-daily) durations. In general, QDF models are generally able to predict out-of-sample durations with a relatively moderate loss in accuracy when compared to in-sample estimates for the same durations. However, we found the QDF framework to be highly sensitive to the choice of durations used to fit the models. In particular, care should be taken to fit the QDF models with the minimum number of durations needed for the inference algorithm to converge. Generating too many sets of dependent data to fit the model can produce results that are both biased and overconfident. The extended QDF model has an improved ability to simultaneously model a wider range of durations when compared to the original QDF model.

CRediT authorship contribution statement

Danielle M. Barna: Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Kolbjørn Engeland:** Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. **Thordis L. Thorarinsdottir:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Chong-Yu Xu:** Conceptualization, Methodology, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kolbjorn Engeland and Chong-Yu Xu reports financial support was provided by Research Council of Norway.

Data availability

The flood and hydrological data were extracted from the National Hydrological Database (Hydra II) hosted by the Norwegian Water Resources and Energy Directorate (NVE). The 12 stations used in this analysis are published at https://doi.org/10.5281/zenodo.7085557.

Acknowledgments

This work was supported by the Research Council of Norway through grant nr. 302457 “Climate adjusted design values for extreme precipitation and flooding” (ClimDesign) and FRINATEK Project 274310. The authors would like to thank Thea Roksvåg and Alex Lenkoski for valuable discussions and Mads-Peter Dahl for help with data selection.

Appendix A. Shape parameter values for QDF and reference models

See Tables A.3 and A.4.

Appendix B. β Parameter values for reference models

See Table B.5.

Appendix C. Mean absolute percent error for out-of-sample sub-daily durations

See Table C.6.

Appendix D. Catchment properties for selected catchments

See Table D.7.

Table A.3
Posterior mean shape parameter values with 90% credible intervals for QDF model fit on durations (24, 36, 48, 60 h) and posterior mean shape parameter values for individually fit GEV distributions. Stations are in order of catchment area.

Station	Individually fit GEV						QDF					
	Duration (h)						Model type					
	1	12	24	36	48	60	DD		MM		J	
Dyrdalsvatn	0.14	0.08	0.06	0.09	0.09	0.08	0.05	[-0.06, 0.17]	0.05	[−0.07, 0.17]	0.05	[−0.07, 0.17]
Gravå	0.18	0.12	0.10	0.07	0.06	0.05	0.04	[−0.07, 0.16]	0.04	[−0.06, 0.16]	0.04	[−0.06, 0.16]
Grosettjern	0.07	0.06	0.05	0.01	−0.01	−0.02	−0.04	[−0.11, 0.04]	−0.04	[−0.1, 0.04]	−0.03	[−0.1, 0.04]
Elgtjern	0.17	0.16	0.17	0.17	0.16	0.15	0.22	[0.1, 0.33]	0.22	[0.1, 0.33]	0.22	[0.1, 0.33]
Gryta	0.14	0.07	0.03	0	−0.02	−0.03	−0.07	[−0.16, 0.02]	−0.07	[−0.16, 0.02]	−0.07	[−0.16, 0.03]
Røykenes	−0.02	−0.03	−0.05	−0.06	−0.07	−0.07	−0.13	[−0.2, −0.06]	−0.13	[−0.19, −0.06]	−0.13	[−0.19, −0.06]
Manndalen Bru	0.03	0.04	0.05	0.05	0.06	0.05	0.01	[−0.08, 0.12]	0.01	[−0.08, 0.12]	0.01	[−0.08, 0.11]
Øyungen	0.03	0.03	0.04	0.05	0.05	0.07	0.02	[−0.04, 0.10]	0.02	[−0.04, 0.10]	0.02	[−0.04, 0.10]
Sjodalsvatn	0.11	0.1	0.11	0.11	0.11	0.12	0.11	[0.01, 0.22]	0.11	[0.01, 0.23]	0.12	[0.01, 0.23]
Viksvatn	−0.08	−0.08	−0.08	−0.09	−0.1	−0.11	−0.13	[−0.17, −0.08]	−0.13	[−0.17, −0.08]	−0.13	[−0.17, −0.08]
Hugdal Bru	0.02	0.05	0.05	0.09	0.09	0.09	0.05	[−0.04, 0.15]	0.05	[−0.04, 0.15]	0.05	[−0.04, 0.15]
Etna	−0.04	−0.05	−0.06	−0.06	−0.07	−0.08	−0.11	[−0.16, −0.05]	−0.11	[−0.16, −0.05]	−0.11	[−0.16, −0.05]

Table A.4
Posterior mean shape parameter values with 90% credible intervals for QDF model fit on durations (1, 24, 48, 72 h) and posterior mean shape parameter values for individually fit GEV distributions. Stations are in order of catchment area.

Station	Individually fit GEV				QDF					
	Duration (h)				Model type					
	1	24	48	72	DD		MM		J	
Dyrdalsvatn	0.14	0.06	0.09	0.06	0.06	[−0.05, 0.17]	0.06	[−0.04, 0.17]	0.06	[−0.04, 0.17]
Gravå	0.18	0.10	0.06	0.05	0.13	[0.03, 0.24]	0.14	[0.05, 0.26]	0.15	[0.03, 0.25]
Grosettjern	0.07	0.05	−0.01	−0.03	−0.01	[−0.09, 0.07]	−0.01	[−0.08, 0.07]	−0.01	[−0.08, 0.07]

(continued on next page)

Table A.4 (*continued*).

Station	Individually fit GEV				QDF					
	Duration (h)				Model type					
	1	24	48	72	DD		MM		J	
Elgtjern	0.17	0.17	0.16	0.14	0.21	[0.10, 0.33]	0.21	[0.10, 0.32]	0.21	[0.10, 0.33]
Gryta	0.14	0.03	−0.02	−0.04	0.02	[−0.07, 0.11]	0.02	[−0.04, 0.12]	0.04	[−0.06, 0.11]
Røykenes	−0.02	−0.05	−0.07	−0.07	−0.11	[−0.17, −0.04]	−0.11	[−0.16, −0.04]	−0.10	[−0.17, −0.04]
Manndalen Bru	0.03	0.05	0.06	0.04	0.003	[−0.09, 0.11]	0.002	[−0.09, 0.1]	0.002	[−0.09, 0.1]
Øyungen	0.03	0.04	0.05	0.08	0.02	[−0.04, 0.09]	0.02	[−0.05, 0.09]	0.02	[−0.05, 0.09]
Sjodalsvatn	0.11	0.11	0.11	0.12	0.12	[0.01, 0.22]	0.12	[0.01, 0.23]	0.12	[0.01, 0.22]
Viksvatn	−0.08	−0.08	−0.10	−0.12	−0.13	[−0.17, −0.08]	−0.12	[−0.17, −0.08]	−0.12	[−0.17, −0.08]
Hugdal Bru	0.02	0.05	0.09	0.07	0.03	[−0.06, 0.13]	0.03	[−0.06, 0.13]	0.03	[−0.06, 0.13]
Etna	−0.04	−0.06	−0.07	−0.07	−0.10	[−0.15, −0.04]	−0.10	[−0.15, −0.04]	−0.10	[−0.15, −0.04]

Table B.5
Posterior mean beta parameter values for individually fit GEV distributions. Stations are in order of catchment area.

Station	Individually fit GEV						
	Duration (h)						
	1	12	24	36	48	60	72
Dyrdalsvatn	−1.56	−1.51	−1.4	−1.47	−1.5	−1.51	−1.55
Gravå	−1.19	−1.37	−1.46	−1.5	−1.53	−1.53	−1.55
Grosettjern	−1.22	−1.25	−1.28	−1.32	−1.34	−1.37	−1.37
Elgtjern	−0.98	−1.00	−1.02	−1.06	−1.08	−1.09	−1.12
Gryta	−0.92	−0.99	−1.07	−1.14	−1.18	−1.21	−1.25
Røykenes	−1.28	−1.29	−1.31	−1.37	−1.44	−1.49	−1.55
Manndalen Bru	−1.43	−1.47	−1.47	−1.50	−1.52	−1.51	−1.5
Øyungen	−1.06	−1.07	−1.08	−1.10	−1.10	−1.11	−1.13
Sjodalsvatn	−1.39	−1.39	−1.41	−1.42	−1.44	−1.47	−1.49
Viksvatn	−1.59	−1.59	−1.60	−1.60	−1.61	−1.62	−1.63
Hugdal Bru	−1.30	−1.38	−1.35	−1.37	−1.36	−1.34	−1.31
Etna	−1.10	−1.11	−1.13	−1.13	−1.14	−1.15	−1.15

Table C.6
Mean absolute percent error (MAPE) for return levels at the 100 and 1000 year return periods. This is the table version of Fig. 9. The MAPE is calculated in regard to individually fit GEV distributions (see Section 3.4 for details). Here "MM" denotes the mixture model. Stations are in order of catchment area.

Station	Model type											
	DD				MM				J			
	Duration = 1 h		Duration = 12 h		Duration = 1 h		Duration = 12 h		Duration = 1 h		Duration = 12 h	
	Return period (years)		Return period (years)		Return period (years)		Return period (years)		Return period (years)		Return period (years)	
	100	1000	100	1000	100	1000	100	1000	100	1000	100	1000
Dyrdalsvatn	3.1	5.1	0.3	2.5	7.3	16.0	6.7	10.0	9.9	19.0	8.4	12.0
Gravå	49.0	58.0	25.0	33.0	51.0	61.0	27.0	35.0	51.0	61.0	27.0	35.0
Grosettjern	16.0	25.0	16.0	24.0	19.0	28.0	18.0	27.0	20.0	29.0	19.0	27.0
Elgtjern	10.0	20.0	7.6	18.0	2.0	5.5	0.2	8.7	4.5	2.7	1.8	6.9
Gryta	34.0	49.0	24.0	36.0	37.0	52.0	26.0	38.0	39.0	53.0	27.0	39.0
Røykenes	8.1	0.4	8.0	15.0	3.4	5.7	11.0	18.0	0.7	8.5	12.0	20.0
Manndalen Bru	7.5	9.8	7.0	10.0	11.0	14.0	9.8	13.0	12.0	15.0	10.0	13.0
Øyungen	6.4	6.0	0.4	1.2	2.0	1.1	3.3	4.4	0.9	0.0	4.0	5.2
Sjodalsvatn	2.6	3.3	0.5	1.2	1.6	1.1	2.4	1.9	1.6	1.0	2.4	1.7
Viksvatn	2.9	5.7	3.5	6.4	4.3	7.2	4.6	7.4	4.4	7.3	4.6	7.5
Hugdal Bru	14.0	11.0	6.3	5.5	17.0	15.0	9.1	8.6	18.0	16.0	9.5	9.1
Etna	7.6	13.0	6.4	11.0	10.0	16.0	8.3	13.0	10.0	16.0	8.4	13.0

Table D.7
Catchment area, median flood, record length, and fraction of rain for the 12 selected catchments used in this study.

Station name	Catchment area (km^2)	Median flood (m^3/s)	Record length (years)	FGP (fraction of rain)
Dyrdalsvatn	3.31	7.52	32	0.93
Gravå	6.31	2.27	43	0.69
Grosettjern	6.6	1.54	53	0.32
Elgtjern	6.63	1.76	28	0.69
Gryta	7.03	1.99	54	0.59
Røykenes	50.09	65.84	39	0.95
Manndalen Bru	200.48	61.22	42	0.35
Øyungen	239.07	165.08	42	0.76
Sjodalsvatn	479.97	118.31	36	0.44
Viksvatn	508.13	174.01	118	0.76
Hugdal Bru	546.17	172.84	40	0.44
Etna	570.17	104.33	102	0.35

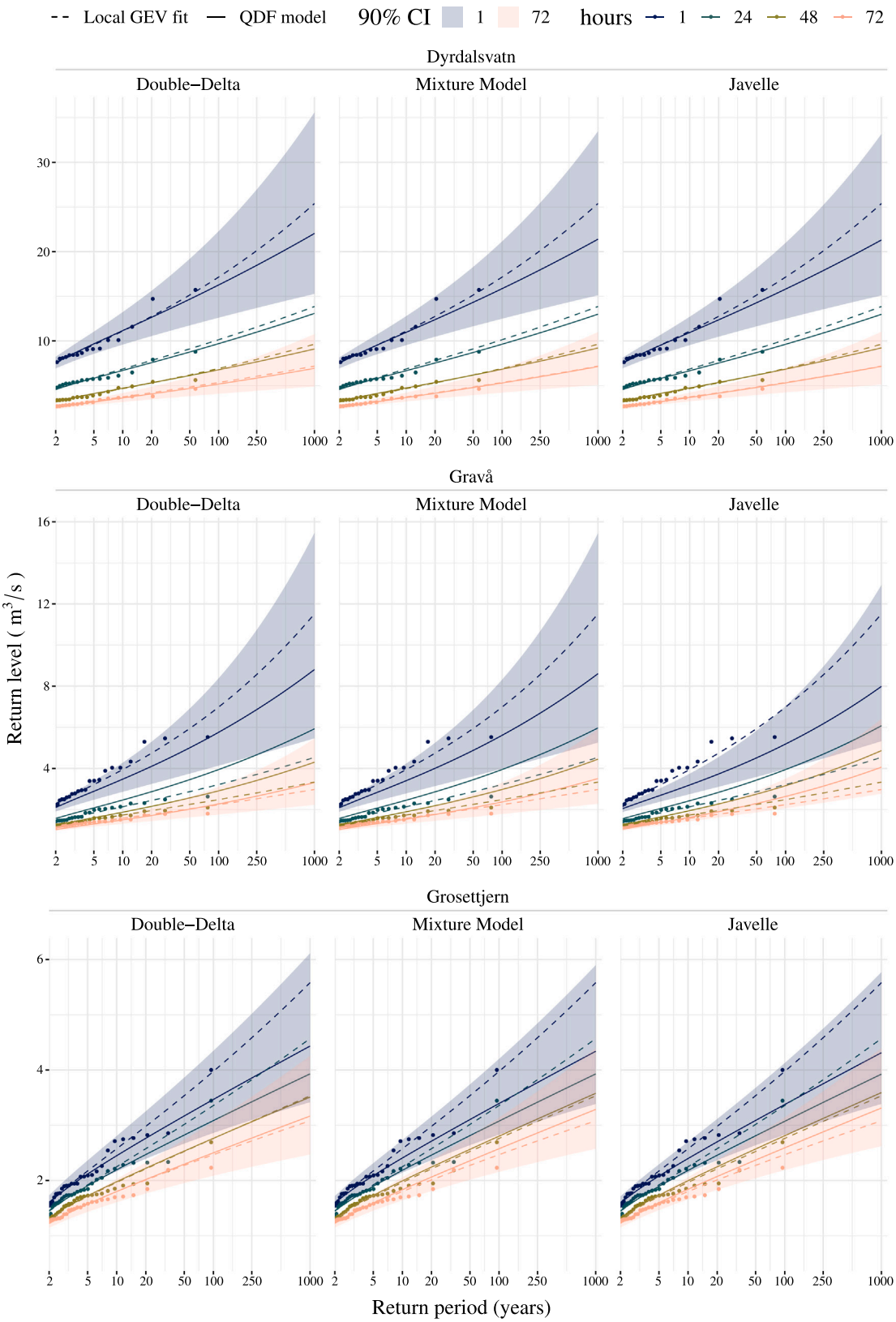

Fig. E.12. In-sample return level plots for stations Dyrdalsvatn, Gravå, and Grosettjern.

Appendix E. In-sample return level plots

See Figs. E.12–E.15.

Appendix F. Out-of-sample return level plots

See Figs. F.16–F.19.

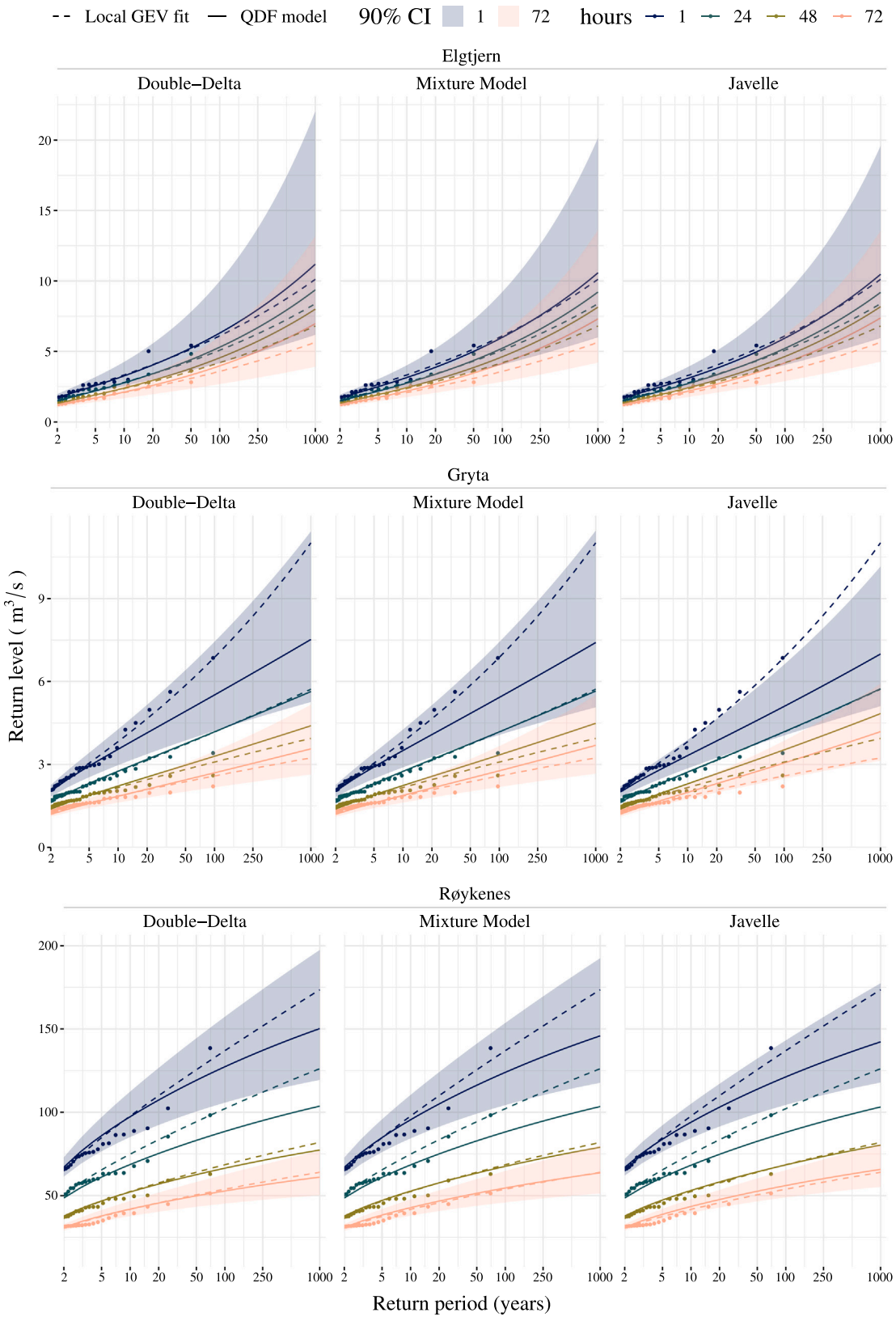

Fig. E.13. In-sample return level plots for stations Elgtjern, Gryta, and Røykenes.

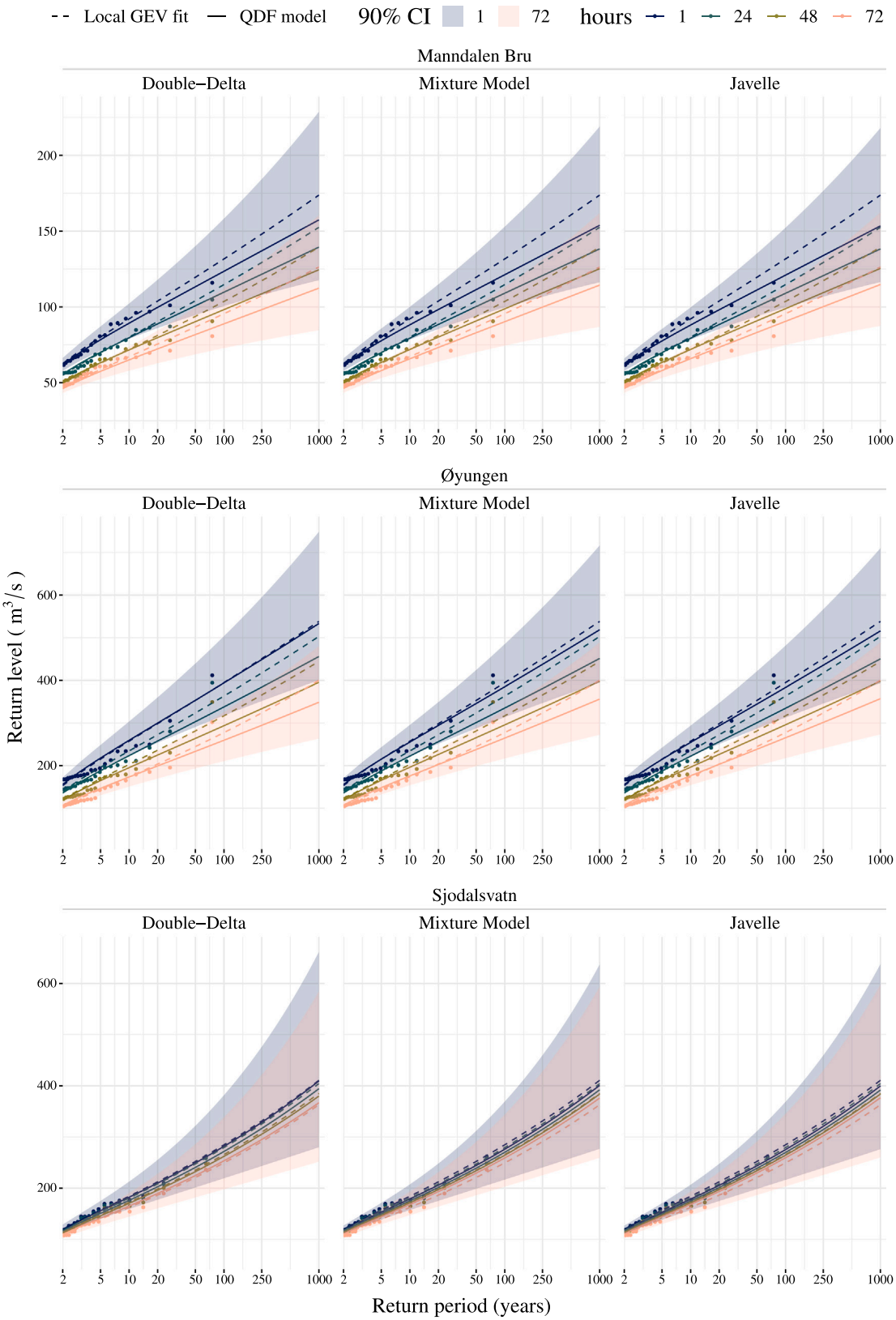

Fig. E.14. In-sample return level plots for stations Manndalen Bru, Øyungen, and Sjodalsvatn.

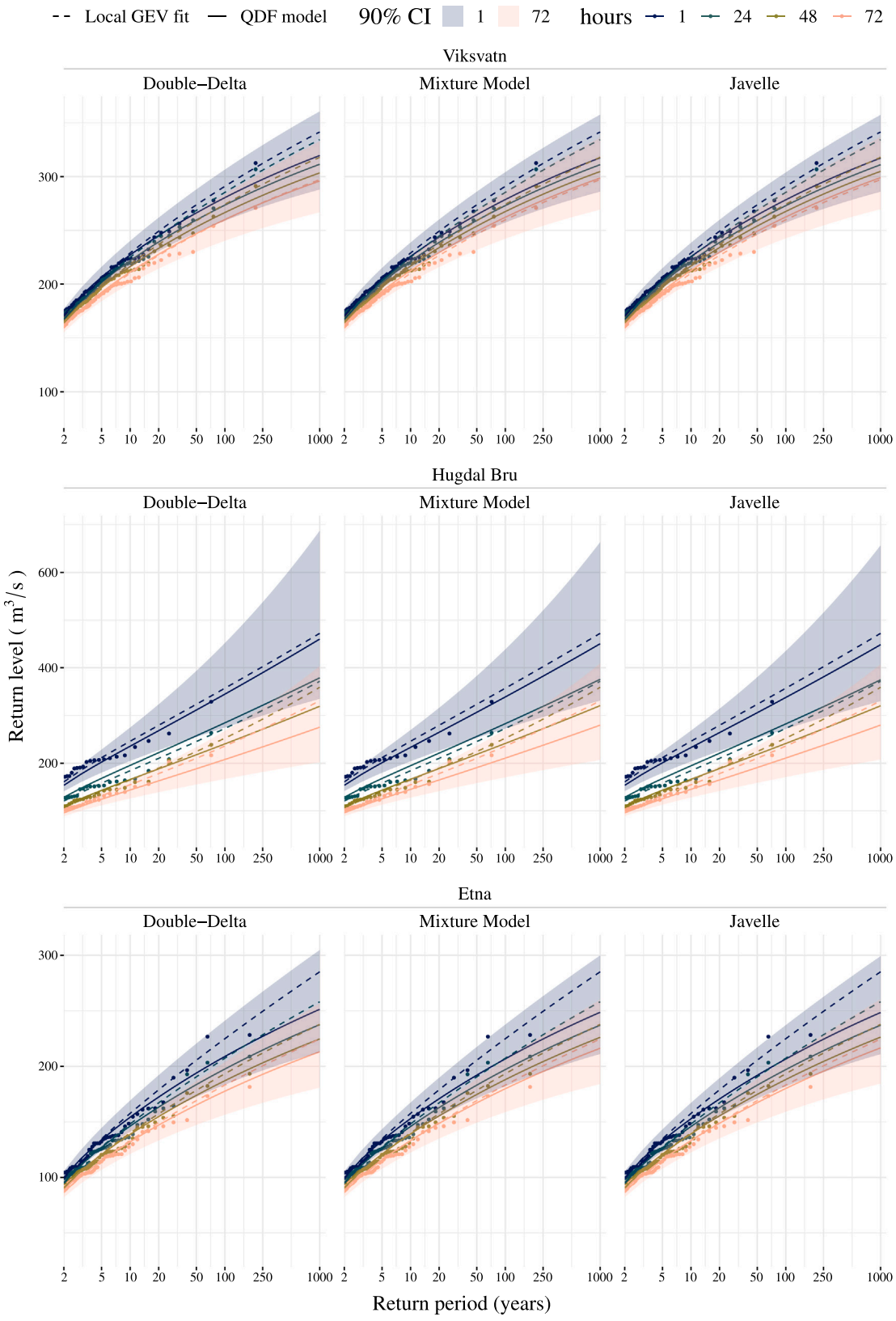

Fig. E.15. In-sample return level plots for stations Viskvatn, Hugdal Bru, and Etna.

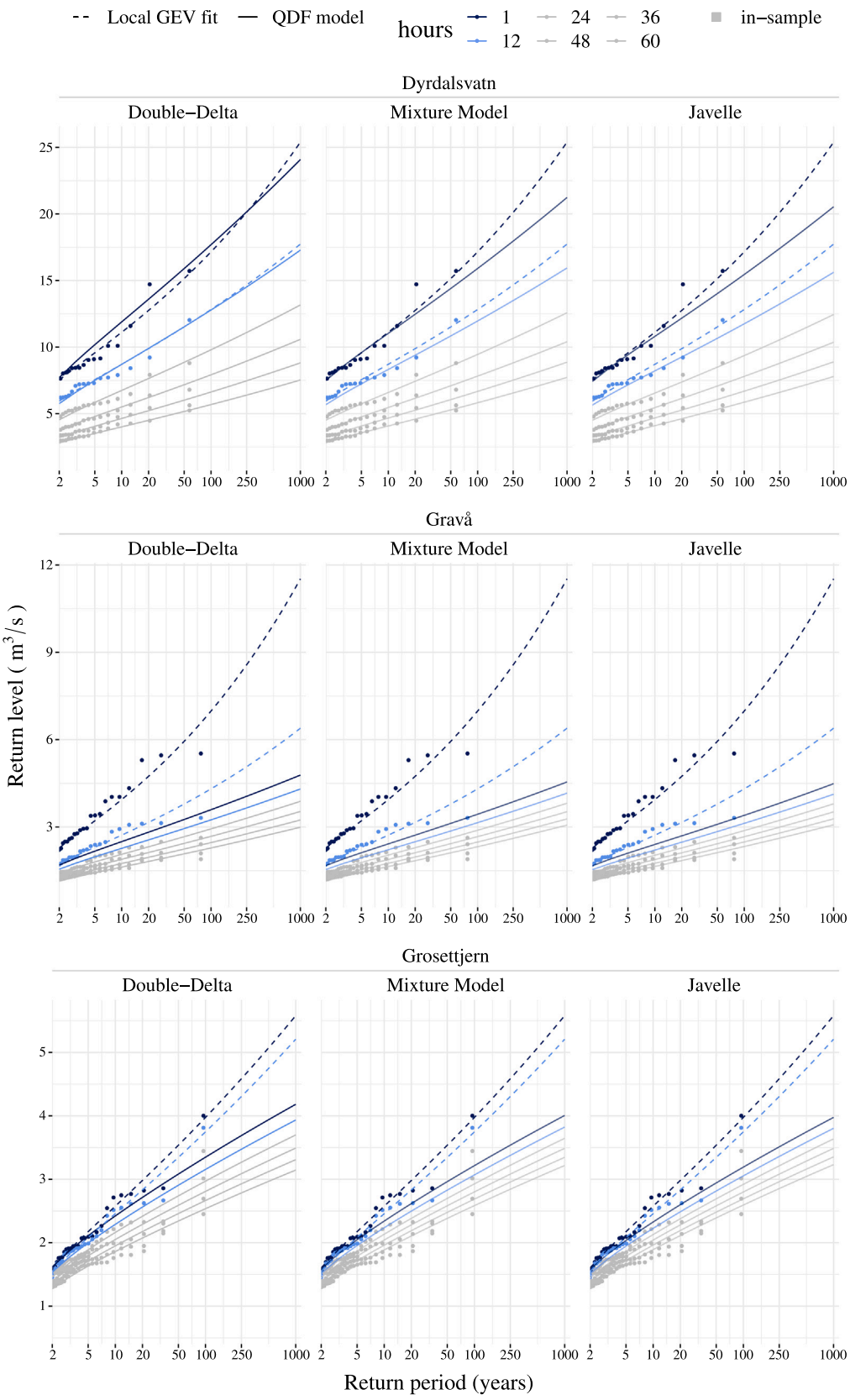

Fig. F.16. Out-of-sample return level plots for stations Dyrdalsvatn, Gravå, and Grosettjern.

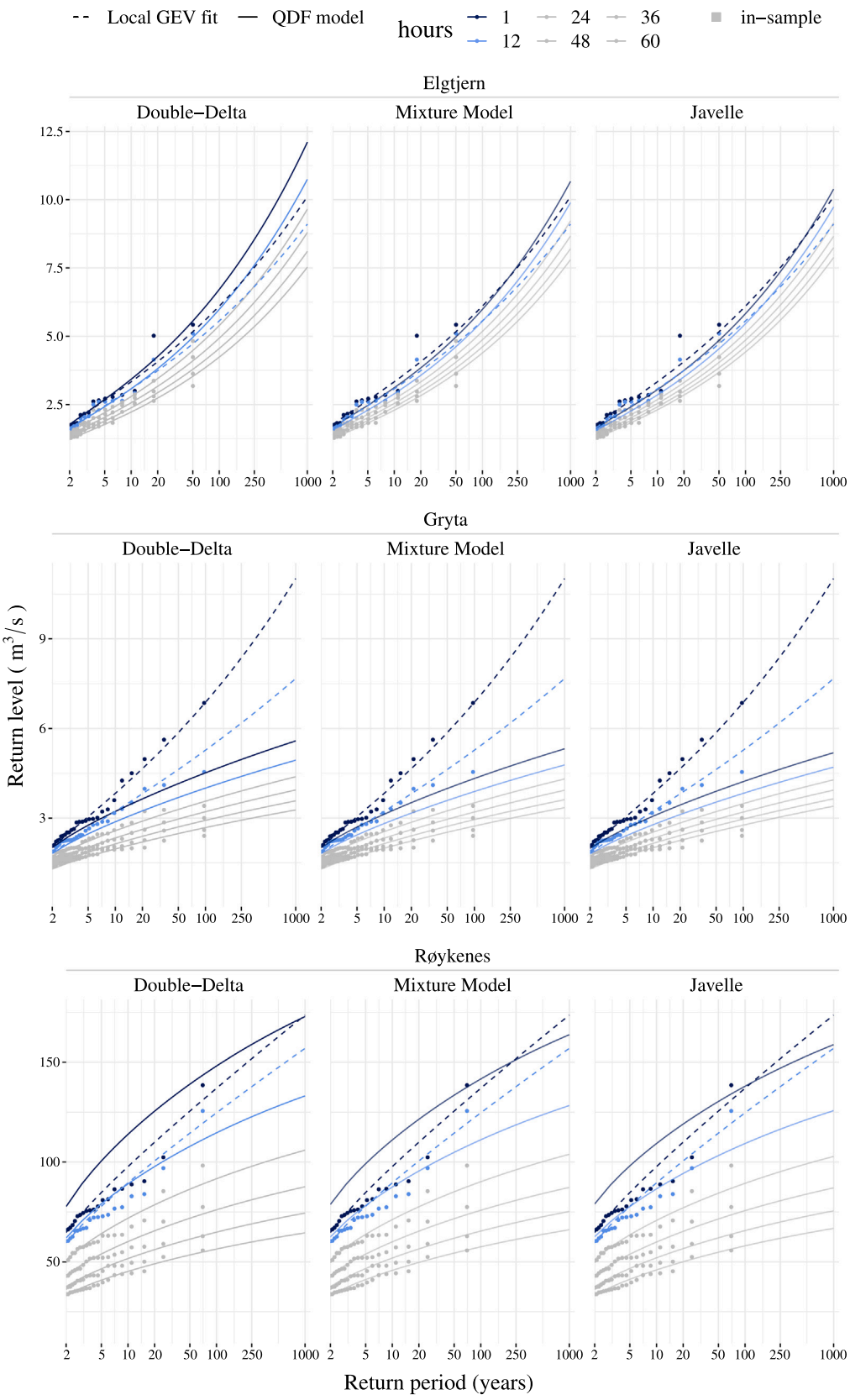

Fig. F.17. Out-of-sample return level plots for stations Elgtjern, Gryta, and Røykenes.

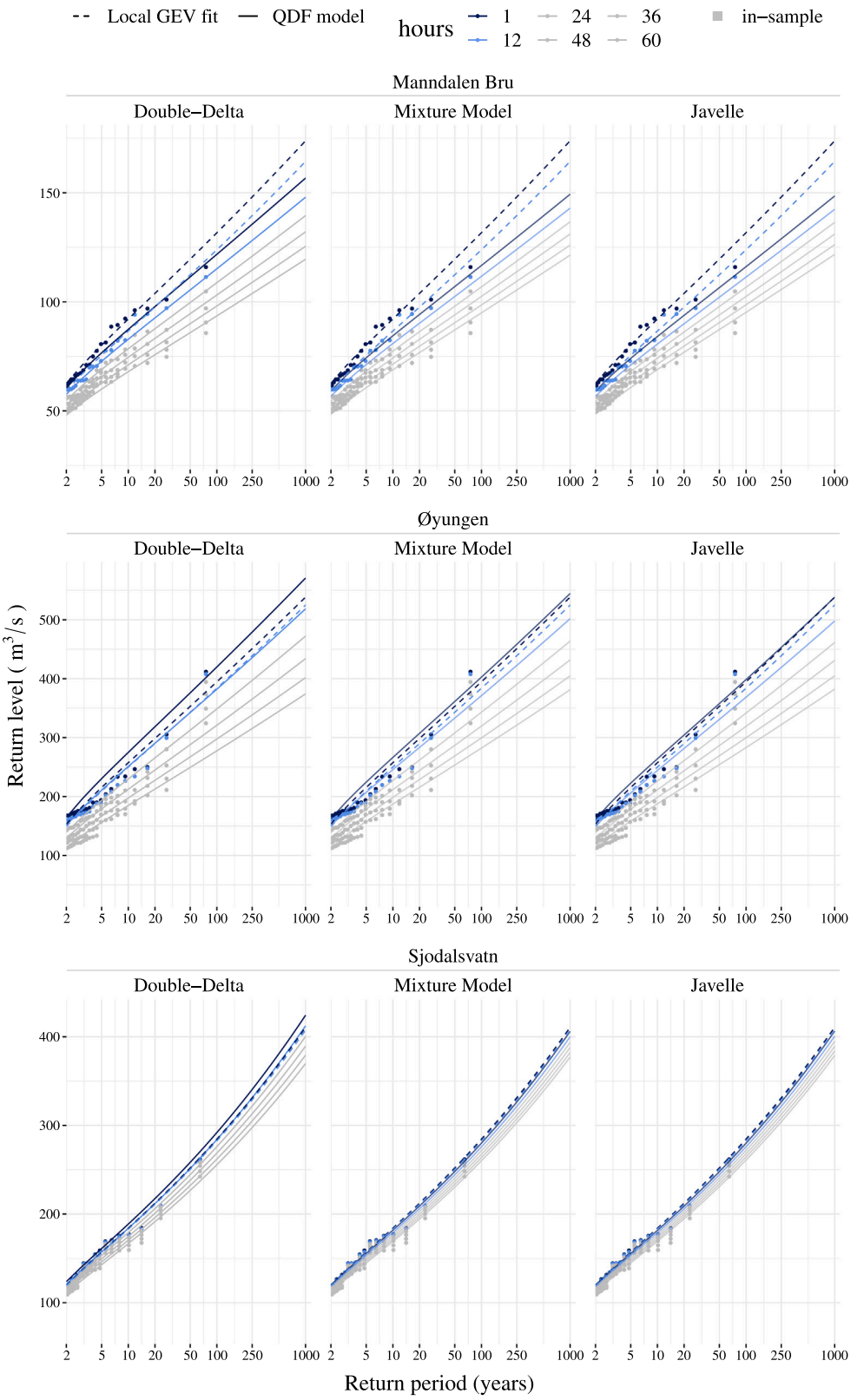

Fig. F.18. Out-of-sample return level plots for stations Manndalen Bru, Øyungen, and Sjodalsvatn.

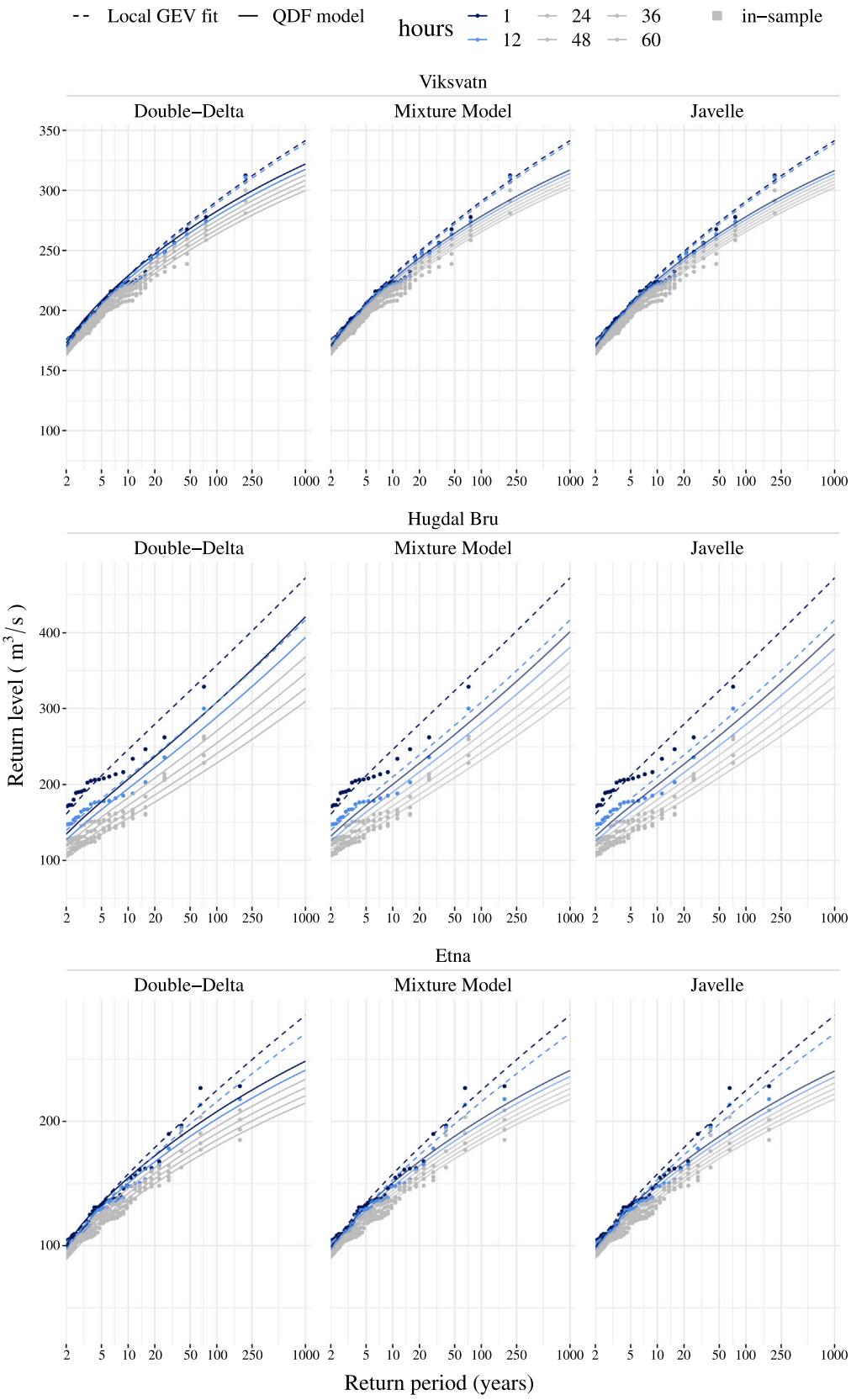

Fig. F.19. Out-of-sample return level plots for stations Viksvatn, Hugdal Bru, and Etna.

References

Alfieri, L., Bisselink, B., Dottori, F., Naumann, G., de Roo, A., Salamon, P., Wyser, K., Feyen, L., 2017. Global projections of river flood risk in a warmer world. Earth's Future 5 (2), 171–182.

Ball, J., Babister, M., Nathan, R., Weeks, W., Weinmann, E., Retallick, M., Testoni, I. (Eds.), 2019. Australian Rainfall and Runoff: A Guide to Flood Estimation. Commonwealth of Australia.

Balocki, J.B., Burges, S.J., 1994. Relationships between n-day flood volumes for infrequent large floods. J. Water Resour. Plan. Manag. 120 (6), 794–818.

Breinl, K., Lun, D., Müller-Thomy, H., Blöschl, G., 2021. Understanding the relationship between rainfall and flood probabilities through combined intensity-duration-frequency analysis. J. Hydrol. 602 (March), 126759. http://dx.doi.org/10.1016/j.jhydrol.2021.126759.

Castellarin, A., Kohnová, S., Gaál, L., Fleig, A., Salinas, J., Toumazis, A., Kjeldsen, T., Macdonald, N., 2012. Review of Applied Statistical Methods for Flood Frequency Analysis in Europe. The Centre for Ecology and Hydrology.

Castro-Camilo, D., Huser, R., Rue, H., 2022. Practical strategies for generalized extreme value-based regression models for extremes. Environmetrics e2742.

Cheng, L., AghaKouchak, A., 2014. Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate. Sci. Rep. 4 (1), 1–6.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer.

Courty, L.G., Wilby, R.L., Hillier, J.K., Slater, L.J., 2019. Intensity-duration-frequency curves at the global scale. Environ. Res. Lett. 14 (8), 084045.

Crochet, P., 2012. Flood-Duration-Frequency Modeling Application to Ten Catchments in Northern Iceland. Technical Report, p. 50, URL: http://www.vedur.is/media/2012_006.pdf.

Cunderlik, J.M., Jourdain, V., Quarda, T.B., Bobée, B., 2007. Local non-stationary flood-duration-frequency modelling. Can. Water Resour. J. 32 (1), 43–58. http://dx.doi.org/10.4296/cwrj3201043.

Cunderlik, J.M., Ouarda, T.B., 2006. Regional flood-duration–frequency modeling in the changing environment. J. Hydrol. 318 (1–4), 276–291.

Dalrymple, T., 1960. Flood-frequency analyses. Manual of hydrology part 3. Flood-flow techniques. Usgpo 1543-A, 80, URL: http://pubs.usgs.gov/wsp/1543a/report.pdf.

Ding, J., Haberlandt, U., Dietrich, J., 2015. Estimation of the instantaneous peak flow from maximum daily flow: a comparison of three methods. Hydrol. Res. 46 (5), 671–688.

Engeland, K., Glad, P., Hamududu, B.H., Li, H., Reitan, T., Stenius, S.M., 2020. Lokal og regional flomfrekvensanalyse. Technical Report, NVE.

Engeland, K., Schlichting, L., Randen, F., Nordtun, K.S., Reitan, T., Wang, T., Holmqvist, E., Voksø, A., Eide, V., 2016. Utvalg og kvalitetssikring av flomdata forflomfrekvensanalyser. Technical Report, NVE.

England, J., Cohn, T., Faber, B., Stedinger, J., Thomas, Jr., W., Veilleux, A., Kiang, J., Mason, Jr., R., 2019. Guidelines for Determining Flood Flow Frequency—Bulletin 17C. U.S. Geological Survey Techniques and Methods, http://dx.doi.org/10.3133/tm4B5.

Fauer, F.S., Ulrich, J., Jurado, O.E., Rust, H.W., 2021. Flexible and consistent quantile estimation for intensity–duration–frequency curves. Hydrol. Earth Syst. Sci. 25 (12), 6479–6494.

Field, C.B., Barros, V., Stocker, T.F., Dahe, Q., 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Filipova, V., Lawrence, D., Skaugen, T., 2019. A stochastic event-based approach for flood estimation in catchments with mixed rainfall and snowmelt flood regimes. Nat. Hazards Earth Syst. Sci. 19 (1), 1–18.

Fill, H.D., Steiner, A.A., 2003. Estimating instantaneous peak flow from mean daily flow data. J. Hydrol. Eng. 8 (6), 365–369. http://dx.doi.org/10.1061/(ASCE)1084-0699(2003)8:6(365).

Fischer, S., 2018. A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons. J. Appl. Stat. 45 (15), 2831–2847.

Fisher, R.A., 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, Vol. 24, No. 2. Cambridge University Press, pp. 180–190.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis Third Edition. Chapman and Hall/CRC, http://dx.doi.org/10.1201/b16018.

Gräler, B., Van Den Berg, M., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., Verhoest, N., 2013. Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. Hydrol. Earth Syst. Sci. 17 (4), 1281–1296.

Guo, S., Xiong, L., Chen, J., Guo, S., Xia, J., Zeng, L., Xu, C.Y., 2022. Nonstationary regional flood frequency analysis based on the Bayesian method. Water Res. Manag. 1–23.

Gupta, V.K., Waymire, E., 1990. Multiscaling properties of spatial rainfall and river flow distributions. J. Geophys. Res.: Atmos. 95 (D3), 1999–2009.

Huard, D., Mailhot, A., Duchesne, S., 2010. Bayesian estimation of intensity–duration–frequency curves and of the return period associated to a given rainfall event. Stoch. Environ. Res. Risk Assess. 24 (3), 337–347.

Javelle, P., Grésillon, J., Galéa, G., 1999. Discharge-duration-frequency curves modeling for floods and scale invariance. Sci. Laterre Des Planet 329, 39–44.

Javelle, P., Ouarda, T.B.M.J., Bob, B., 2003. Spring flood analysis using the flood-duration – frequency approach : application to the provinces of Quebec and Ontario, Canada. 3736 (June 2002), 3717–3736. http://dx.doi.org/10.1002/hyp.1349.

Javelle, P., Ouarda, T.B., Lang, M., Bobée, B., Galéa, G., Grésillon, J.M., 2002. Development of regional flood-duration-frequency curves based on the index-flood method. J. Hydrol. 258 (1–4), 249–259. http://dx.doi.org/10.1016/S0022-1694(01)00577-7.

Jenkinson, A.F., 1955. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Q. J. R. Meteorol. Soc. 81 (348), 158–171.

Jurado, O.E., Ulrich, J., Scheibel, M., Rust, H.W., 2020. Evaluating the performance of a max-stable process for estimating intensity-duration-frequency curves. Water 12 (12), 3314.

Kobierska, F., Engeland, K., Thorarinsdottir, T., 2018. Evaluation of design flood estimates - a case study for Norway. Hydrol. Res. 49 (2), 450–465. http://dx.doi.org/10.2166/nh.2017.068.

Koutsoyiannis, D., Kozonis, D., Manetas, A., 1998. A mathematical framework for studying rainfall intensity-duration-frequency relationships. J. Hydrol. 206 (1–2), 118–135.

Lussana, C., Tveito, O.E., Dobler, A., Tunheim, K., 2019. seNorge_2018, daily precipitation, and temperature datasets over Norway. Earth Syst. Sci. Data 11 (4), 1531–1551.

Markiewicz, I., 2021. Depth–duration–frequency relationship model of extreme precipitation in flood risk assessment in the Upper Vistula Basin. Water 13 (23), 3439.

Martins, E.S., Stedinger, J.R., 2000. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resour. Res. 36 (3), 737–744.

Midtømme, G.H., 2011. Retningslinjer for flomberegninger 2011. Technical Report 4/2011, NVE, pp. 1–66.

Onyutha, C., Willems, P., 2015. Empirical statistical characterization and regionalization of amplitude–duration–frequency curves for extreme peak flows in the Lake Victoria Basin, East Africa. Hydrol. Sci. J. 60 (6), 997–1012. http://dx.doi.org/10.1080/02626667.2014.898846.

Renima, M., Remaoun, M., Boucefiane, A., Sadeuk Ben Abbes, A., 2018. Regional modelling with flood-duration-frequency approach in the middle Cheliff watershed. J. Water Land Dev. 36 (1), 129–141. http://dx.doi.org/10.2478/jwld-2018-0013.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 59 (4), 731–792.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods. Springer, New York, NY.

Robson, A., Reed, D., 1999. Flood Estimation Handbook. Vol. 3: Statistical Procedures for Flood Frequency Estimation. Institute of Hydrology, p. 40.

Sælthun, N.R., Tveito, O.E., Bønsnes, T.E., Roald, L.A., 1997. Regional flomfrekvens-analyse for norsk vassdrag. Technical Report, NVE.

Saloranta, T., 2014. New Version (v.1.1.1) of the seNorge Snow Model and Snow Maps for Norway. Technical Report, Norges Vassdrags og Energidirektorat (NVE).

Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT 10 (1), 33–60.

Sherwood, J.M., 1994. Estimation of volume-duration-frequency relations of ungaged small urban streams in Ohio 1. JAWRA J. Am. Water Resour. Assoc. 30 (2), 261–269.

Thorarinsdottir, T.L., Gneiting, T., Gissibl, N., 2013. Using proper divergence functions to evaluate climate models. SIAM/ASA J. Uncertain. Quant. 1 (1), 522–534. http://dx.doi.org/10.1137/130907550.

Ulrich, J., Fauer, F.S., Rust, H.W., 2021. Modeling seasonal variations of extreme rainfall on different timescales in Germany. Hydrol. Earth Syst. Sci. 25 (12), 6133–6149.

Van de Vyver, H., 2018. A multiscaling-based intensity–duration–frequency model for extreme precipitation. Hydrol. Process. 32 (11), 1635–1647.

Wilson, D., Fleig, A.K., Lawrence, D., Hisdal, H., Pettersson, L.E., Holmqvist, E., 2011. A Review of NVE's Flood Frequency Estimation Procedures, Vol. 9. Technical Report, Norges vassdrags -og energidirektorat.

World Meteorological Organization, 2009. Guide to Hydrological Practices, Volume II: Management of Water Resources and Application of Hydrological Practices. WMO Geneva, Switzerland.

Zaidman, M.D., Keller, V., Young, A.R., Cadman, D., 2003. Flow-duration-frequency behaviour of british rivers based on annual minima data. J. Hydrol. 277 (3–4), 195–213.

Paper II

Regional index flood estimation at multiple durations with generalized additive models

Paper III

Regional flood frequency analysis at multiple durations: a comparison of duration consistency in quantile and parameter regression techniques