Flood-duration-frequency modelling with adaptive tail behaviour: A Bayesian approach

Danielle Barna[1], Kolbjørn Engeland[1], Thordis L. Thorarinsdottir[2], and Chong-Yu Xu[3]

[1]Norwegian Water and Energy Directorate
[2]Norwegian Computing Center
[3]University of Oslo

February 27, 2023

Abstract

Flood frequency analysis is a statistical approach for estimation of design flood values. 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. Flood magnitude is here typically taken as peak flow from an instantaneous discharge series. However, this univariate approach can be somewhat artificial as a flood event is not described by its peak flow alone. A relatively simple extension of traditional flood frequency models can be found in flood-duration-frequency, or QDF, models. QDF models take flood magnitude to be a product of peak flow and duration and are analogous to intensity-duration-frequency curves for precipitation. In an application to 12 locations in Norway, we assess how three different QDF models capture relationships between floods of different duration. Incorporating dependence on return period in the ratio between growth curves 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 flood durations. Overall, we find the choice of durations used to fit the QDF model is a highly influential aspect of the modeling process. Users should be aware that the choice of which durations to fit the model with will always be a qualitative choice that is only partially mitigated by adding extra flexibility to the models.

Flood-duration-frequency modelling with adaptive tail behaviour: A Bayesian approach

D. M. Barna[1,2], K. Engeland[1], T. L. Thorarinsdottir[3], C.-Y. Xu[2]

[1]Norwegian Water Resources and Energy Directorate, Oslo, Norway.
[2]Department of Geosciences, University of Oslo, Oslo, Norway.
[3]Norwegian Computing Center, Oslo, Norway.

Key Points:

- QDF models with duration-dependency in growth curve ratio better estimate event sizes than QDF models without
- QDF models are highly sensitive to the choice of input durations used to fit the models
- A Bayesian inference approach provides direct quantification of flood estimation uncertainty

Corresponding author: D. M. Barna, `daba@nve.no`

Abstract

Flood frequency analysis is a statistical approach for estimation of design flood values. 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. Flood magnitude is here typically taken as peak flow from an instantaneous discharge series. However, this univariate approach can be somewhat artificial as a flood event is not described by its peak flow alone. A relatively simple extension of traditional flood frequency models can be found in flood-duration-frequency, or QDF, models. QDF models take flood magnitude to be a product of peak flow and duration and are analogous to intensity-duration-frequency curves for precipitation. In an application to 12 locations in Norway, we assess how three different QDF models capture relationships between floods of different duration. Incorporating dependence on return period in the ratio between growth curves 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 flood durations. Overall, we find the choice of durations used to fit the QDF model is a highly influential aspect of the modeling process. Users should be aware that the choice of which durations to fit the model with will always be a qualitative choice that is only partially mitigated by adding extra flexibility to the models.

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 flooding. Such estimation can be undertaken in one of three general ways (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 rainfall records 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.

Traditionally, the 'flood events' in FFA are simply taken to be the annual maximum values from an instantaneous or mean daily streamflow series (Cunderlik et al., 2007). However, this univariate approach can be somewhat artificial as a flood event is not described by its peak flow alone—the volume and duration of the flood also matter in terms of its impact (Hettiarachchi et al., 2018) and are routinely needed in applications such as reservoir operation and flood damage assessment (Merz et al., 2010). Multivariate frequency analysis of flood return periods often requires large amounts of data and can be prohibitively complex (Gräler et al., 2013) but a relatively simple extension of FFA can be found in QDF, or Flood-Duration-Frequency, models. QDF models take flood magnitude to be a product of peak flow and duration and are based in the literature surrounding Intensity-Duration-Frequency (IDF) curves for precipitation (Javelle et al., 2002).

In the QDF approach, annual maxima are sampled from discharge series averaged over different durations. An extreme value distribution (usually the generalized extreme value, or GEV, distribution) is fit to these annual maxima, and a relationship between the durations and the fitted distributions is quantified 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 flood 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 flood durations is of interest. In addition, the parametric nature of the QDF model allows for extrapolation to unobserved flood 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 & Willems, 2015). 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.

Despite its similarities to the widely adopted IDF model (Cheng & AghaKouchak, 2014), a review of flood estimation practices in Europe, Australia and the USA reveals QDF models are not often applied for design flood estimation (Ball et al., 2019; England et al., 2019; Castellarin et al., 2012; Robson & Reed, 1999). In Australia, estimates of extreme floods are derived using rainfall-based flood estimation methods, where design floods are calculated separately for each duration by utilizing critical rainfall durations that produce the maxima for flood characterizations of interest (Nathan & Weinmann, 2019). In the USA, flood frequency estimates from durations other than the instantaneous flood are obtained by statistically estimating the flood frequency relationship on aggregated data (England et al., 2019; Lamontagne et al., 2012). In Europe, a wide variety of methodologies are used—most of which parallel those in Australia and the USA—to accommodate differing flood durations, but only France makes mention of QDF models (Castellarin et al., 2012). In Norway specifically, analysis of critical flood durations (typically longer duration events for dam safety analyses) is carried out via rainfall-runoff models where the appropriate storm duration is selected (Wilson et al., 2011). Note that consistency between return levels of different flood durations is not enforced by any of these methods apart from the QDF model; the consistency issue is generally addressed by noting the need to defer to expert judgement (Castellarin et al., 2012; England et al., 2019), by performing a comparison of flood frequency analysis and rainfall-runoff output (Wilson et al., 2011; Ball et al., 2019) or by post-processing of the design flood values (Nathan & Weinmann, 2019).

Design flood estimation is often most concerned with estimation of the flood stemming from the instantaneous streamflow series, since this is the scenario that typically produces the highest return level. Statistical estimation of this design value, however, poses a challenge since flood series of length appropriate for flood frequency analysis 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). However, there exist methodologies for scaling daily data to approximate the instantaneous peak flow (Ding et al., 2015; Fill & Steiner, 2003). In Norway, scaling between daily and instantaneous peak flows is performed by assuming similarity between the growth curve for daily flow at the site of interest and the peak flow curve at another site with comparable properties. At ungauged locations, peak flows can be estimated via regression equations on relevant catchment properties. Wilson et al. (2011) notes the uncertainty in both of these methods is likely to be large and difficult to reconcile with the uncertainty inherent to FFA. While QDF models seek to consistently estimate a range of durations simultaneously and thus are formulated to address a slightly different question than daily to instantaneous peak flow estimation, their performance when estimating floods at subdaily unobserved flood durations is of particular interest.

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 the ratio between peak and daily values to dependent on return period to different degrees. For comparison, three-parameter GEV distributions are fit independently to each flood duration in line with the current guidelines (Midtømme, 2011; England et al., 2019).

Estimation methodologies for QDF models have to cope with the typical challenges that come with fitting extreme value models. Extreme value models are prone to parameter estimation difficulties stemming from the inherent sparsity of threshold excess data (Scarrott & MacDonald, 2012), and, when the GEV distribution is used, enforcement of a support condition that depends on all parameters and the data. This last condition is particularly problematic under QDF modelling since the introduction of multiple durations means the support must be enforced at each duration individually. In the QDF literature this has typically been dealt with by introducing a two-step estimation procedure where a single parameter representing the "characteristic duration" is estimated beforehand and then used in tandem with standard frequentist estimation techniques to estimate the remaining parameters of the extreme value distribution (Javelle et al., 2002; Cunderlik et al., 2007). However, such two-step estimation does not allow for easily accessible uncertainty estimates, and, moreover, requires additional assumptions if the model is to be used in a regional context (Cunderlik & Ouarda, 2006). Since the models presented here (1) require uncertainty estimates to inform discussion around flood design values and (2) are intended to form the basis of a regional flood model, we adopt a Bayesian estimation approach. Bayesian estimation of IDF models is well established and the advantages particular to this approach (accessible uncertainty estimation, scaling to regional models via hierarchical Bayesian approaches, ability to add information through prior distributions) have been shown to be relevant in estimation of precipitation extremes (Cheng & AghaKouchak, 2014; Huard et al., 2010).

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).

2 Data

The flood data came from 12 streamflow stations in Norway that have at least 20 years of quality-controlled data with minimal influence from reservoirs and other installations that might alter the natural streamflow. 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 and relevant catchment properties are shown in Figure 1. The selected stations show a diversity of locations, catchment sizes and flood generating processes, allowing us to evaluate the QDF models on a diversity of flood behaviours. 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 Figure 1 this is illustrated as the average fractional rain contribution to each flood event. The average rain contribution was estimated by calculating the ratio of accumulated rain and snowmelt 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.

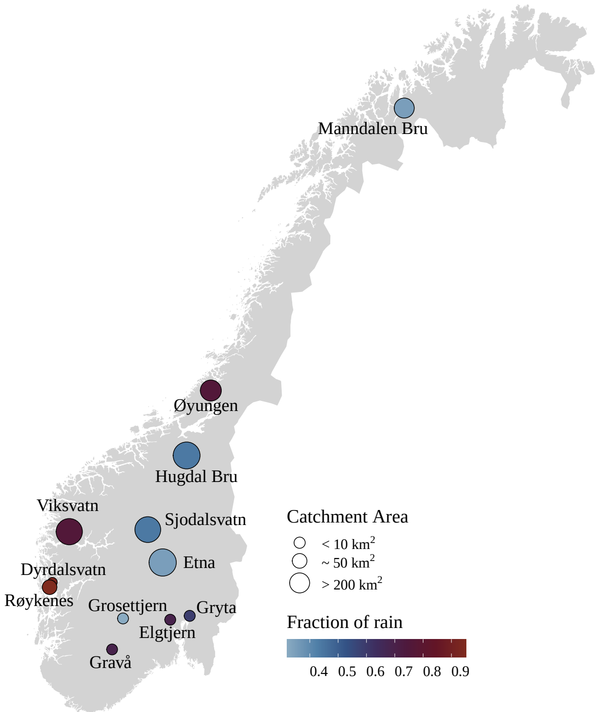


Figure 1: Locations of twelve gauging stations used in study. Catchment area and fraction of rain contribution to flood are also indicated.

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. We used the daily data in addition to the more high-resolution data from the last five decades for only two stations (Etna and Viksvatn, both large, primarily snowmelt driven catchments). For all the remaining stations we used data from approximately 1970 to present day, which is collected via a combination of limnigraph and digital readings. 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 primary quality control by the hydrometric section at NVE and are corrected for ice jams. Any year with less than 300 days of data was discarded.

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 flood 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.

These sets of maxima produced under different d are dependent. The QDF model does not account for this dependency. This is an intentional modeling decision. While methodologies exist to capture the dependence structure between extreme events in these types of models–for example, the copula-based methods of Singh and Zhang (2007), the max-stable based model of Jurado et al. (2020) and the stochastic process theory based model of Van de Vyver (2018), all of which are discussed in Section 5–Figure 2 illustrates several artifacts introduced by QDF data processing that confound our ability to model the dependencies between maxima, particularly at ungauged locations. The model proposed in this paper is intended to form the basis of a regional model and thus needs a methodology that can be extended to ungauged catchments.

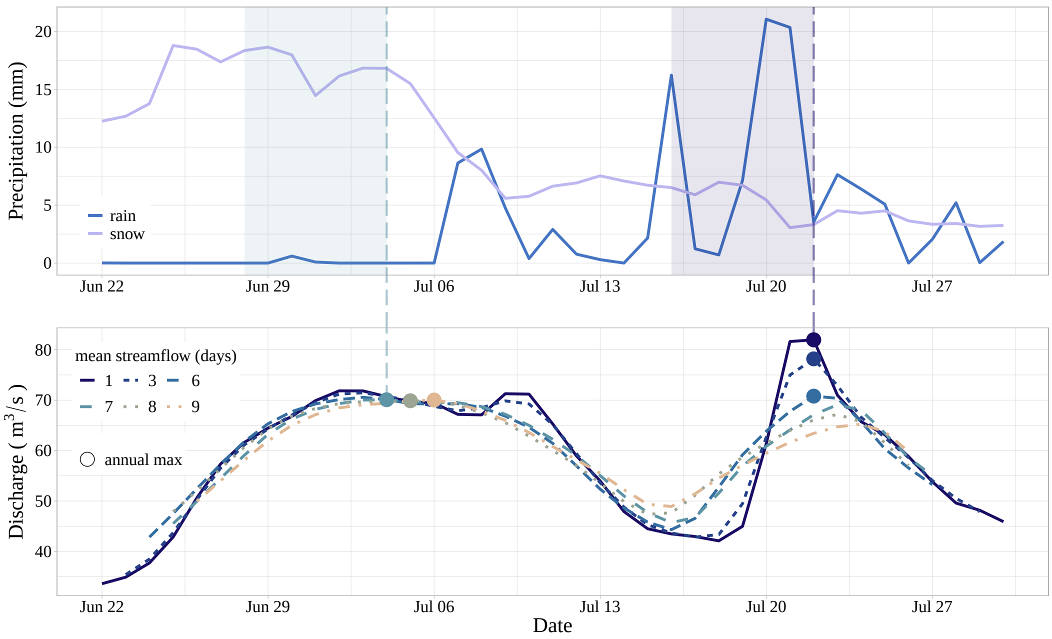


Figure 2: Figure showing two artifacts introduced by QDF data processing: (1) annual maxima are not guaranteed to decrease as the duration of the averaging window is increased and (2) annual maxima for each duration are not always issued from the same flood event. Here we see the annual maxima from durations less than 7 days originate in a primarily rain-driven flood in mid-July (top panel) while annual maxima from durations greater than 7 days come from a smoother, snowmelt-driven flood at the beginning of July. The shaded areas in the top panel show the window of time from which the flood generating process is calculated. Data is from Sjodalsvatn gauging station, for the year 2009.

First, 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 peaks of similar volume occurring within a short time period of each other, or a particularly wide and flat-topped flood) are not directly relatable to catchment properties. Secondly, the floods for different durations are in some cases based on 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 flood events 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 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 & 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 with long data series. 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 $\xi = 0$ 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

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

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

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

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 Equation 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.

3.2.1 Original QDF model

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

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

where

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

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 Δ from Javelle's original QDF model is used here for numerical stability during estimation. The value of the Δ parameter reflects the shape of the hydrograph. A high value for Δ indicates a flashy/peaked hydrograph with a pronounced duration dependency for the median flood, whereas a value close to zero indicates 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 Equation 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 deals only with a first-order distribution of precipitation events and does not rely on this theory base (Koutsoyiannis et al., 1998).

Since only the magnitude of the flood (η) is duration-dependent in the model in Equation 9, the underlying assumption is that the slope of the growth curve does not change with flood duration. Breaking this assumption (as the extended QDF model in the next section does) requires breaking the separable functional dependence. However, as discussed in Section 5, flood events are unlikely to follow a single stochastic process (Viglione et al., 2010; Gaál et al., 2012), relaxing the need to draw on the related theory base.

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 flood 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 (Equation 3). Under Double-Delta the annual flood 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

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

with constraint

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

The constraint on the Delta parameters reflects the relationship between sets of flooding events; 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; 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.

The model is a weighted average of the Double-Delta and Javelle models such that the density of the annual maximum flood events is given by

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

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 Double-Delta and Javelle models, respectively.

Thus Equation 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 Equation 10 and the Javelle model in Equation 7.

3.3 Bayesian Framework

For the Javelle and Double-Delta models, Bayesian inference is performed using a Metropolis-Within-Gibbs algorithm (Robert & Casella, 2004). 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, 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

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

where $p(\cdot|\cdot)$ is the generic conditional distribution consistent with this joint specification and $m \in \{\text{DD}, \text{J}\}$, $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, 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 Equation 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 $\boldsymbol{\theta}$. To account for this, we employ a reversible jump MCMC algorithm, similar to the reversible jump methodology for normal mixtures described in Richardson and Green (1997). The reversible jump MCMC proceeds as follows:

1. updating $\boldsymbol{\theta}$:
   (a) if $m = \text{DD}$ update η^{DD}, else update η^{J};
   (b) if $m = \text{DD}$ update β^{DD}, else update β^{J};
   (c) if $m = \text{DD}$ update ξ^{DD}, else update ξ^{J};
   (d) if $m = \text{DD}$ 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

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

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

3.3.1 Posterior return levels

The Markov chains detailed above return a collection of R samples

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

where R is the total number of iterations in the MCMC with a suitable number of burn-in samples removed. Under the mixture model, $\boldsymbol{\theta}$ can be either $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$ dependent on iteration r, while posterior samples under Double-Delta or Javelle will return only $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$, 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 flood 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 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 Equation 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 hour) durations from long durations (greater than or equal to 24 hours). 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 flood 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 flood 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.

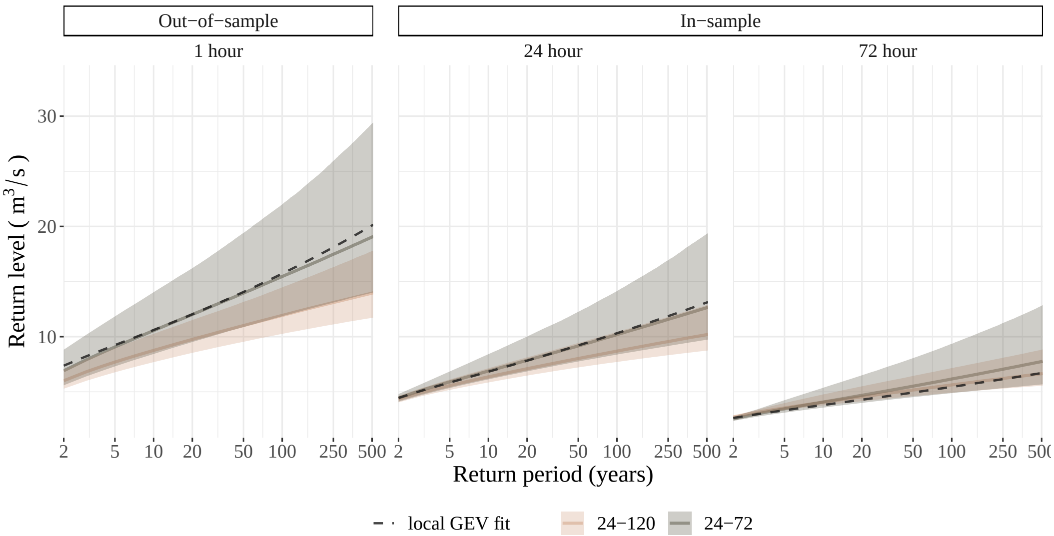


Figure 3: Return level plots from the Dyrdalsvatn gauging station using the Double-Delta model fit to two different data sets: one set with six durations [24, 36, 48, 72, 96, 120 hours] and one set with four durations [24, 36, 48, 72 hours]. The model fit to the six duration set is both overconfident and biased at shorter durations; the posterior mean return level estimates are consistently underestimated when compared to locally fit GEV models (dashed grey lines) and the 90% credible interval is artificially narrow and fails to capture the locally fit model for the 24 and 1 hour durations.

To test this, the models were fit under three different sets of data: two durations (24 and 36 hours); four durations (24, 36, 48, 72 hours); and six durations (24, 36, 48, 72, 96, 120 hours). 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 Figure 3. 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 grey lines) for the 24 and 1 hour 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 flood duration, where the in-sample flood durations are 1, 24, 48, and 72 hours. 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 Figures C1-C4.

4.2.1 Assessing model behavior using IQD

A comparison of in-sample IQD scores across stations, durations and methods is given in Figure 4. The scores are relatively similar across models–most points fall on or along the diagonals in the two plots in Figure 4. 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.

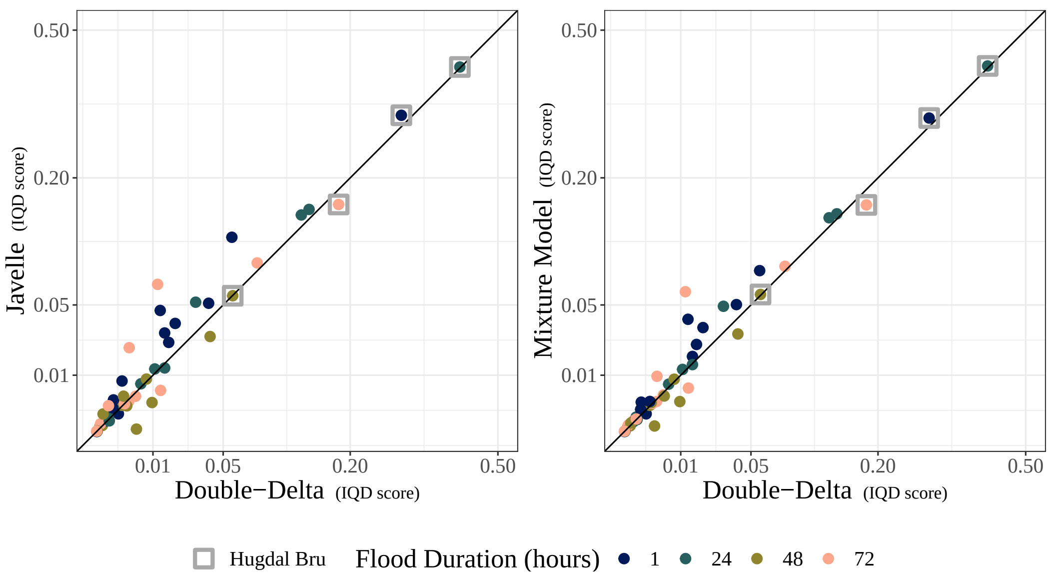


Figure 4: Model-to-model comparison of IQD scores for each station and in-sample duration. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, left panel) and the mixture model (right panel). Notable values are indicated.

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

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 hour	10/12	10/12
24 hours	9/12	9/12
48 hours	7/12	7/12
72 hours	7/12	8/12

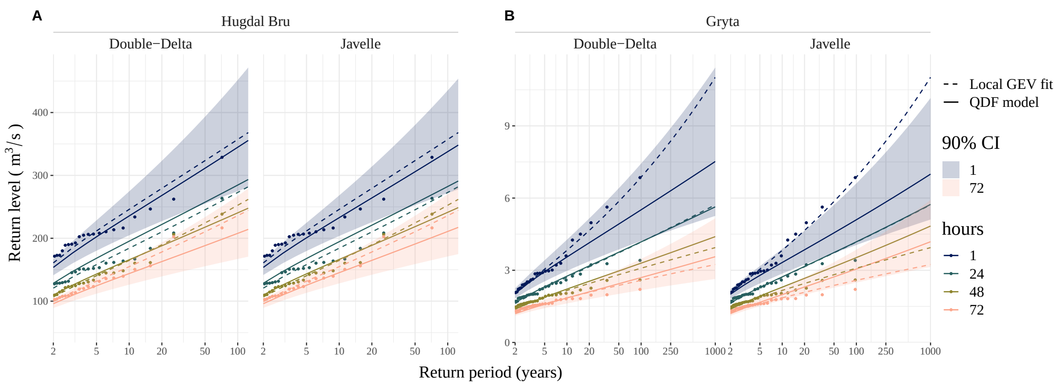


Figure 5: Return level plots showing two selected stations where QDF models differ substantially from the reference model on in-sample durations. (**A**) Hugdal Bru: the 1-hour floods with return period under 5 years are characterized by a diurnal melt-freeze cycle at this snowmelt-driven catchment; 1-hour 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.

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 hour durations at Hugdal Bru, displayed in panel A of Figure 5. 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

Equation 14). Thus the behavior of 1-hour 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 as such have a more regular relationship between durations. Flood durations above 24 hours (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 Figure 5, 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 hour duration to -0.037 for the 72 hour 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, under Javelle the 1 hour duration is underestimated and the 48 and 72 hour 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.

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	Comparison model		T
duration	Javelle	MM	
1 hour	11/12	11/12	100
24 hours	10/12	9/12	
48 hours	4/12	4/12	
72 hours	7/12	6/12	
1 hour	11/12	11/12	1000
24 hours	9/12	9/12	
48 hours	4/12	4/12	
72 hours	6/12	6/12	

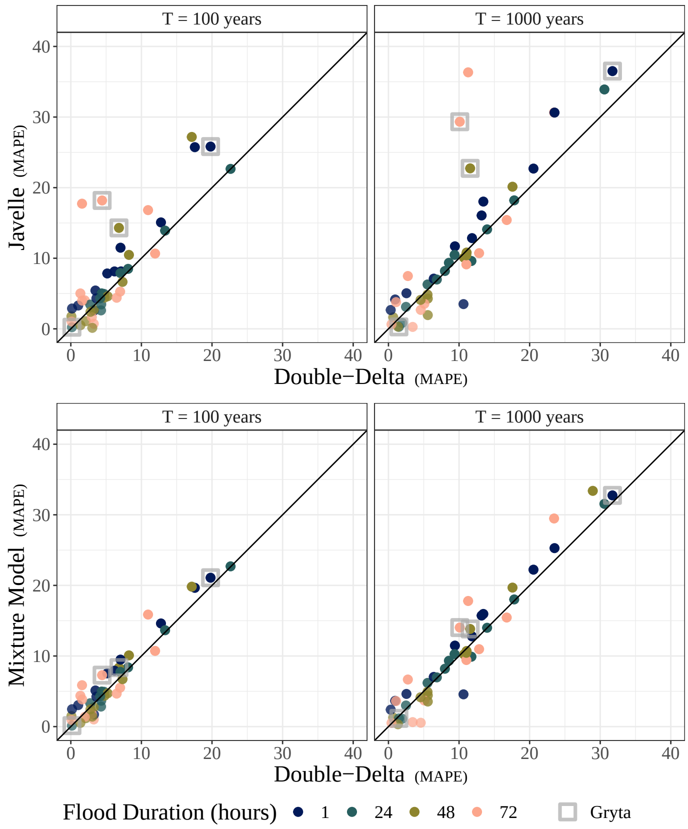


Figure 6: Model-to-model comparison of MAPE scores for each station and in-sample duration. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, top panels) and the mixture model (bottom panels). Notable values are labeled.

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 Figures 4 and 6. Javelle and the mixture model appear more similar when evaluated by the IQD than they do under the MAPE; 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 Figure 6 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 in Figure 6). The return level plots in panel (B) of Figure 5 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 hour 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 hours) and the resulting parameter estimates were used to predict the 1 and 12 hour 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 Figures D1-D4.

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. Figure 7 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 hour durations at Sjodalsvatn and the 1 hour duration at Dyrdalsvatn and Øyungen) where Double-Delta performs worse, as measured by the IQD, than the other two models. 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 hour duration at Røykenes. Difficulties fitting the sub-daily durations of Hugdal Bru are discussed in Section 4.2.1. The 1 hour duration at Røykenes exhibits a large change in shape parameter with an increase in duration like the station Gryta shown in panel B of Figure 5.

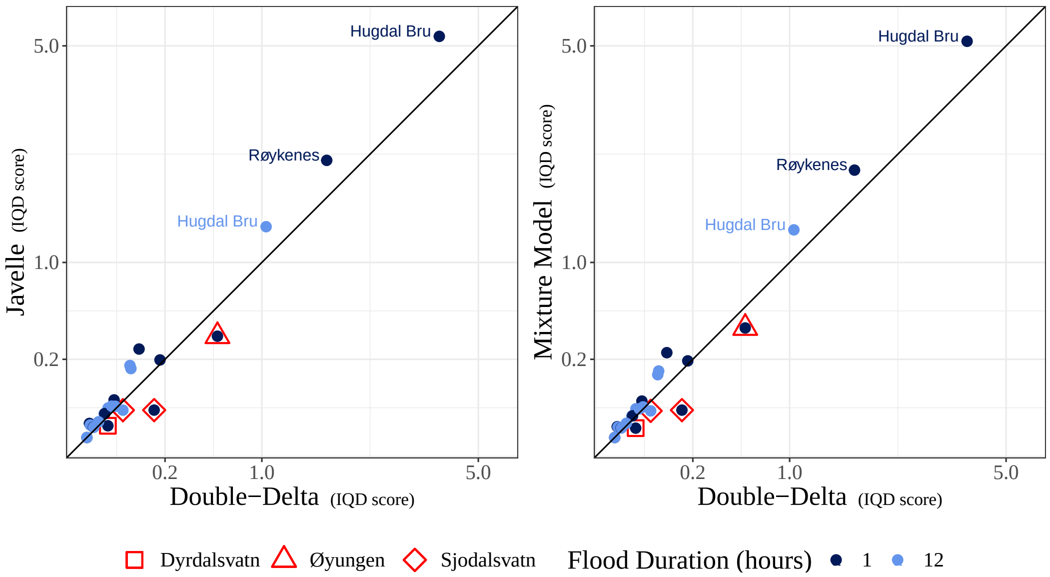


Figure 7: Model-to-model comparison of IQD scores for each station and both out-of-sample durations. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, left panel) and the mixture model (right panel). Notable values are labeled.

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 marked in red in Figure 8.

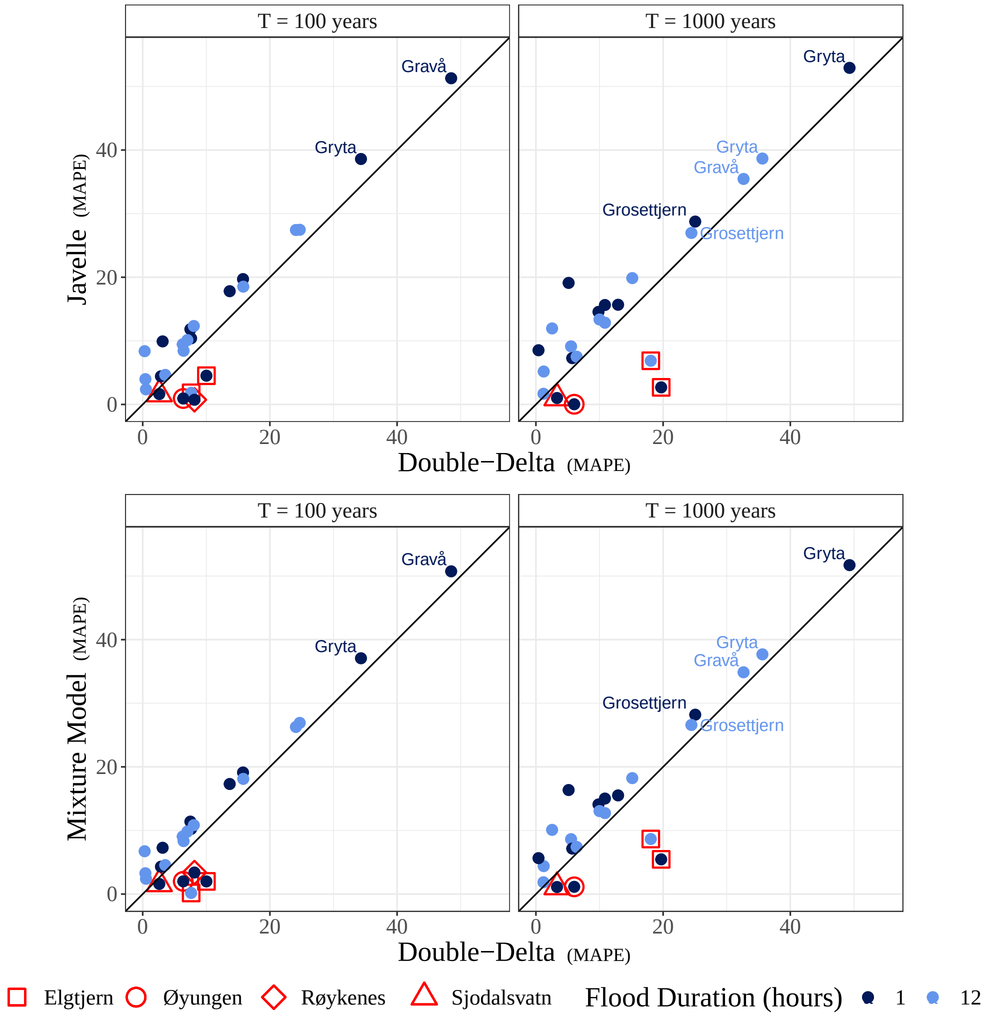


Figure 8: Model-to-model comparison of MAPE scores for each station and both out-of-sample durations. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, top panels) and the mixture model (bottom panels). Notable values are labeled.

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 A1 and B1, 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 Figure 5) 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 hour duration or the 12 and 24 hour duration while the rate of change of growth curve slope is less for durations above 24 hours; 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.

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 hours) where the 1 and 12 hour 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 hours) that are used to predict the 1 and 12 hour durations. The performance of each of these sets is evaluated at the 1 and 12 hour durations using both the IQD, as shown in Figure 9, and MAPE, as shown in Figure 10.

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 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 Figures 9 and 10). For the MAPE, this difference is on the order of ±5%.

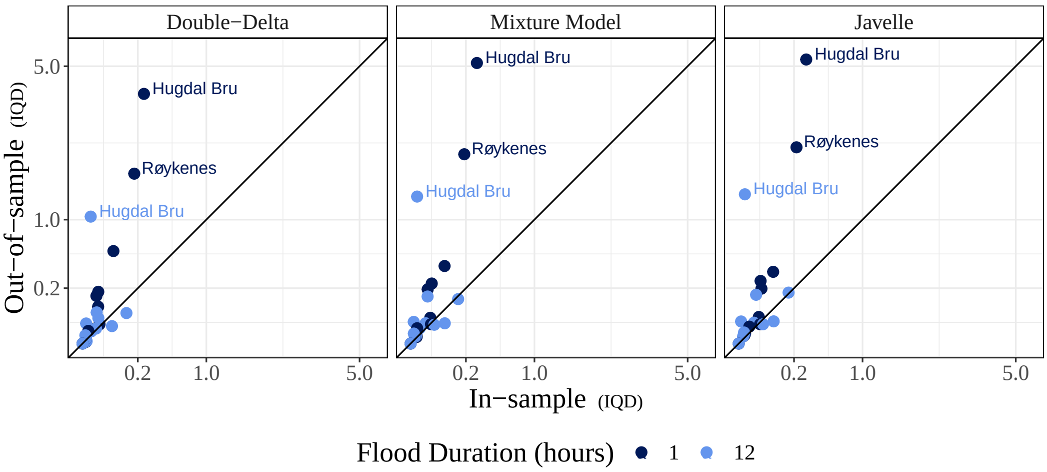


Figure 9: Comparison of IQD score when durations are either predicted (out of sample) or included in the model fitting set (in sample). The out-of-sample set was fit with durations 24, 36, 48, 60 hours and used to predict the 1 and 12 hour durations. The in-sample set was fit with durations 1, 12, 24, 36, 48, 60 hours. Notable values are labeled.

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 iden-

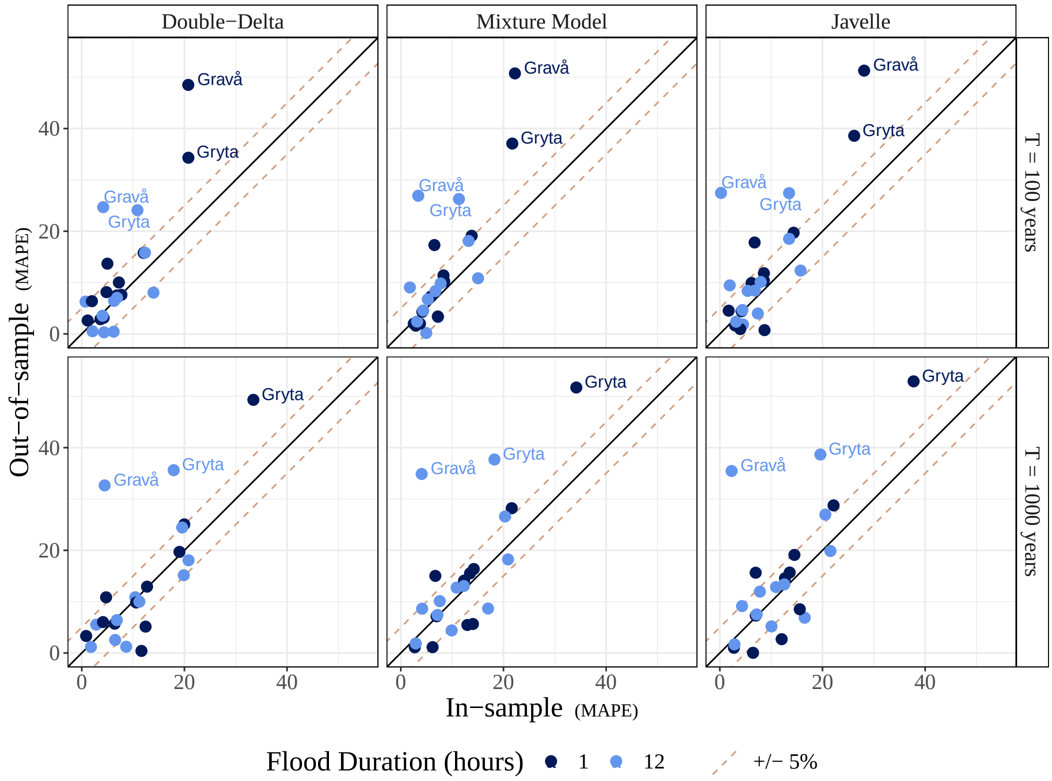


Figure 10: Comparison of mean absolute percent error when durations are either predicted (out of sample) or included in the model fitting set (in sample). The out-of-sample set was fit with durations 24, 36, 48, 60 hours and used to predict the 1 and 12 hour durations. The in-sample set was fit with durations 1, 12, 24, 36, 48, 60 hours. Notable values are labeled and dashed lines indicate ±5% difference from the diagonal.

tified 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 events with long return periods and/or short flood durations. Additionally, we tested the out-of-sample performance of QDF models on sub-daily floods 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 main contribution of the proposed Double-Delta 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 Figures C1-C4, the assumption of a constant ratio will commonly not hold, in particular, if the shortest duration of 1 hour 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.

The Double-Delta model assumes that the magnitude of the differences in return level either stays the same or increases with increasing return period. However, this behavior is not the only dependence structure we observe in our data, as illustrated in panel A of Figure 5. The example from Hugdal Bru shows that cases exist where the magnitude of differences between return levels of different duration may decrease rather than increase. This is one of two scenarios (the other being duration dependence in the shape parameter) where we observe large discrepancies between the QDF models and the reference model.

Methods exist to model the dependence structure between durations. These methods—which typically build dependence relationships into a dGEV model via use of a copula or a max-stable process—are an active area of research in the IDF modeling community (Jurado et al., 2020; Tyralis et al., 2019; Vinnarasi & Dhanya, 2019; Singh & Zhang, 2007). However, none of these methods explicitly address changes in dependence structure with return period, which is our observed source of difficulty with QDF models. It is possible that, if the dependency between events is greatest at high return periods, max-stable or copula based methods could provide an improvement in accuracy at these high quantiles as a by-product of modeling the dependence structure at all quantiles. Unfortunately, artifacts introduced by processing of the data for QDF modeling (described in Section 2.2, Figure 2) limit our ability to make statements about the increase or decrease of event dependence with return period and with duration. Other copula-based approaches to multivariate flood frequency analysis sidestep these artifacts by avoiding data averaging and instead work with extensive observed data series (Zhang & Singh, 2006) or long series of synthetic data (Gräler et al., 2013); peak discharge events from these series are then characterized by their discharge and duration, where the association between discharge and duration is then modelled by the copula. These approaches, however, are data-intensive, reliant on observing events of relevant duration, and sensitive to the choice of copula (Gräler et al., 2013; Zhang & Singh, 2006). This increases the burden on the practitioner and complicates the extension to ungauged catchments and unobserved flood durations.

In addition to the scenario described above, a second scenario where QDF models struggle to fit the data are situations where the shape parameter changes with duration. This situation, illustrated in panel B of Figure 5, is a known limitation of QDF modeling as these models assume a constant shape parameter across all durations. It would be technically possible to add duration dependence to the shape parameter of the models in Equations 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 ability of QDF models to predict sub-daily unobserved flood durations using daily or longer data is of particular interest due to the relevance of the instantaneous, or hourly, flood to design flood estimation and the prevalence of daily data in many of the longer flood records. When this behavior was tested in Sections 4.3 and 4.4, most stations returned results we find promising: the out-of-sample results are similar to the results obtained when the sub-daily durations are in-sample. However, a few stations demonstrated that the ability of QDF models to predict out-of-sample durations can be severely limited. Simply put, in certain situations the relationship between the in-sample floods does not inform the relationship at the out-of-sample floods. We suspect such situations arise in connection with the temporal scaling properties of flooding events.

The ways in which flood properties change with increasing event duration are complex. Flood duration incorporates many aspects of runoff generation and precipitation characteristics. This relationship is further complicated under FFA since all flood events are grouped regardless of their generating process and under QDF modeling since averaging introduces the possibility that flood events from different durations are not necessarily from the same flooding event.

The disconnect between sub-daily and long-duration flood events observed at some stations in this study parallels the work of Viglione et al. (2010). They found that short duration flood events tend to be controlled by temporal and spatial components working in concert with properties of the specific storm that generated the flood. Longer duration events, by contrast, are primarily controlled by temporal components. That is, different processes—that likely produce different dependency relationships between flood events—control floods at different durations. Corroborating this is the work of Gaál et al. (2012), which found different generating processes for the shortest and longest floods in an analysis of nearly 10,000 flood events at a variety of catchments in Austria. What is "short" and what is "long" will be specific to the catchment in question and defies an easy definition; however, it seems likely that we have found the boundary between "long" and "short" floods for the three stations that struggle to use QDF models to estimate sub-daily floods from daily data.

We note that this observed disconnect in temporal scaling properties of flood events, along with the work of Viglione et al. (2010) and Gaál et al. (2012), indicate that it is unlikely floods at different timescales are generated from the same stochastic process. As such a "multiscaling" model that attempts to relate the probabilistic properties of floods at two different timescales (such as the IDF model proposed in Van de Vyver (2018)) is not appropriate here.

Importantly, 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; as discussed in the previous paragraphs, 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 hour flood duration on the smallest catchments in this study and can therefore avoid the somewhat contrived scenarios where we use what are, for these catchments, only long-duration flood events to estimate the shortest durations.

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). Additionally, 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 (Figure 3).

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. 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. Extending the analysis presented in this paper to include more gauging stations is also a priority and an important component of developing a regional model. In addition to regionalization of the model, 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. And, finally, a natural follow-up question to this analysis using QDF models to predict sub-daily out-of-sample durations is "How good are QDF models at predicting short durations when compared to other methodologies designed for the purpose of estimating the instantaneous design flood?".

6 Conclusions

This paper proposes a five parameter (Double-Delta) QDF model based on the GEV distribution, where both flood magnitude and the ratio between growth curves may vary across flood durations. A Bayesian inference algorithm is developed where a four parameter QDF model, a five parameter QDF model, or a mixture of the two may be estimated. In a case study comprising 12 study locations in Norway, we analyze how the different QDF models capture the relationship between floods of different duration. The results suggest it is advantageous to include an adaptive tail behaviour in the QDF model. This advantage is particularly pronounced for events with long return periods and/or short flood durations. The Double-Delta model is also better at handling changes in the underlying statistical properties of floods at different durations, allowing for a wider range of durations to be included in the analysis. Overall, we found the QDF framework to be highly sensitive to the choice of durations used to fit the models. Users should be aware that the choice of input durations will always be a qualitative choice that is only partially mitigated by adding extra flexibility to 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. On the other hand, generating too many sets of dependent data to fit the model can produce results that are both biased and overconfident. When care is taken with these aspects, the 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.

Data Availability Statement

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). The authors would like to thank Thea Roksvåg and Alex Lenkoski for valuable discussions and Mads-Peter Dahl for help with data selection.

References

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

Ball, J., et al. (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. *Journal of Water Resources Planning and Management*, *120*(6), 794–818.

Barna, D. M. (2022, September). *12 selected stations from NVE Hydra II for QDF analysis* [Dataset]. Zenodo. doi: 10.5281/zenodo.7085557

Breinl, K., Lun, D., Müller-Thomy, H., & Blöschl, G. (2021). Understanding the relationship between rainfall and flood probabilities through combined intensity-duration-frequency analysis. *Journal of Hydrology*, *602*(March), 126759. doi: 10.1016/j.jhydrol.2021.126759

Castellarin, A., Kohnová, S., Gaál, L., Fleig, A., Salinas, J., Toumazis, A., ... Macdonald, N. (2012). Review of applied statistical methods for flood frequency analysis in europe.

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. *Scientific reports*, *4*(1), 1–6.

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

Crochet, P. (2012). *Flood-Duration-Frequency modeling Application to ten catchments in Northern Iceland* (Tech. Rep.). Retrieved from `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. *Canadian Water Resources Journal*, *32*(1), 43–58. doi: 10.4296/cwrj3201043

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

Dalrymple, T. (1960). Flood-Frequency Analyses. Manual of Hydrology Part 3. Flood-flow techniques. *Usgpo*, *1543-A*, 80. Retrieved from `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. *Hydrology Research*, *46*(5), 671–688.

Engeland, K., Glad, P., Hamududu, B. H., & Li, H. (2020). *Lokal og regional flomfrekvensanalyse* (Tech. Rep.). NVE.

England, J., Cohn, T., Faber, B., Stedinger, J., Thomas, J., W.O., Veilleux, A., ... Mason, J., R.R. (2019). *Guidelines for determining flood flow frequency—bulletin 17c.* U.S. Geological Survey Techniques and Methods. doi: https://doi.org/10.3133/tm4B5

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. *Natural Hazards and Earth System Sciences*, *19*(1), 1–18.

Fill, H. D., & Steiner, A. A. (2003). Estimating instantaneous peak flow from mean daily flow data. *Journal of Hydrologic Engineering*, *8*(6), 365-369. doi: 10.1061/(ASCE)1084-0699(2003)8:6(365)

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. , *24*(2), 180–190.

Gaál, L., Szolgay, J., Kohnová, S., Parajka, J., Merz, R., Viglione, A., & Blöschl, G. (2012). Flood timescales: Understanding the interplay of climate and catchment processes through comparative hydrology. *Water Resources Research*,

48(4).
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. doi: https://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. *Hydrology and Earth System Sciences*, *17*(4), 1281–1296.
Hettiarachchi, S., Wasko, C., & Sharma, A. (2018). Increase in flood risk resulting from climate change in a developed urban watershed–the role of storm temporal patterns. *Hydrology and Earth System Sciences*, *22*(3), 2041–2056.
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. *Stochastic Environmental Research and Risk Assessment*, *24*(3), 337–347.
Javelle, P., Gresillon, J., & Gala, G. (1999). Discharge-duration-frequency curves modeling for floods and scale invariance. *Sciences de laterre et des planet*, *329*, 39-44.
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. *Journal of Hydrology*, *258*(1-4), 249–259. doi: 10.1016/S0022-1694(01)00577-7
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. doi: 10.1002/hyp.1349
Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, *81*(348), 158–171.
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. *Hydrology Research*, *49*(2), 450–465. doi: 10.2166/nh.2017.068
Koutsoyiannis, D., Kozonis, D., & Manetas, A. (1998). A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of hydrology*, *206*(1-2), 118–135.
Lamontagne, J. R., Stedinger, J. R., Berenbrock, C., Veilleux, A. G., Ferris, J. C., & Knifong, D. L. (2012). Development of regional skews for selected flood durations for the central valley region, california, based on data through water year 2008. *US Geological Survey Scientific Investigations Report*, *5130*, 60.
Lussana, C., Tveito, O. E., Dobler, A., & Tunheim, K. (2019). senorge_2018, daily precipitation, and temperature datasets over norway. *Earth System Science 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 Resources Research*, *36*(3), 737–744.
Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article "assessment of economic flood damage". *Natural Hazards and Earth System Sciences*, *10*(8), 1697–1724.
Midtømme, G. H. (2011). *Retningslinjer for flomberegninger 2011* (Tech. Rep. No. 4/2011).

Nathan, R., & Weinmann, E. (2019). *Estimation of very rare to extreme floods, book 8 in australian rainfall and runoff - a guide to flood estimation.* Commonwealth of Australia.

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. *Hydrological Sciences Journal*, *60*(6), 997-1012. doi: 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. *Journal of Water and Land Development*, *36*(1), 129–141. doi: 10.2478/jwld-2018-0013

Richardson, S., & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), 731–792.

Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods.* New York, NY: Springer.

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

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

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

Sherwood, J. M. (1994). Estimation of volume-duration-frequency relations of ungaged small urban streams in ohio 1. *JAWRA Journal of the American Water Resources Association*, *30*(2), 261–269.

Singh, V. P., & Zhang, L. (2007). Idf curves using the frank archimedean copula. *Journal of hydrologic engineering*, *12*(6), 651–662.

Thorarinsdottir, T. L., Gneiting, T., & Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522-534. doi: 10.1137/130907550

Tyralis, H., Force, H. A., & Langousis, A. (2019). Estimation of intensity-duration-frequency curves using max-stable processes. (January). doi: 10.1007/s00477-018-1577-2

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

Viglione, A., Chirico, G. B., Komma, J., Woods, R., Borga, M., & Blöschl, G. (2010). Quantifying space-time dynamics of flood event types. *Journal of Hydrology*, *394*(1-2), 213–229.

Vinnarasi, R., & Dhanya, C. (2019). Bringing realism into a dynamic copula-based non-stationary intensity-duration model. *Advances in Water Resources*, *130*, 325–338.

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; Tech. Rep.). Norges vassdrags -og energidirektorat.

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

Zhang, L., & Singh, V. (2006). Bivariate flood frequency analysis using the copula method. *Journal of hydrologic engineering*, *11*(2), 150–164.

Appendix A Shape parameter values for QDF and reference models

Table A1: Posterior mean shape parameter values with 90% credible intervals for QDF model fit on durations (24, 36, 48, 60 hours) and posterior mean shape parameter values for individually fit GEV distributions. Stations are in order of catchment area.

Station	Individually fit GEV						QDF					
	Duration (hours)						Model Type					
	1	12	24	36	48	60	DD		RJ		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 A2: Posterior mean shape parameter values with 90% credible intervals for QDF model fit on durations (1, 24, 48, 72 hours) and posterior mean shape parameter values for individually fit GEV distributions. Stations are in order of catchment area.

Station	Individually fit GEV				QDF					
	Duration (hours)				Model Type					
	1	24	48	72	DD		RJ		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]
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]

976 ## Appendix B β parameter values for reference models

Table B1: Posterior mean beta parameter values for individually fit GEV distributions. Stations are in order of catchment area.

Station	Individually fit GEV						
	Duration (hours)						
	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

977 ## Appendix C In-sample return level plots

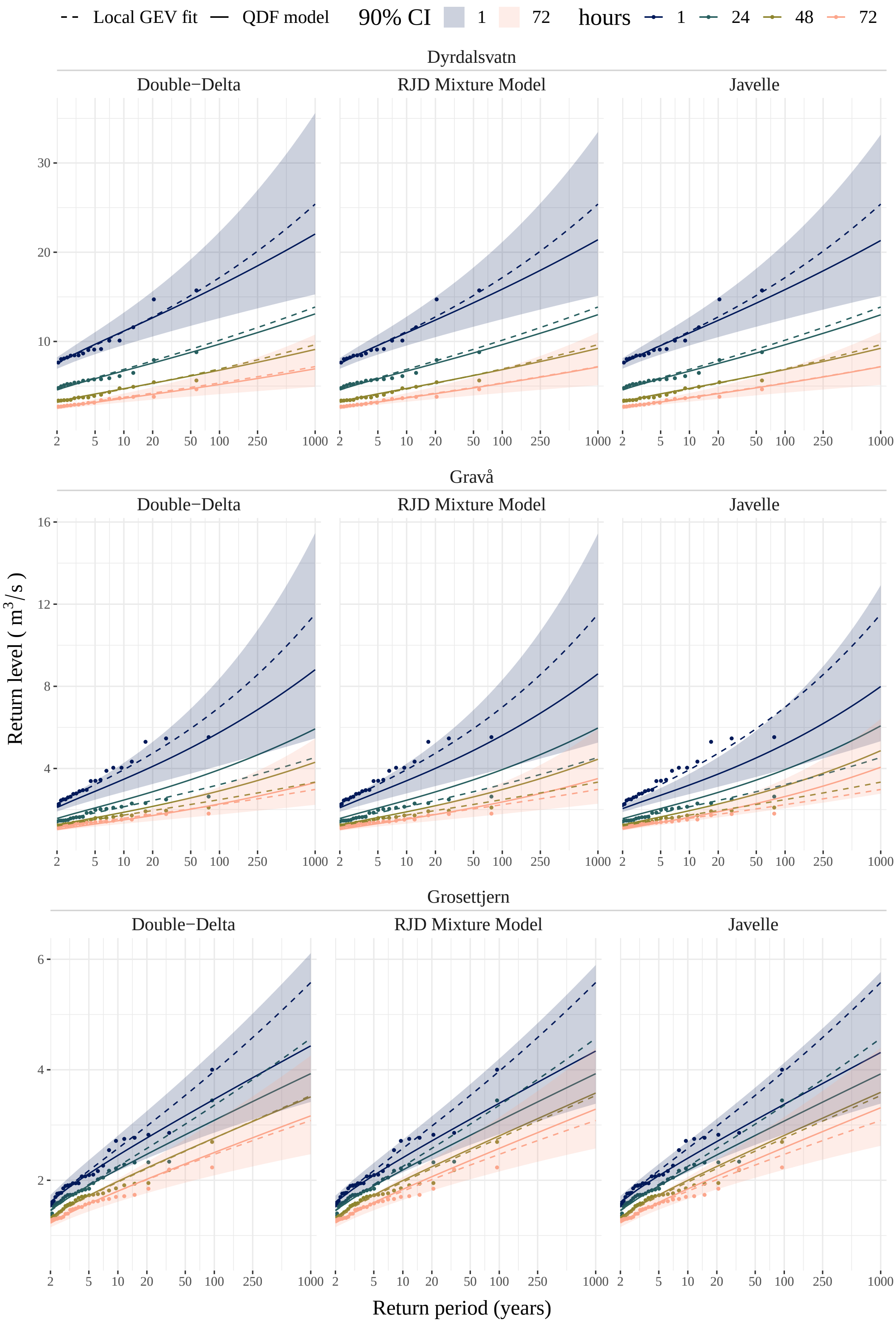
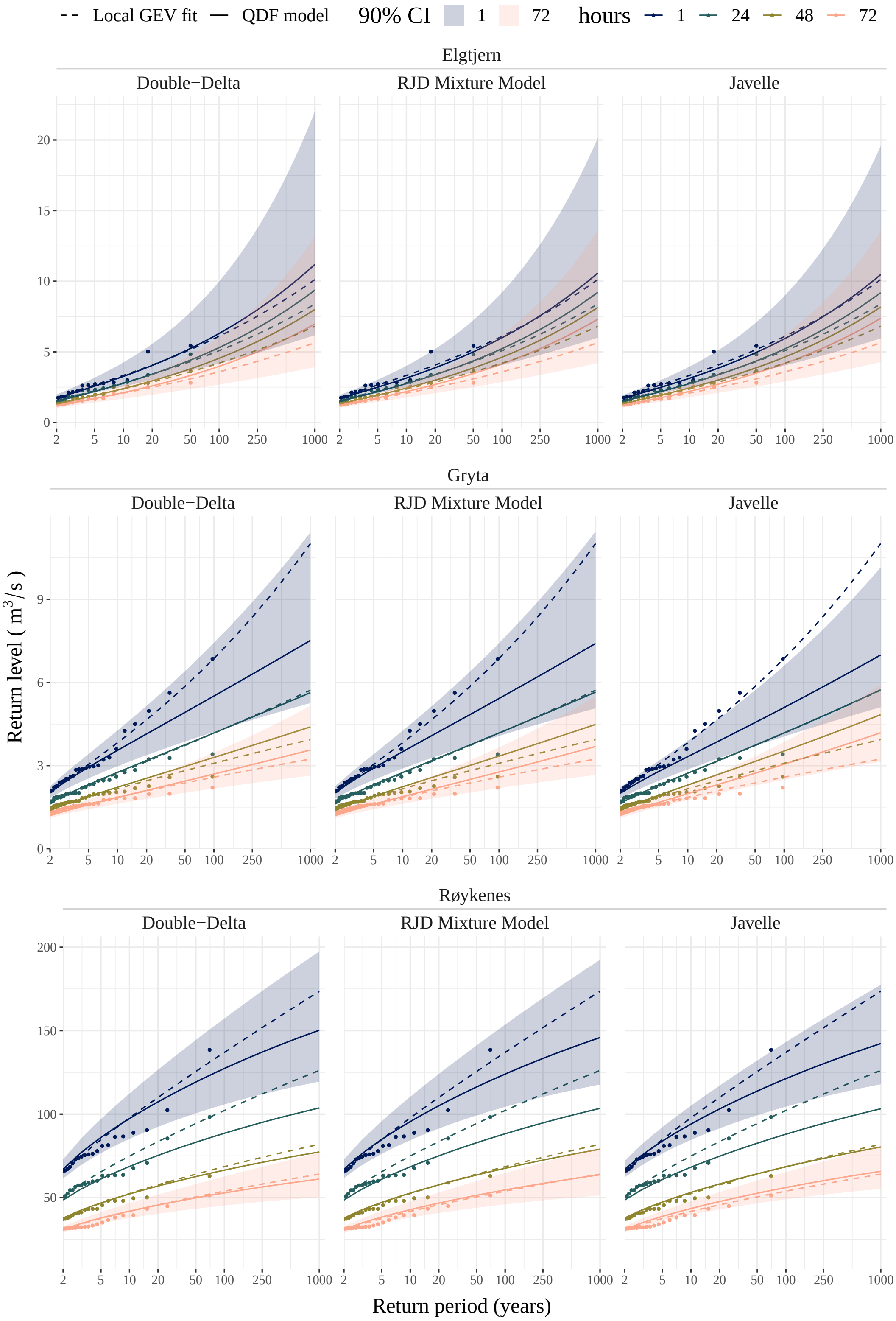
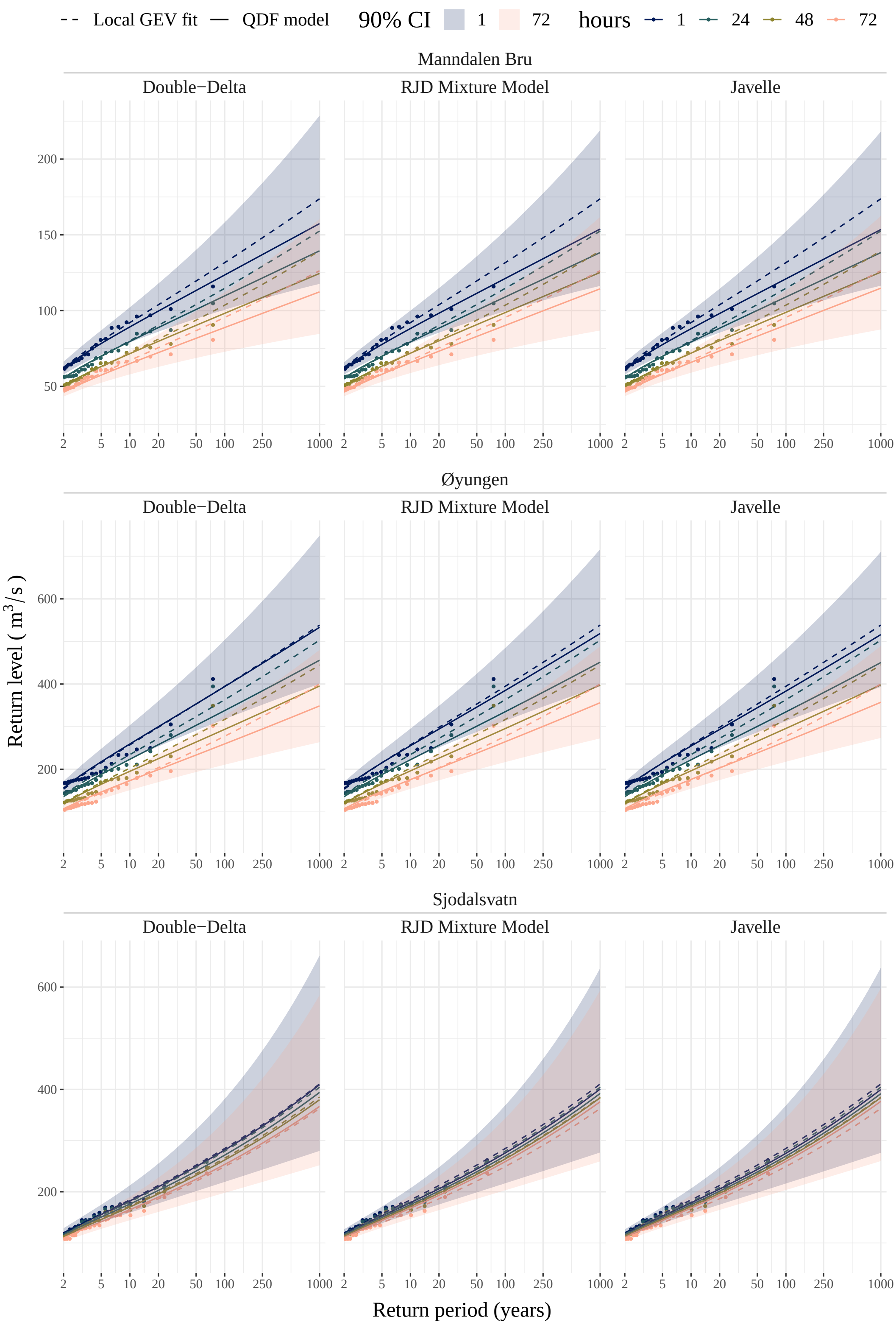


Figure C1



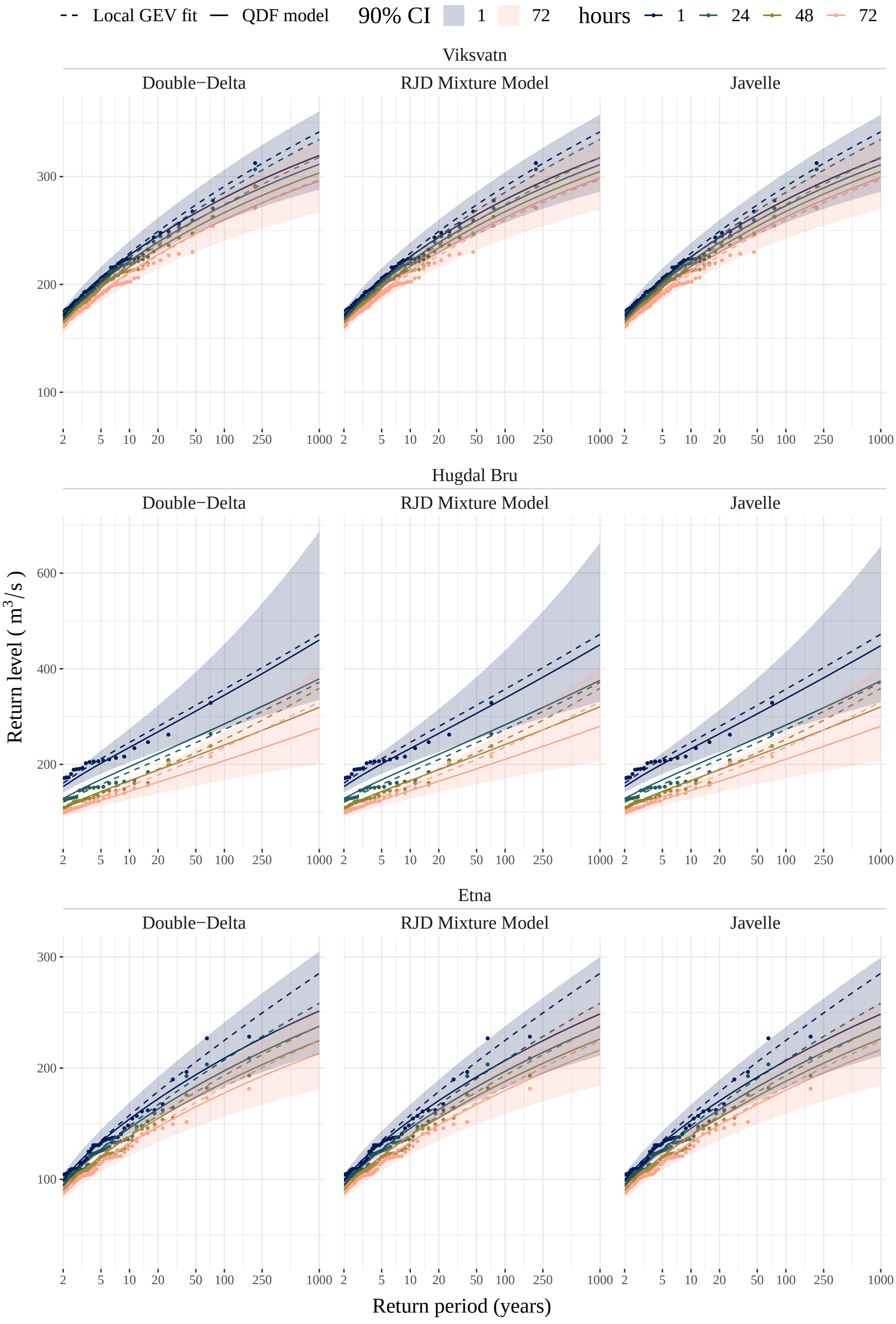
Local GEV fit
QDF model
90% CI
1
72
hours
1
24
48
72
Elgtjern
Gryta
Røykenes
Double−Delta
RJD Mixture Model
Javelle
Return level (m3/s)
Return period (years)

Figure C2



Local GEV fit
QDF model
90% CI
1
72
hours
1
24
48
72
Manndalen Bru
Øyungen
Sjodalsvatn
Double−Delta
RJD Mixture Model
Javelle
Return level (m³/s)
Return period (years)
2
5
10
20
50
100
250
1000
50
100
150
200
400
600

Figure C3



Local GEV fit
QDF model
90% CI
1
72
hours
1
24
48
72
Viksvatn
Hugdal Bru
Etna
Double−Delta
RJD Mixture Model
Javelle
Return level (m3/s)
Return period (years)

Figure C4

978 ## Appendix D Out-of-sample return level plots

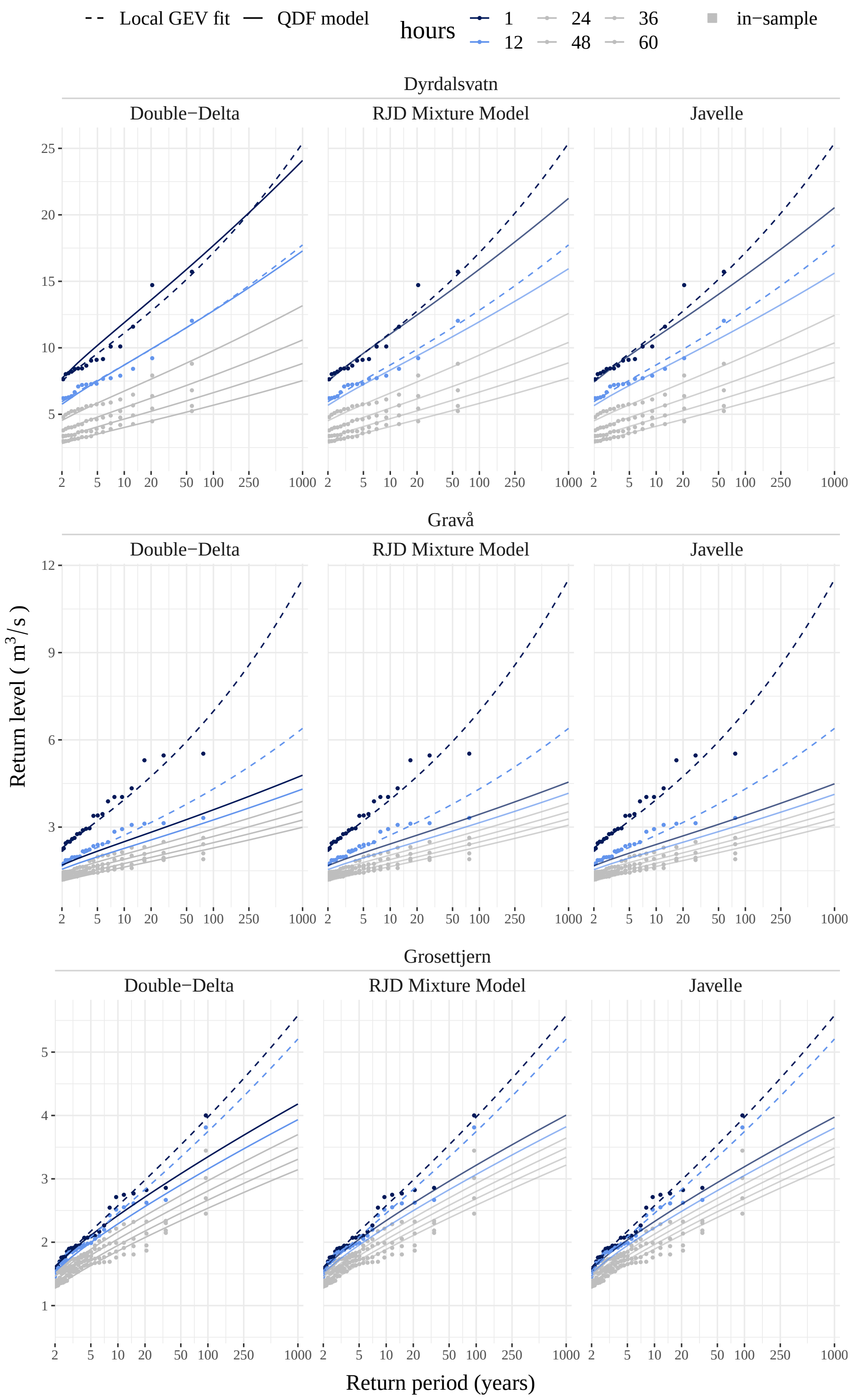


Figure D1

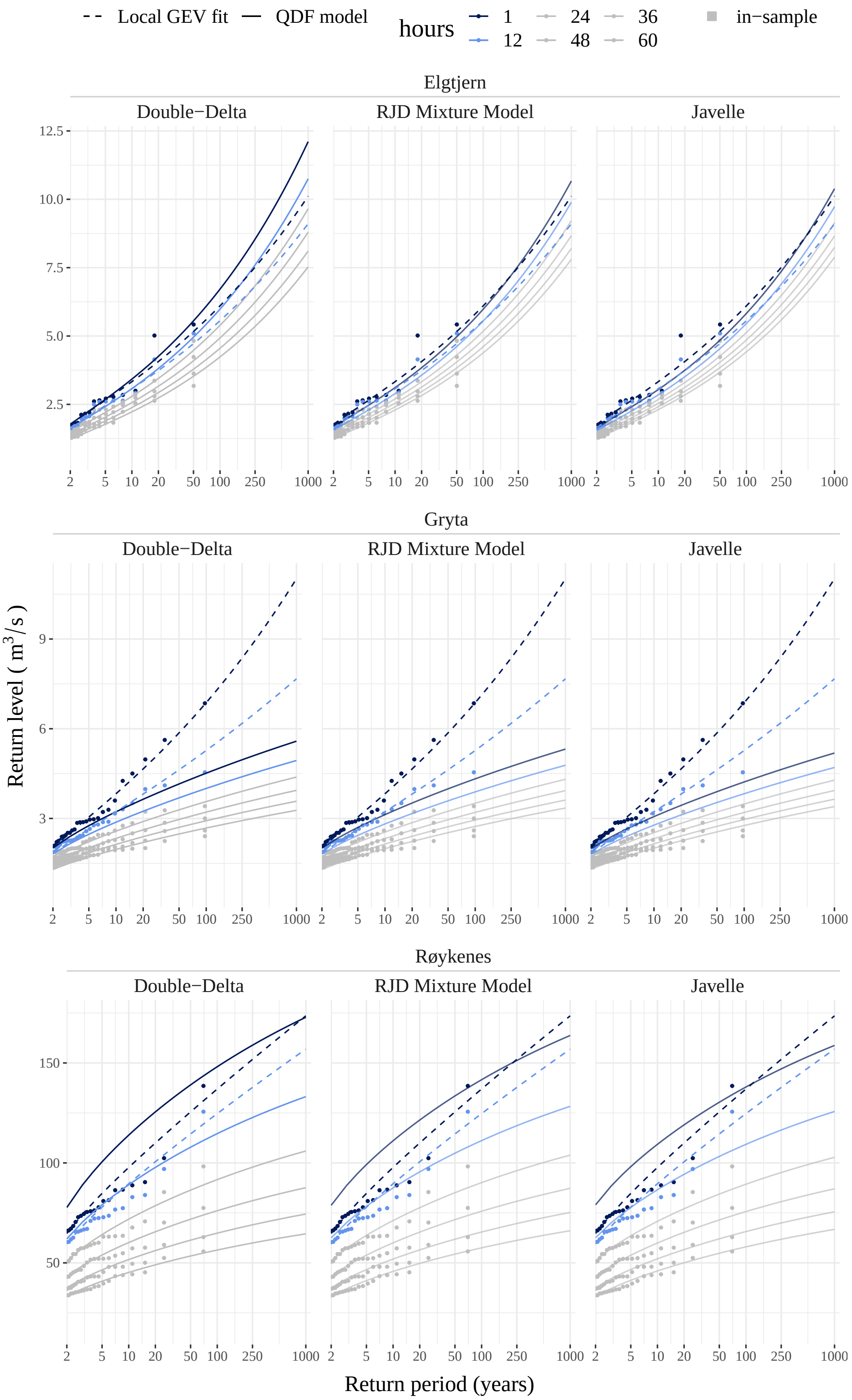
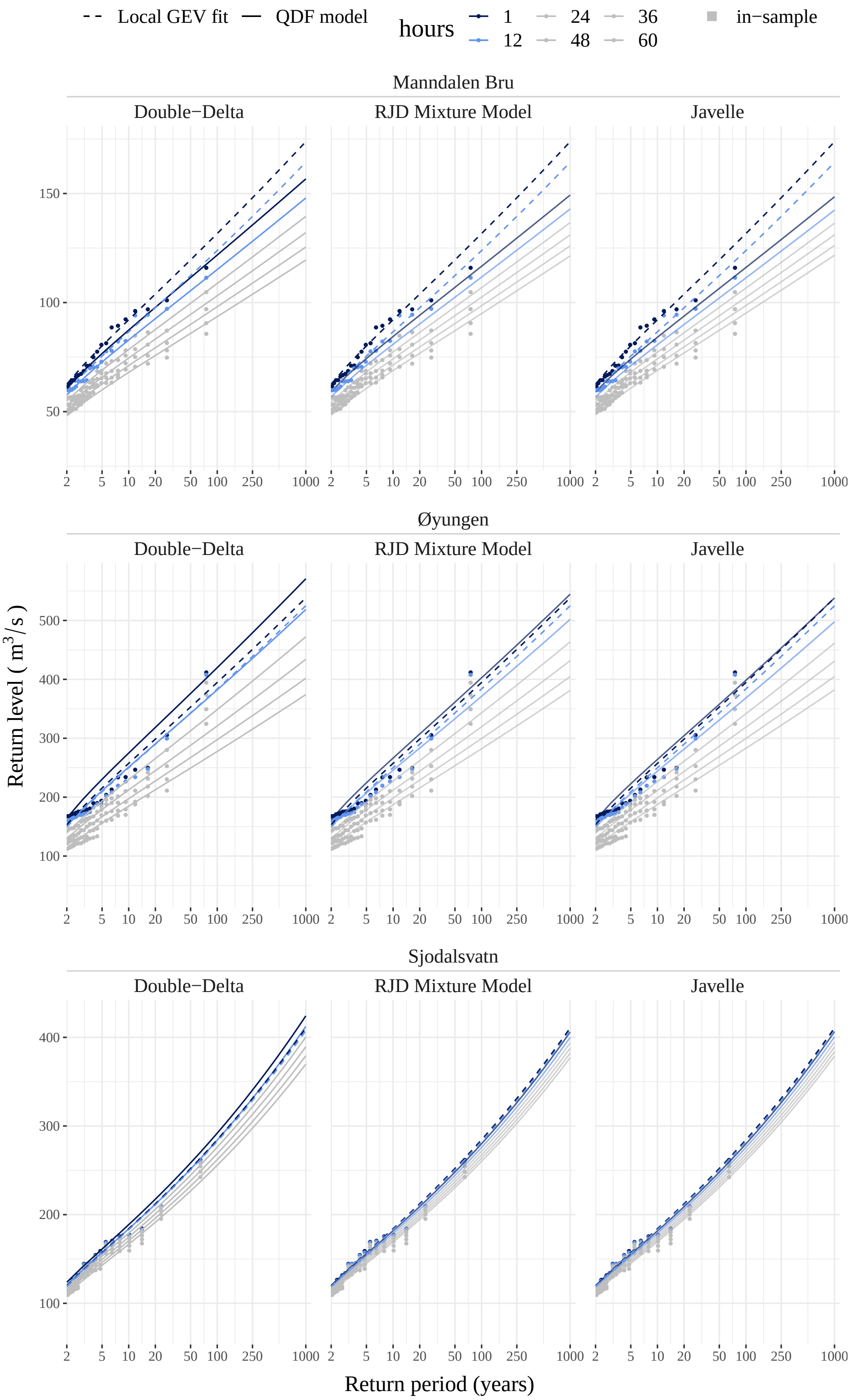
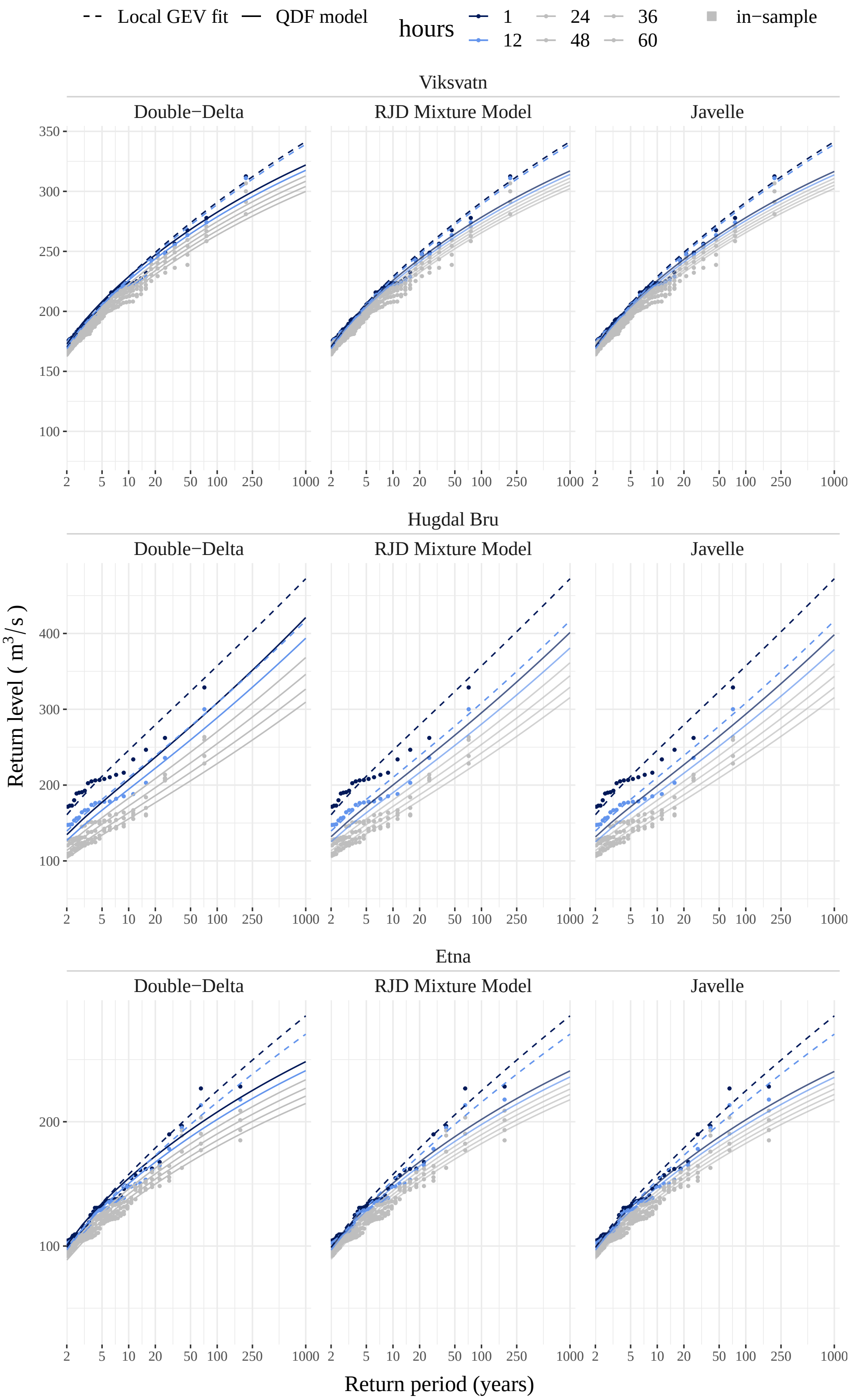

Local GEV fit
QDF model
hours
1
12
24
48
36
60
in−sample
Elgtjern
Double−Delta
RJD Mixture Model
Javelle
Gryta
Røykenes
Return level (m³/s)
Return period (years)
12.5
10.0
7.5
5.0
2.5
9
6
3
150
100
50
2
5
10
20
50
100
250
1000

Figure D2



Local GEV fit
QDF model
hours
1
12
24
48
36
60
in−sample
Manndalen Bru
Øyungen
Sjodalsvatn
Double−Delta
RJD Mixture Model
Javelle
Return level (m3/s)
Return period (years)

Figure D3



Local GEV fit
QDF model
hours
1
12
24
48
36
60
in−sample
Viksvatn
Hugdal Bru
Etna
Double−Delta
RJD Mixture Model
Javelle
Return level (m³/s)
Return period (years)

Figure D4