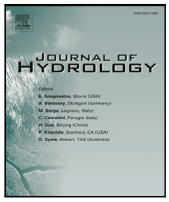
Journal of Hydrology 620 (2023) 129448

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).

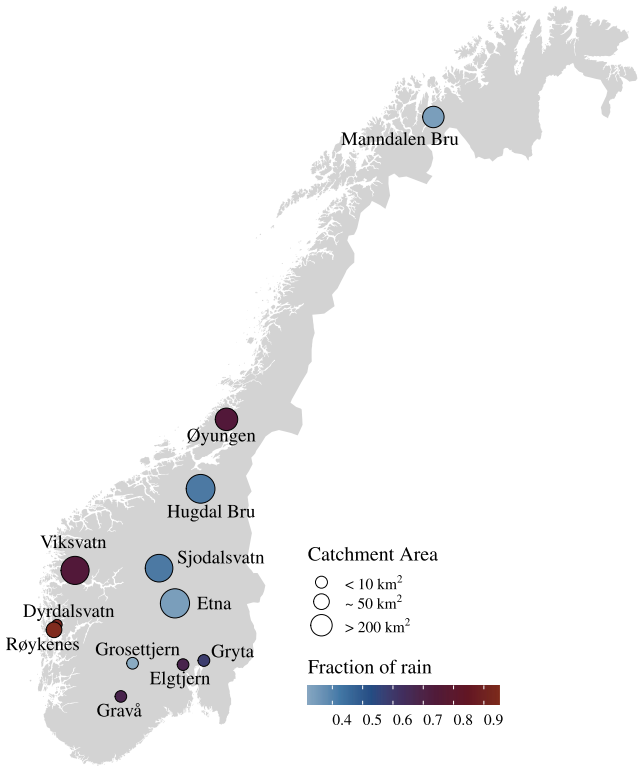


Fig. 1. Locations of twelve gauging stations used in study. Catchment area and fraction of rain contribution to flood are also indicated.

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

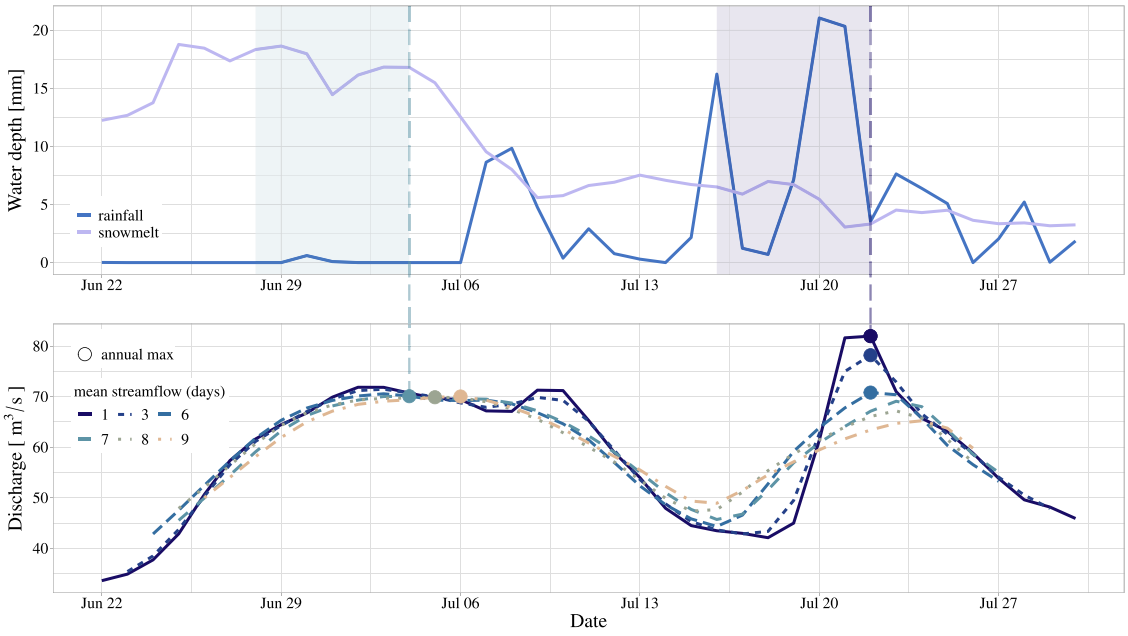


Fig. 2. In QDF modeling, the duration d represents the total flow volume for a time span of d hours, not flood events that lasted precisely d hours. This means longer duration series are always aggregated from series of shorter durations. This creates a dependency structure that is artificial yet not 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

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

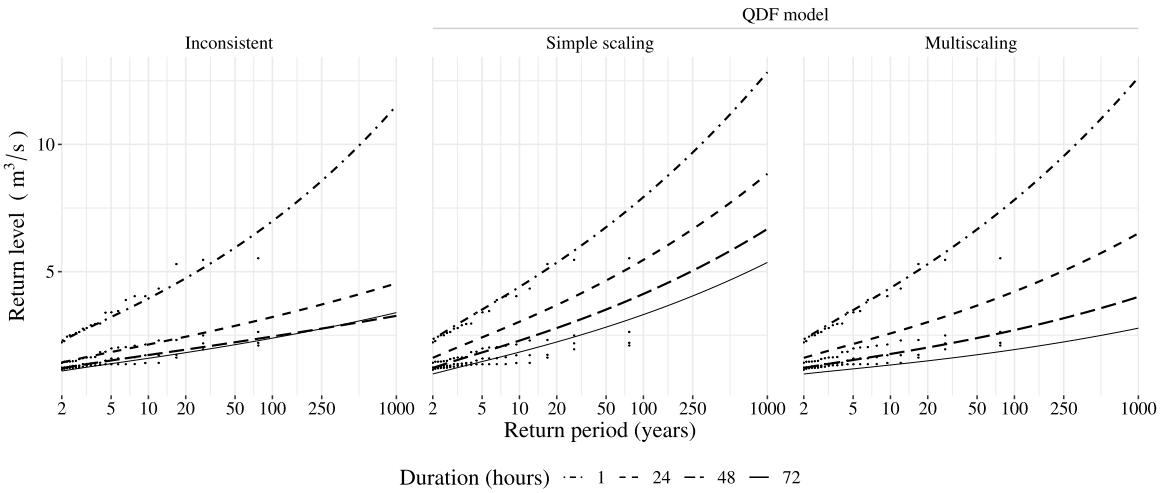


Fig. 3. Return level plots from a synthetic data set showing (i) flood frequency curves estimated independently for four durations (left panel), (ii) output from a simple scaling QDF model (middle panel), and (iii) output from a multiscaling QDF model (right panel). The independent fits do not account 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

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

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. 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 $\boldsymbol{\theta}_1$ 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 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 $\boldsymbol{\theta}$. 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 = 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

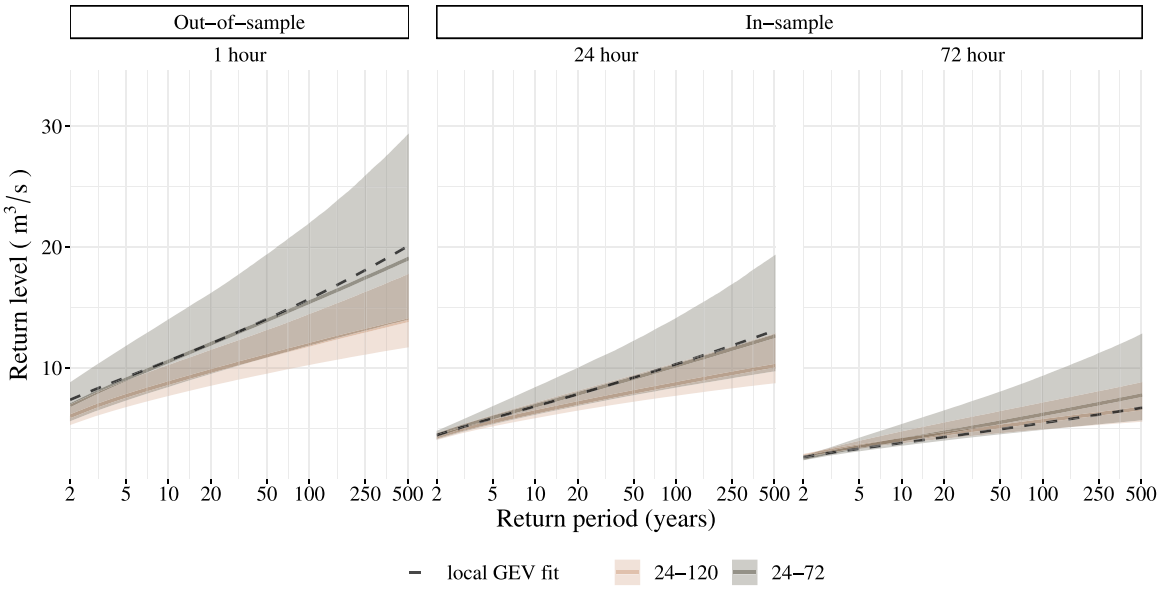


Fig. 4. Return level plots from the Dyrdalsvatn gauging station using the Double-Delta model fit to two different data sets: one set with six durations [24, 36, 48, 72, 96, 120 h] and one set with four durations [24, 36, 48, 72 h]. The model fit to the six duration set is both overconfident and biased 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.

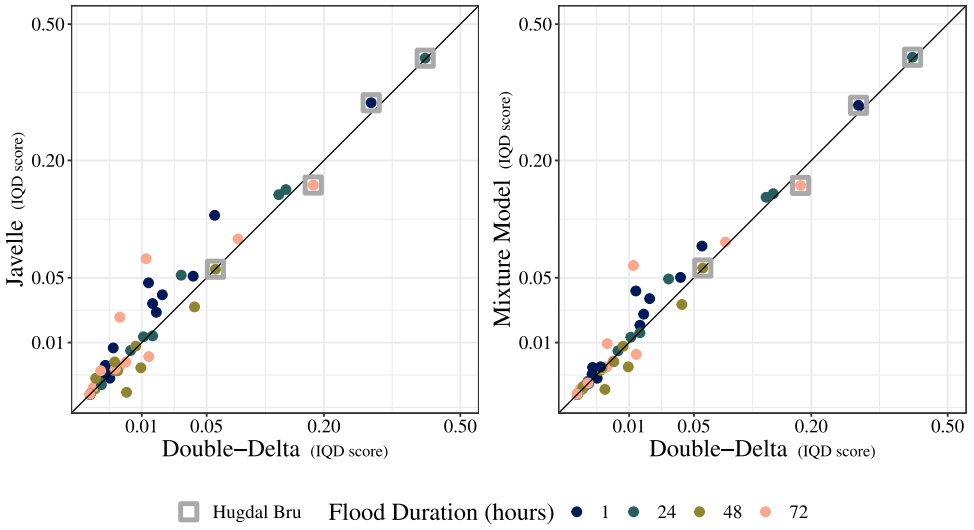


Fig. 5. Model-to-model comparison of interquantile distance (IQD) scores for each station and in-sample duration. Lower values of the IQD indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, left panel) and the mixture model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

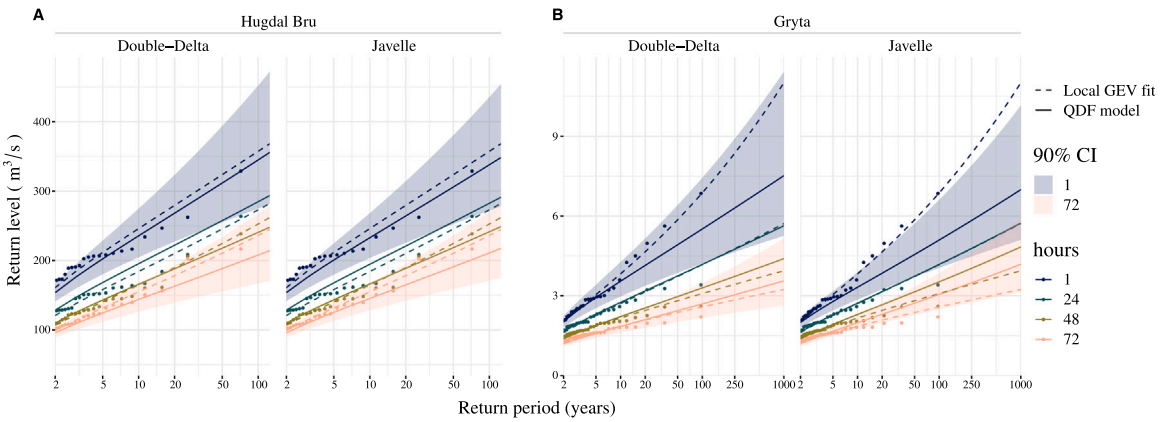


Fig. 6. Return level plots showing two selected stations where QDF models differ substantially from the reference model on in-sample durations. (**A**) Hugdal Bru: the 1 h floods with return period under 5 years are characterized by a diurnal melt-freeze cycle at this snowmelt-driven catchment; 1 h floods 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.

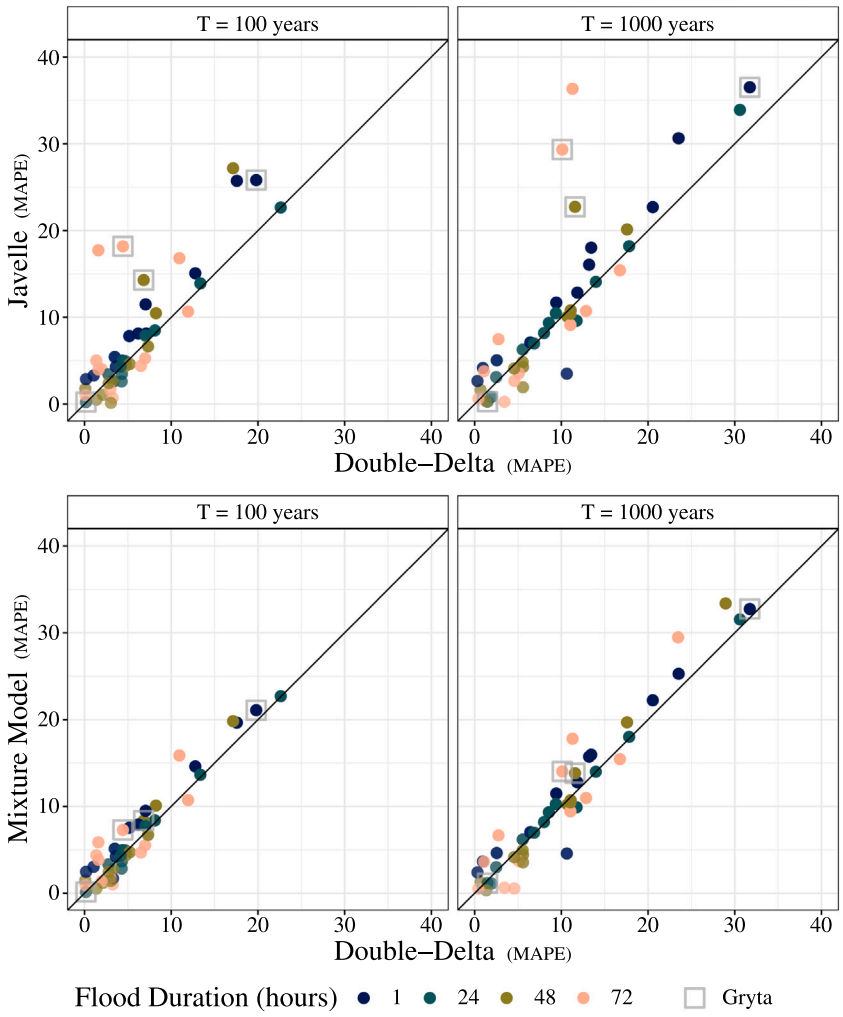


Fig. 7. Model-to-model comparison of the mean absolute percent error (MAPE) scores for each station and in-sample duration. Lower values of the MAPE indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, top panels) and the mixture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

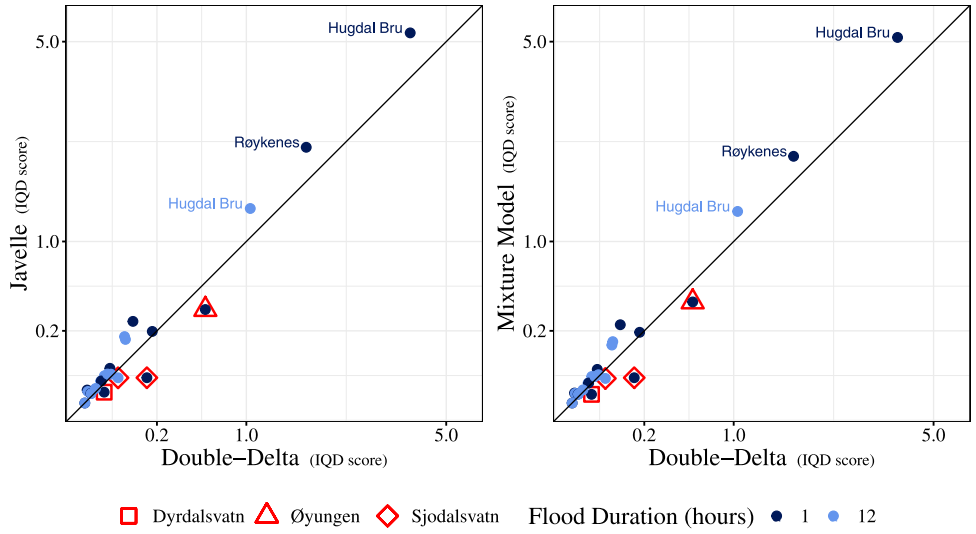


Fig. 8. Model-to-model comparison of interquantile distance (IQD) scores for each station and both out-of-sample durations. Lower values of the IQD indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, left panel) and the mixture 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.)

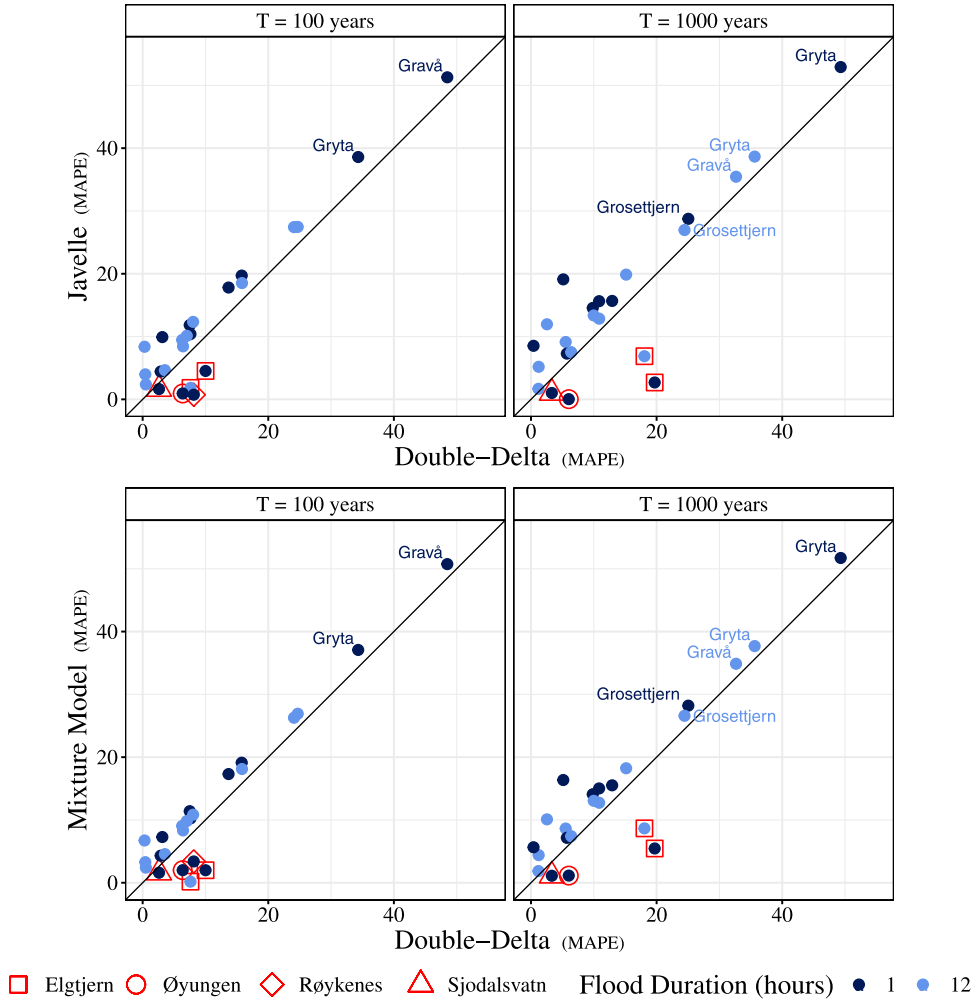


Fig. 9. Model-to-model comparison of mean absolute percent error (MAPE) scores for each station and both out-of-sample durations. Lower values of the MAPE indicate better performance. The extended QDF model (Double-Delta) serves as a reference to both the original QDF model (Javelle, top panels) and 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.)

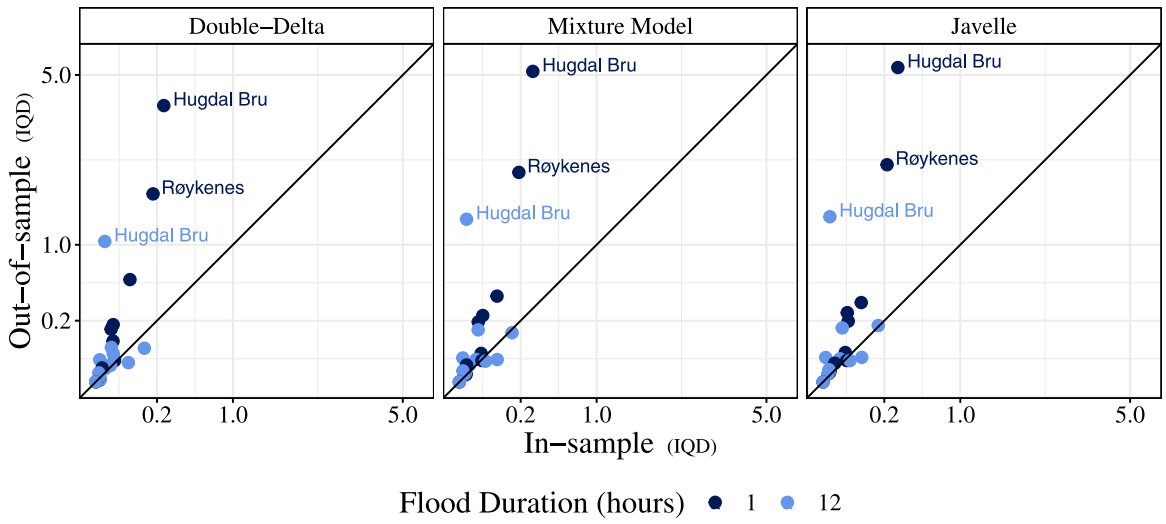


Fig. 10. Comparison of interquantile distance (IQD) score when durations are either predicted (out of sample) or included in the model fitting set (in sample). Lower values of the IQD indicate better performance. The out-of-sample set was fit with durations 24, 36, 48, 60 h and used to predict the 1 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

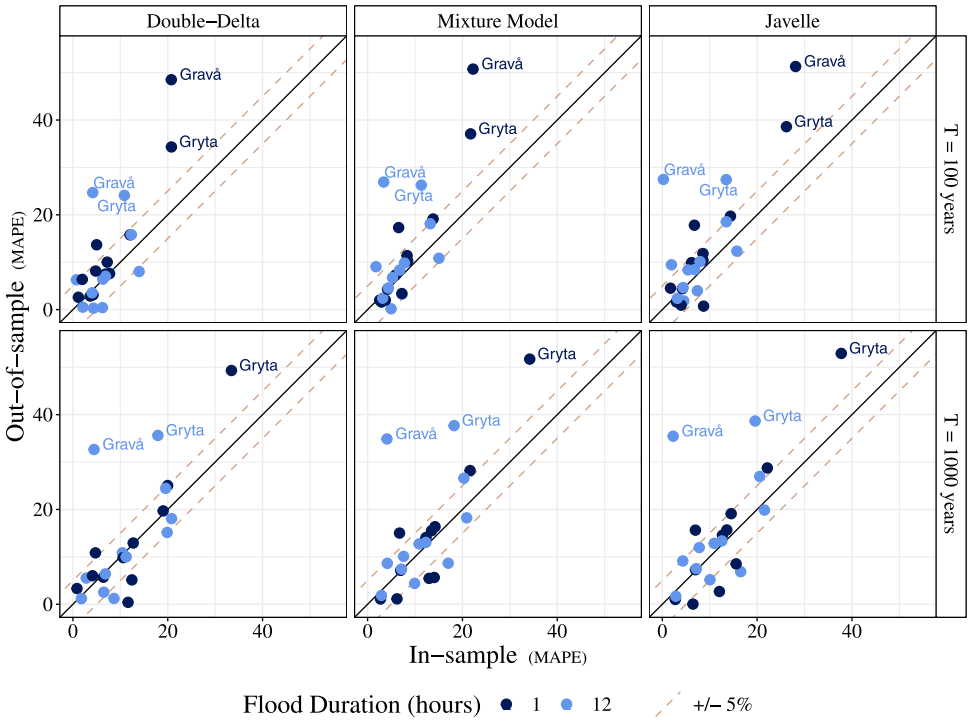


Fig. 11. Comparison of mean absolute percent error (MAPE) when durations are either predicted (out of sample) or included in the model fitting set (in sample). Lower values of the MAPE indicate better performance. The out-of-sample set was fit with durations 24, 36, 48, 60 h and used to predict the 1 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

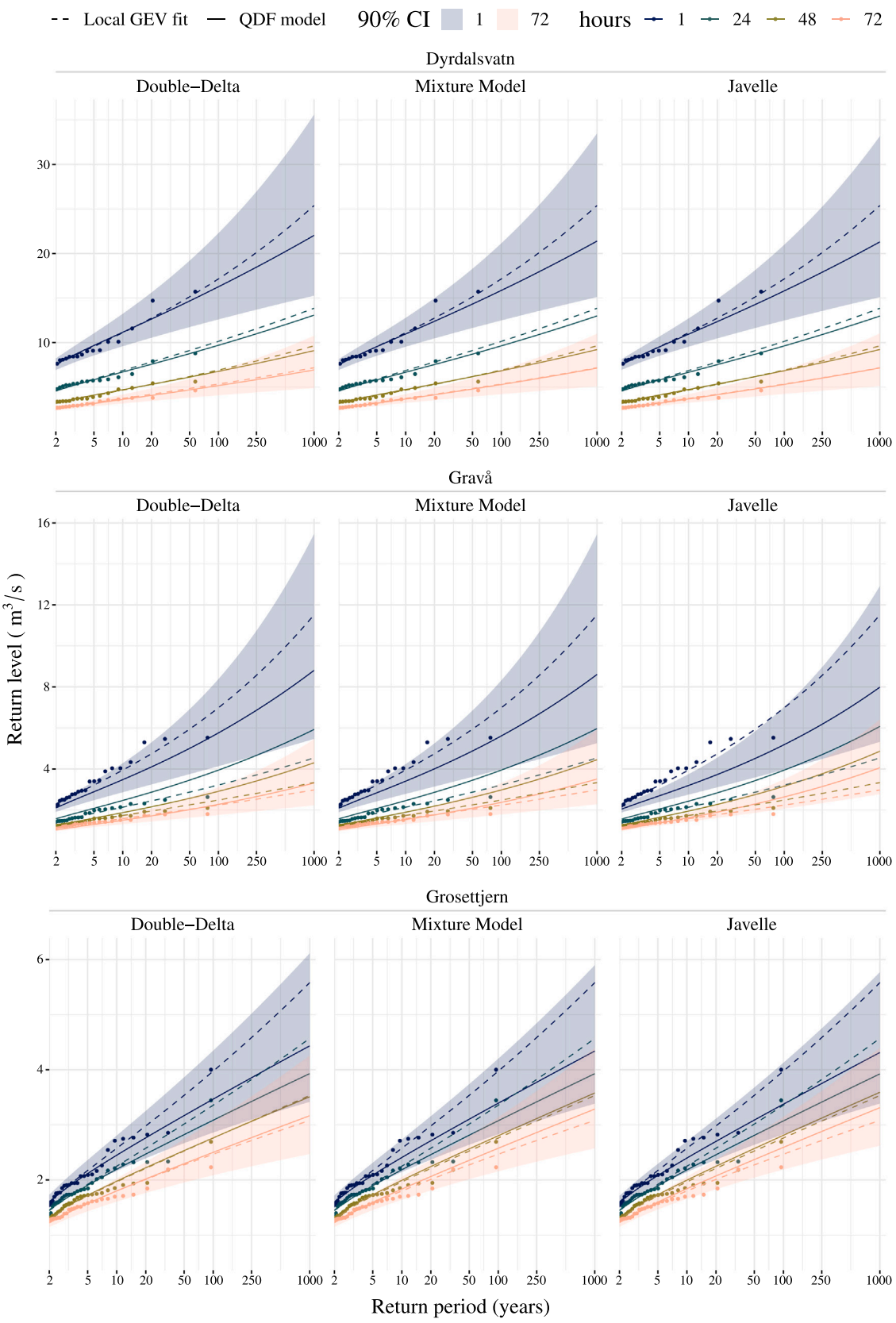


Fig. E.12. In-sample return level plots for stations Dyrdalsvatn, Gravå, and Grosettjern.

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.

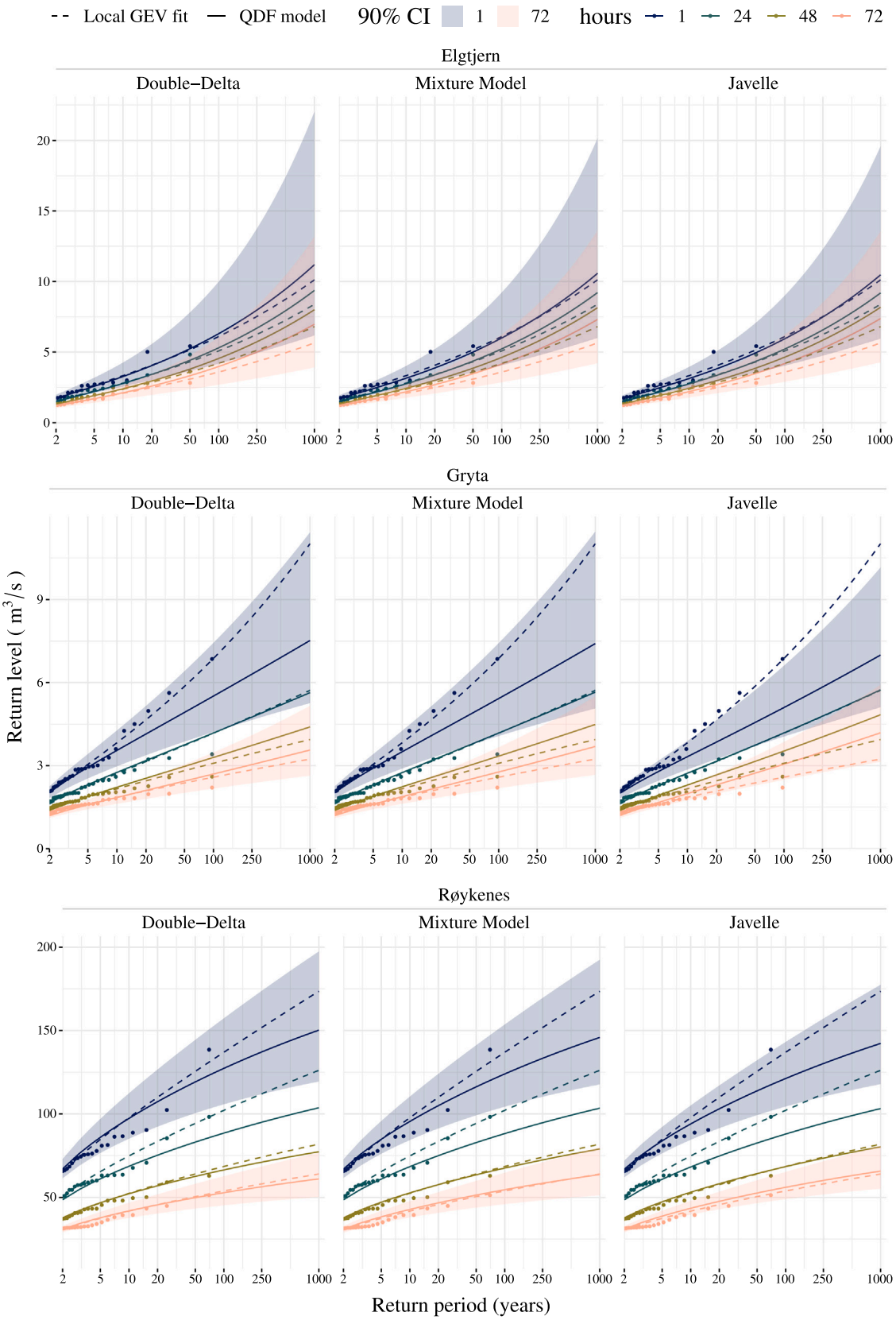


Fig. E.13. In-sample return level plots for stations Elgtjern, Gryta, and Røykenes.

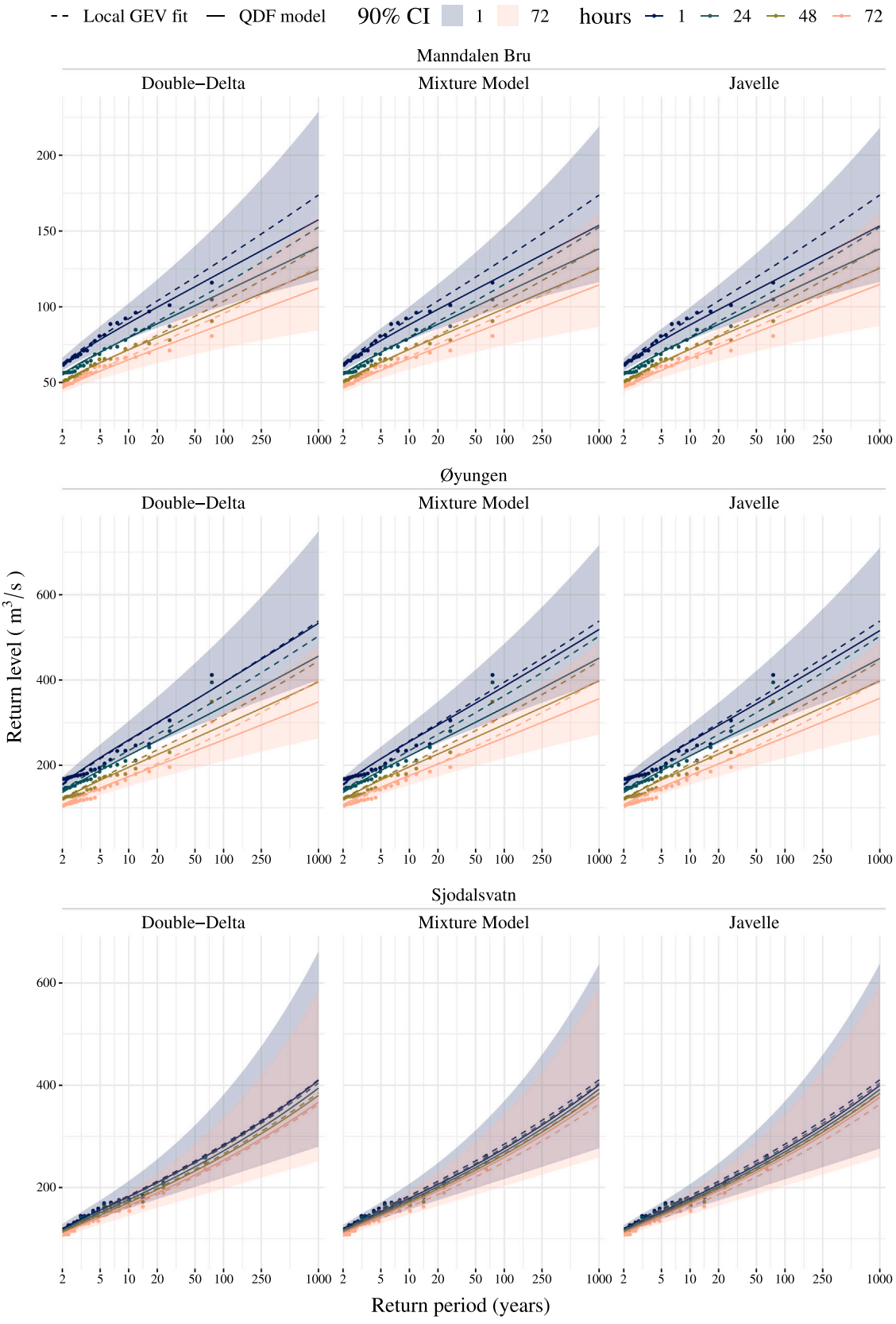


Fig. E.14. In-sample return level plots for stations Manndalen Bru, Øyungen, and Sjodalsvatn.

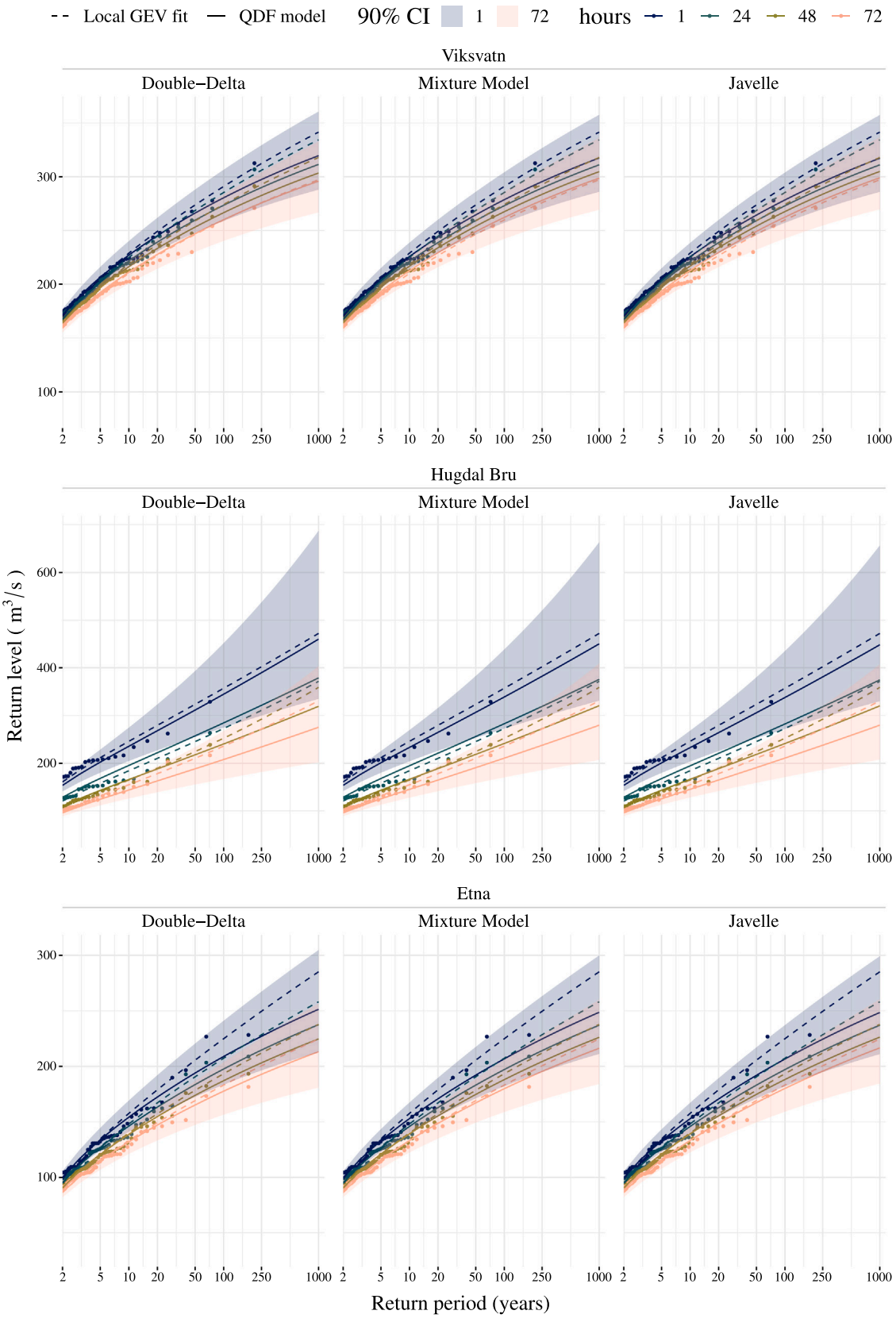


Fig. E.15. In-sample return level plots for stations Viskvatn, Hugdal Bru, and Etna.

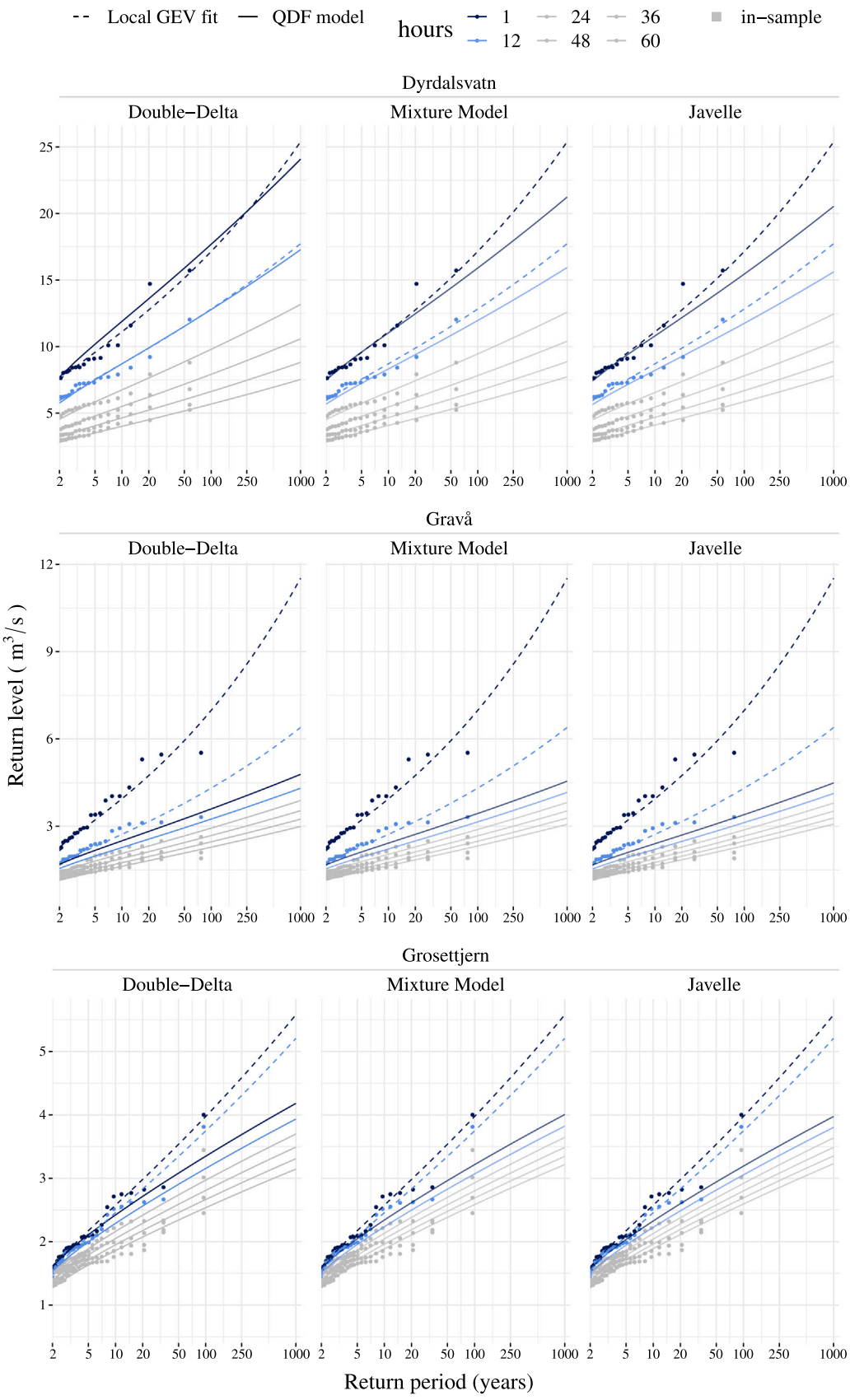


Fig. F.16. Out-of-sample return level plots for stations Dyrdalsvatn, Gravå, and Grosettjern.

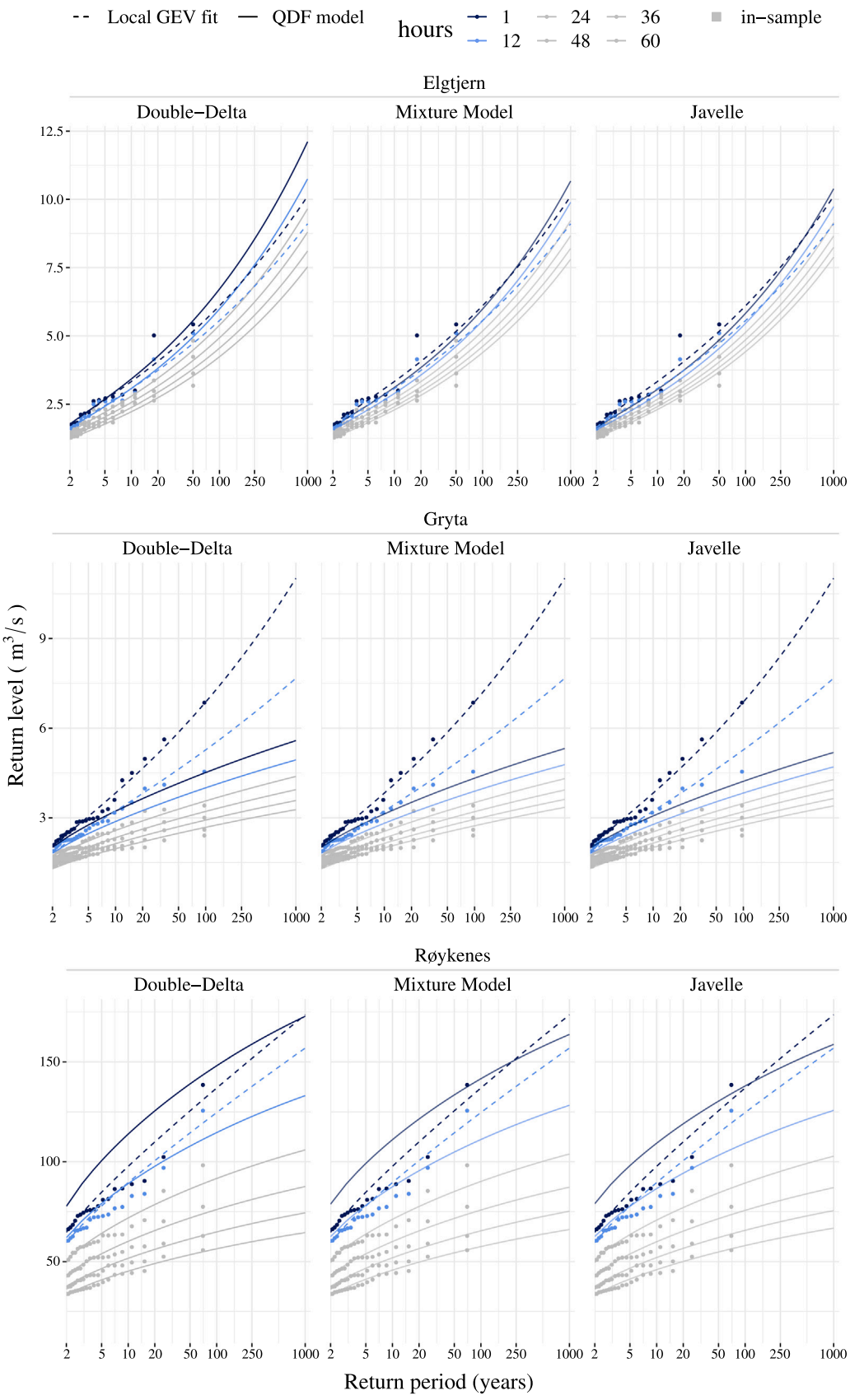


Fig. F.17. Out-of-sample return level plots for stations Elgtjern, Gryta, and Røykenes.

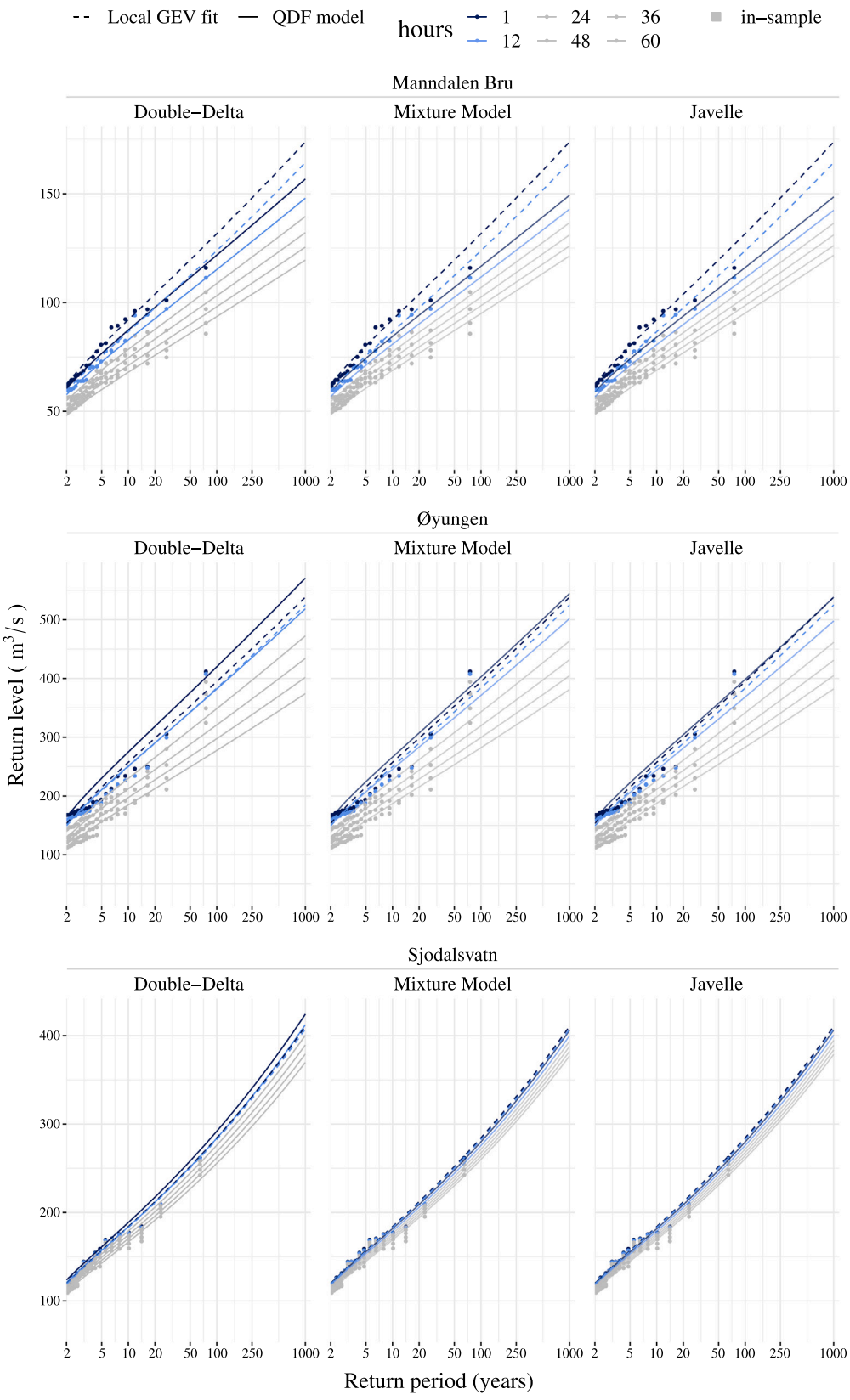


Fig. F.18. Out-of-sample return level plots for stations Manndalen Bru, Øyungen, and Sjodalsvatn.

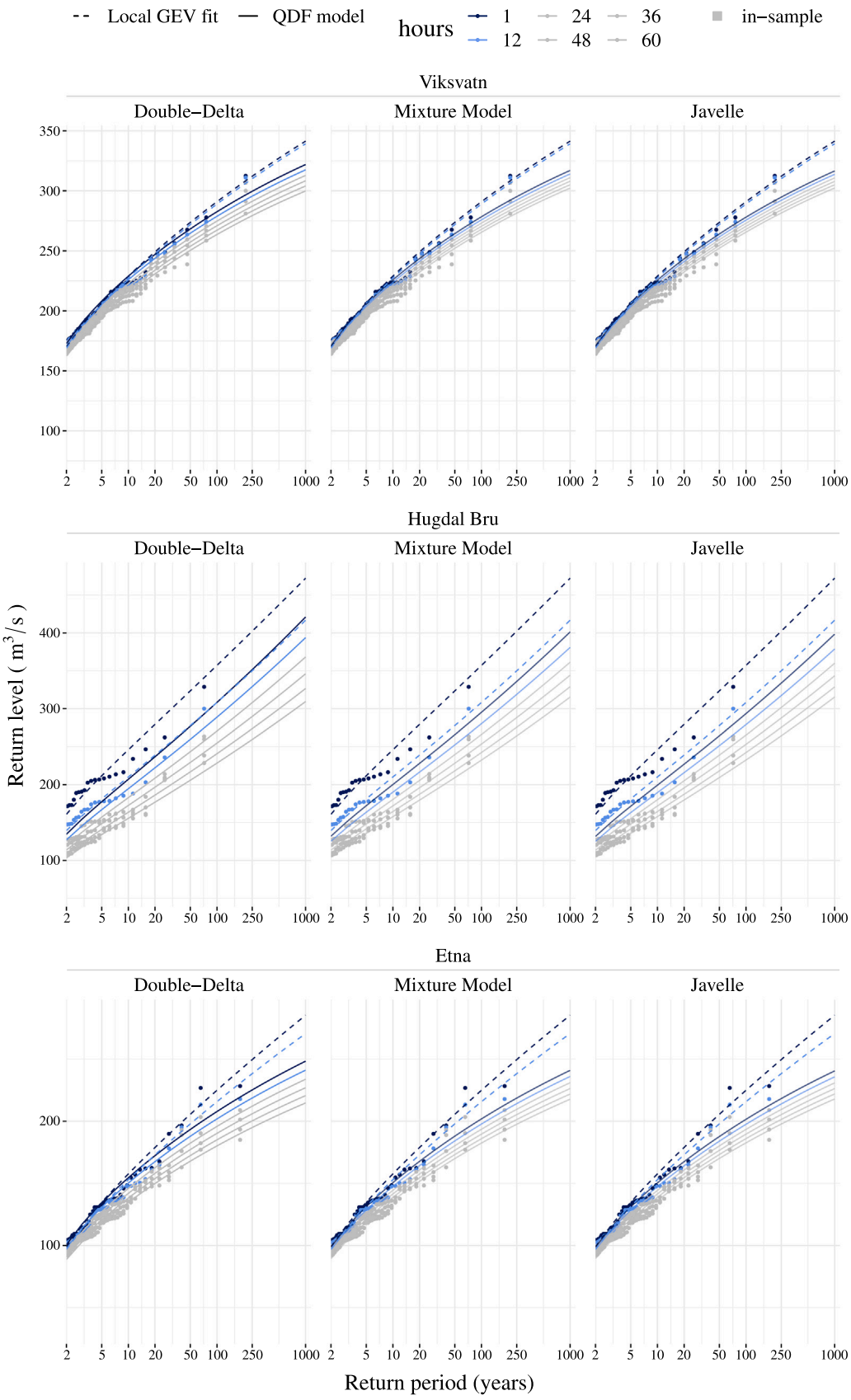


Fig. F.19. Out-of-sample return level plots for stations Viksvatn, Hugdal Bru, and Etna.

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 flomfrekvensanalyse 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.