Use of historical data in flood frequency analysis: a case study for four catchments in Norway
Kolbjørn Engeland, Donna Wilson, Péter Borsányi, Lars Roald and Erik Holmqvist

ABSTRACT

There is a need to estimate design floods for areal planning and the design of important infrastructure. A major challenge is the mismatch between the length of the flood records and needed return periods. A majority of flood time series are shorter than 50 years, and the required return periods might be 200, 500, or 1,000 years. Consequently, the estimation uncertainty is large. In this paper, we investigated how the use of historical information might improve design flood estimation. We used annual maximum data from four selected Norwegian catchments, and historical flood information to provide an indication of water levels for the largest floods in the last two to three hundred years. We assessed the added value of using historical information and demonstrated that both reliability and stability improves, especially for short record lengths and long return periods. In this study, we used information on water levels, which showed the stability of river profiles to be a major challenge.

Key words | Bayesian estimation, flood frequency analysis, historical floods, Norway

INTRODUCTION

The motivation of this study is the need to estimate design floods for important infrastructure. According to Norwegian dam safety regulations (Lovdata 2010), dam safety should be evaluated for floods with 500 or 1,000 years return periods, depending on an individual dam safety class. According to building regulations (TEK10 2016), buildings and infrastructure should resist or be protected from floods with 20, 200, or 1,000 year return periods, depending on the consequence of flooding. To support municipalities in land use planning and to meet the obligations of the regulations, the Norwegian Water Resources and Energy Directorate (NVE) produce flood zone maps of floods with 10 to 500 year return periods, for selected locations. Since the typical length of a streamflow record is 40–50 years and the longest streamflow record in Norway is 123 years, the estimated high flood quantiles (i.e., for 200, 500, and 1,000 year return periods) are based on a large degree of extrapolation, and the estimates are highly uncertain. In order to reduce the estimation uncertainty for high flood quantiles, the amount of data might be increased by following three different strategies (Merz & Blöschl 2008; Gaál et al. 2010; Kjeldsen et al. 2014): (i) use flood data from several gauging stations within a region (e.g., Dalrymple 1960; Hosking & Wallis 1997); (ii) use historical data (e.g., Benson 1950; Brázdil et al. 2006; Viglione et al. 2013; Macdonald et al. 2014; Schendel & Thongwichian 2017) and/or paleohydrological data (e.g., Benito & O’Connor 2013); or (iii) use causal information, i.e., by combining precipitation statistics with precipitation-runoff models (e.g., Paquet et al. 2015).

In this study, we focused on the use of historical information. Such information might be included in at least three ways (Reis & Stedinger 2005; Gaál et al. 2010). For a certain period, we need to know either (i) the exact magnitude, (ii) the upper and lower limits, or (iii) the total...
number of all floods over a threshold. Below follows a brief summary of the methods and challenges related to the use of historical data. More comprehensive reviews are found in the literature (Stedinger & Cohn 1986; Brázdil et al. 2006; Kjeldsen et al. 2014).

Parameter estimation based on a combination of historical and systematic data can be classified into graphical, moment-based, maximum likelihood, or Bayesian methods. The graphical methods are based on establishing an empirical distribution via the use of suitable plotting positions, which can include historical data, and then a line is fitted to the empirical data (e.g., Benson 1950; Gerard & Karpuk 1976; Zhang 1982; Bayliss & Reed 2001). Two different moment-based methods that take into account historical flood information, have been developed. The weighted moment technique was first developed for ordinary moments (US Water Resources Council 1982) and extended to probability weighted moments in Wang (1990). For this method we need to know the magnitude of the historical floods, and the moment estimates are based on weighting the historical and systematic data according to the length of the time periods they represent. The expected moment method is developed for ordinary moments (Cohn et al. 1997) and probability weighted moments (Jeon et al. 2011). Also, this method requires that the magnitudes of all historical floods above a threshold are known. The maximum likelihood method combines likelihoods for the systematic data, the historical data above a threshold, and the number of years in the historical period the threshold was not exceeded (Stedinger & Cohn 1986). In a Bayesian framework, the likelihood is defined in a similar way, and combined with prior information, posterior distributions of parameters and return levels are estimated (Reis & Stedinger 2005; Gaál et al. 2010). Most references mentioned above combine annual maximum series (AMS) from the systematic record with the historical information, and AMS-type distributions are used (e.g., generalized extreme value, generalized logistic and the Pearson type III distributions). An alternative is to model both the systematic and the historical data as peaks over threshold, and apply a generalized Pareto distribution (Macdonald & Black 2010). In this study, we combined annual maxima from the systematic data with the historical information, and we used the Bayesian approach since it is flexible with respect to the type of historical information that can be included. The Bayesian approach is also more flexible when, in a later stage, several data sources might be combined in a regional-historical flood frequency analysis.

Sources for historical data include: (i) annals, chronicles, memory books and memoirs; (ii) weather diaries; (iii) correspondence (letters); (iv) special prints; (v) official economic and administrative records; (vi) newspapers and journals; (vii) sources of a religious nature; (viii) chronograms; (ix) early scientific papers, compilations and communications; (x) stall-keepers’ and market songs; (xi) pictorial documentation; and (xii) epigraphic sources (Brázdil et al. 2010, 2012). In several European countries there are central depositories of historical flood data (Brázdil et al. 2006; Kjeldsen et al. 2014). In Norway, a central depository is still under development, and an informative overview is given in Roald (2015). One important source of information for historical floods in Norway is official economic and administrative records provided as public reviews of flood damages for individual farms resulting in tax reduction. A systematic analysis of these data has the potential to reveal both the magnitudes and spatial extent of large historical floods (Roald 2015). The use of historical floods is challenging due to changes in the environment and society. First, historical flood data are based on all floods which exceed a perception threshold for floods, and this perception level might evolve dependent on how vulnerable a society is to flooding (Kjeldsen et al. 2014; Macdonald & Sangster 2017). Time-dependent thresholds can be used to meet this challenge (e.g., Brázdil et al. 2006). Second, flood generating processes might evolve in time due to changes in climate and land use, which can make the flood frequencies themselves non-stationary (Benito et al. 2004). Finally, changes in the river channel might limit the possibility to estimate the magnitude of historical floods (Brázdil et al. 2011). Several studies demonstrate that provided the perception threshold is sufficiently high, it is sufficient to know the number of floods exceeding this threshold in order to improve flood quantile estimates (Stedinger & Cohn 1986; Martins & Stedinger 2001; Payrastre et al. 2011).

When using historical data, it is important to set a reasonable length of the period for which we assume that all events above a threshold are known. Strupczewski et al. (2014) assess the sensitivity of flood quantile estimates to the
length of the historical period and show that if one large historical flood event is known, the optimal choice is to set the start of the historical period so that this event is in the middle of the historical period. In Prosdocimi (2017), it is shown that this estimate equals to the ‘maximum spacing estimator’ which sets the start of the historical period to be the time point that precedes the first historical flood event by the average time spacing between the historical events. The same approach is used in Schendel & Thongwichian (2017), the only difference being that both the historical and the systematic data are used to calculate the average time spacing between flood events. The aims of this paper are to (i) assess the added value of using historical data for flood quantile estimation, how the added value depends on data availability and the estimated length of the historical period, and (ii) to demonstrate the use of historical information in selected Norwegian catchments. To address this first aim, a 123-year-long annual maximum flood series was used, and it was evaluated how the added value of historical information depends on (a) the systematic record length relative to the length of the historical period, (b) threshold for historical events, and (c) the estimated length of the historical period. To address the second aim, four catchments were selected where epigraphic sources indicating the water levels of large historical floods close to gauging stations are available.

Guidelines and/or recognized approaches for the inclusion of historical information in flood frequency analysis have not yet been developed for Norway (Kjeldsen et al. 2014), and as far as the authors know, no scientific studies have addressed this topic for Norwegian catchments.

This paper continues with a presentation of the study area, and data include a brief description of the largest historical floods. The methodology used for parameter estimation and the evaluation strategy are detailed, before the results are presented and discussed. Finally, some conclusions are drawn.

**STUDY AREA AND DATA**

**Study catchments**

The location of the four selected study catchments: Bulken, Lalm, Strandefjord, and Labru, and their upstream catchment areas are shown in Figure 1. Important catchment characteristics are listed in Table 1, and Figure 2 illustrates the seasonality of both monthly streamflows and floods, and shows the time series of annual maximum floods. Bulken is located in western Norway (Figure 1). The largest floods in this region are generated from heavy precipitation, typically in late autumn/winter, caused by westerly winds combined with high atmospheric moisture transport known as atmospheric rivers, that meet the mountain range at the Norwegian west coast (e.g., Stohl et al. 2008). Lalm, Strandefjord, and Labru are all located in the geographic region of eastern Norway (Figure 1). For these catchments, the largest floods are caused by a combination of heavy precipitation and snow melt. The largest precipitation amounts in this region are generated by south easterly winds that transport high moisture from the Mediterranean which streams around the eastern part of the Alps (Roald 2008). This storm trajectory is known as the type Vb in the van Bebber classification (Van Bebber 1882; Roald 2008). The Vb type is also linked to the largest floods in eastern and central Europe (Roald 2008; Brázdil et al. 2012).
Table 1 | Characteristics and flood statistics for the four study catchments

<table>
<thead>
<tr>
<th>Name</th>
<th>Bulken</th>
<th>Lalm</th>
<th>Strøndefjord</th>
<th>Labru</th>
</tr>
</thead>
<tbody>
<tr>
<td>St. number</td>
<td>62.5</td>
<td>2.25</td>
<td>12.80</td>
<td>15.1</td>
</tr>
<tr>
<td>Catch. area (km²)</td>
<td>1,092</td>
<td>2,982</td>
<td>1,793</td>
<td>4,285</td>
</tr>
<tr>
<td>Obs. period</td>
<td>1892–2014</td>
<td>1907–1941</td>
<td>1908–1918</td>
<td>1874–1907</td>
</tr>
<tr>
<td>Eff. lake %</td>
<td>0.89</td>
<td>0.42</td>
<td>1.65</td>
<td>0.72</td>
</tr>
<tr>
<td>M.A. flow (l/s/km²)</td>
<td>60.2</td>
<td>35.5</td>
<td>21.4</td>
<td>18.6</td>
</tr>
<tr>
<td>M.A. flood (m³/s)</td>
<td>319</td>
<td>262</td>
<td>137</td>
<td>126</td>
</tr>
<tr>
<td>Flood peak/daily flood</td>
<td>1.16</td>
<td>1.04</td>
<td>1.02</td>
<td>1.04</td>
</tr>
<tr>
<td>Reservoir capacity (%)</td>
<td>0</td>
<td>12</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td># Reservoirs</td>
<td>0</td>
<td>4</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

Note: Reservoir capacity is given as percentage of annual runoff.

Information about the ratio between flood peaks and daily floods, reservoir capacity and number of reservoirs was obtained for Bulken (Holmquist 2003), Lalm (Dragset 2000), Strøndefjord (Kleivane 2013), and Labru (Dragset 2001).

Figure 2 | Average seasonal variation in flows (left column), time series of annual maximum floods (right column), and the seasonal variation of floods (middle column) for the four study catchments.
**Systematic flood observations**

We extracted daily streamflow observations for our four study sites from the national hydrological database at NVE. In Figure 2, time-series of annual maximum floods \((Q)\) are shown (in black) together with the historical floods (as circles). Table 1 lists the length of the available systematic data as well as the mean annual flood. Bulken has the longest record with 123 years of data. For the other three catchments, we used data from the time-period prior to river regulations, which have influenced discharge. We see that Bulken had the largest observed flood in 2014, Lalm in 1938, Labru in 1879, and Strandefjord in 1917.

**Historical flood information**

A recent program of documenting the largest floods in Norway during recent centuries is found in Roald (2013) and summarized in Brázdil et al. (2012). More detailed reports systemizing the historical information are under development at NVE. The following descriptions are based on this material.

The oldest event for which we have historical information was at Bulken in 1604. There is a carving in the stone church at Voss, which indicates the water level (Figure 3). There were also large floods in western Norway (including Bulken) in 1719, 1743, 1745, and 1790, but information about the water levels are less certain, and most likely they were smaller than the 1604 flood. The major flood (Storeflaumen), which occurred 3–11 December 1743, covered catchments in the inland part of western Norway (Figure 4). The autumn of 1743 was cold, with the ground freezing early, followed by an early snowfall in the mountains. The temperature rose abruptly in early December, causing the snow to melt. The flood peaked on 4–5 December in most locations. Heavy rainfall started on 3 December, continued until 11 December, with the second peak on 10–11 December. Floods reached almost the same level in 1719 and in 1790. The fact that four floods exceeding 56 m³/s occurred in the 18th century, and only two exceedances occurred during the most recent century, indicates that the flood regime may have changed.

The flood event known as ‘Storofsen’, 21–23 July 1789, covered the entire Glomma River and the upper parts of the Begna and Numedalslågen watersheds. The rivers Driva, Surna, and Orkla which flow towards the northwest were also largely affected (Figure 4). Storofsen is considered to be the largest flood event since 1345 in these areas (Roald...
The combination of a substantial amount of snow which accumulated in the mountains, a deep soil frost, and heavy rainfall that saturated the soil, constituted the initial conditions. Warm and humid air masses from the southeast were contained by much colder air masses towards the northwest and caused high rainfall over the entire region. The rainfall intensity peaked on Wednesday 22 July. The flood started on 21 July in small brooks and culminated on the 22 July. The main rivers at the bottom of the valleys flooded to unprecedented levels, and were accompanied by landslides. The water levels of this flood are known from several markings cut into rocks, and many flood levels have later been transferred to monuments erected at locations near the major rivers. For this flood we have information at Strandefjord and Lalm.

Another major flood (Storflaumen), which occurred in June 1860, was extreme in the western branch of the Glomma basin and Drammenselva including the Begna, Numedalslågen, and Skien watersheds (Figure 4). This flood was mainly caused by an exceptional snowfall accumulation in the winter of 1859/60. The spring of 1860 was cold, and snowmelt started in mid-May. An intensive rainfall event occurred between 15 and 17 June with thunderstorms, strong winds from the south, and high temperatures even in alpine areas. The peak was slightly lower than during ‘Storofsen’, but the duration of the flood event was in excess of a month. The total volume was probably considerably larger than the volume of ‘Storofsen’ which lasted only 3 days.

For Bulken we used historical flood estimates presented in Holmqvist (2015). Bulken is located at the outlet of the lake Vangsvatnet. The village of Voss and its stone church are located at the opposite end of the lake. Since the outlet of the lake has been modified several times (in the mid-19th and late 20th centuries), we used historical information about river profile to get the magnitude of the floods. Details are given in a subsection below.

At Lalm, the level of the 1789 flood was carved in a stone that later was removed for the construction of a road. The mark was leveled and transferred to a flood stone that is located close to the gauging station and a flood volume was calculated by Klæboe (1938). The 1860 flood was reported to culminate 6 inches lower than the 1789 flood at Lalm (Kleiven 1908). Since Lalm has a relatively stable river profile, we used the oldest rating curve to translate flood levels into flood volumes.

At Strandefjord in the Begna catchment, flood levels from the 1860 flood were obtained from flood marks and used to estimate flood volumes (Otnes 1983). Reports of damage indicate that the 1789 flood was even larger than

Figure 4 | The geographical extent of the catchments affected by the floods Storeflaumen in 1743 (left) Storofsen in 1789 (middle), and Storflaumen 1860 (right).
the 1860 flood (Otne\textsuperscript{s} 1983), but there are no flood marks for this event. Similarly, for these two locations, we used the oldest rating curve since the river profiles are relatively stable.

For Labru, descriptions of flood levels are found in Kristensen (19\textsuperscript{4}) and were also transferred to flood volumes using the oldest rating curve (Roald 20\textsuperscript{3}).

Table 2 below summarizes the available information for each of the historical flood events used in this study.

### Historical river profile information

Bulken gauging station is located at the outlet of the lake Vangsvatnet. The water was originally conveyed into a channel with limited capacity, which caused large fluctuations of the water level in the lake upstream. In order to avoid this, the channel was deepened and widened in 1865, and assessments from the Norwegian channel directorate (Kanalvæsenet) indicate that the flood levels were reduced by up to 6 feet (1.9 meters) (Kanalvæsenet 1888). The outlet of the lake was again modified in 1990 in order to reduce flood water levels. The gauging station at Voss therefore has two rating curves, one for the period 1892–1990 and one for the period 1990 to today. In order to assess flood discharge for historical floods before 1865, the rating curve from 1892 to 1990 was simply adjusted by 1.9 meters when compared to historical water levels.

For the Lalm, Strandefjord, and Labru gauging stations, we used the oldest available rating curves when translating water levels to river flows. Their profiles are assumed to be stable since no information regarding changes have been found.

### MODELING

#### Flood frequency modeling

For flood frequency modeling we applied the GEV distribution which is shown to be a limiting distribution for block maxima (Fisher & Tippett 1928; Gnedenko 1943; Embrechts \textit{et al.} 1997):

\[
F(x) = \begin{cases} 
\exp\left\{-\left[1 - k \left(\frac{x - m}{\alpha}\right)^{1/h}\right]\right\} \\
\exp\left\{-\exp\left(-\frac{x - m}{\alpha}\right)\right\}
\end{cases}
\] (1)

### Table 2 | Information about historical floods for the four study catchments

<table>
<thead>
<tr>
<th>Year</th>
<th>Date</th>
<th>Flood information</th>
<th>Peak flood (m\textsuperscript{3}/s)</th>
<th>Daily flood (m\textsuperscript{3}/s)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulken</td>
<td>1604</td>
<td>Flood mark on the church wall</td>
<td>900</td>
<td>776</td>
<td>Holmqvist (20\textsuperscript{5})</td>
</tr>
<tr>
<td></td>
<td>1719</td>
<td>13 August Almost equal to the 1743 flood</td>
<td>700</td>
<td>603</td>
<td>Holmqvist (20\textsuperscript{5})</td>
</tr>
<tr>
<td></td>
<td>1743</td>
<td>4–5 December Water level reaching the church choir</td>
<td>700</td>
<td>603</td>
<td>Holmqvist (20\textsuperscript{5})</td>
</tr>
<tr>
<td></td>
<td>1745</td>
<td>– Water level reaching the church</td>
<td>650</td>
<td>560</td>
<td>Holmqvist (20\textsuperscript{5})</td>
</tr>
<tr>
<td></td>
<td>1790</td>
<td>– Equal to the 1743 flood</td>
<td>700</td>
<td>603</td>
<td>Holmqvist (20\textsuperscript{5})</td>
</tr>
<tr>
<td></td>
<td>1884</td>
<td>– Water level reached 26 inches on the road towards Bergen</td>
<td>700</td>
<td>603</td>
<td>Holmqvist (20\textsuperscript{5})</td>
</tr>
<tr>
<td>Lalm</td>
<td>1789</td>
<td>21–23 July Flood marks transferred to a flood stone</td>
<td>1,648</td>
<td>1,585</td>
<td>K\textae{}boe (19\textsuperscript{3})</td>
</tr>
<tr>
<td></td>
<td>1860</td>
<td>15–22 June 6 inches lower than 1789</td>
<td>1,585</td>
<td>1,524</td>
<td>Kleiven (19\textsuperscript{0})</td>
</tr>
<tr>
<td>Strandefjord</td>
<td>1789</td>
<td>21–23 July Larger than 1860</td>
<td>&gt;550</td>
<td>539</td>
<td>Otne\textsuperscript{s} (1983)</td>
</tr>
<tr>
<td></td>
<td>1860</td>
<td>15–22 June Two flood marks</td>
<td>550</td>
<td>539</td>
<td>Otne\textsuperscript{s} (1983)</td>
</tr>
<tr>
<td>Labru</td>
<td>1860</td>
<td>15–22 June 6.0 m</td>
<td>1,285</td>
<td>1,236</td>
<td>Kristensen (19\textsuperscript{1}) and Roald (20\textsuperscript{3})</td>
</tr>
</tbody>
</table>
where $m$ is a location parameter, $\alpha$ a scale parameter, and $k$ a shape parameter. If $k = 0$ we get a Gumbel distribution. If $k < 0$ we have a Frechet distribution and if $k > 0$ we have a Weibull distribution. For estimating the parameters and the return levels, we used a Bayesian approach, where the parameters are considered as random variables. Their posterior density is calculated as:

$$p'(m, \alpha, k|\hat{x}) = \frac{l(\hat{x}|m, \alpha, k)\pi(m, \alpha, k)}{\int l(\hat{x}|m, \alpha, k)\pi(m, \alpha, k)dm \, d\alpha \, dk}$$  \hspace{1cm} (2)$$

where $\pi$ is the prior and $l(\hat{x}|m, \alpha, k)$ is the likelihood of the observation vector $\hat{x}$ given the parameters $m, \alpha, k$. The denominator makes the integral under the pdf equal one.

For the prior $\pi(m, \alpha, k)$, we used non-informative priors for the location and scale parameters, where the priors for the location parameter and the log-transformed scale parameter were uniform. In order to avoid absurd estimates of the shape parameter, we added a prior likelihood as recommended by Coles & Dixon (1999) and Martins & Stedinger (2000). Following Renard et al. (2015), a normal distribution with standard deviation 0.2 and expectation 0.0 was used as prior for the shape parameter $k$.

For calculating the likelihood using historical data we followed Gaál et al. (2010) and Stedinger & Cohn (1986). The likelihood for the systematic data was calculated as:

$$I_s = \prod_{i=1}^{n} f(x_i|m, \alpha, k)$$  \hspace{1cm} (3)$$

where $f(x_i)$ is the probability density function for the GEV distribution with the parameter values $m, \alpha, k$ evaluated for the observation $x_i$.

For historical floods it is assumed all $t$ floods exceeding a threshold $x_0$ for a period with duration $h$ are known. The likelihood of $h-t$ number of floods not exceeding $x_0$ during the period $h$ is then:

$$I_b = [F(x_0|m, \alpha, k)]^{h-t}$$  \hspace{1cm} (4)$$

where $F$ is the GEV distribution given in Equation (1).

Then, we need to include our knowledge on the floods exceeding $x_0$. In the simplest case we know only that $t$ floods exceeded $x_0$, and then this likelihood is given as:

$$I_{a1} = [1 - F(x_0|m, \alpha, k)]^t$$  \hspace{1cm} (5)$$

Alternatively, we might know that the floods that exceeded $x_0$ where within an interval defined by an upper $x_U$ and lower $x_L$ limit:

$$I_{a2} = \prod_{j=1}^{t} [F(x_U|m, \alpha, k) - F(x_L|m, \alpha, k)]$$  \hspace{1cm} (6)$$

And in the best case we know the exact magnitude of all floods exceeding $x_0$:

$$I_{a3} = \prod_{j=1}^{t} f(y_j|m, \alpha, k)$$  \hspace{1cm} (7)$$

Depending on which data we have, the total likelihood is given as a product of the three major likelihood terms:

$$I_t = I_s \cdot I_b \cdot I_{a1}$$  \hspace{1cm} (8)$$

The posterior distribution of the parameters were estimated using a MCMC method implemented in the R-package nRFA (Viglione 2012). For estimating return levels, we used the posterior modal values of the parameters.

A challenge for all catchments was to set the perception threshold $x_0$ and length of the historical flood period $h$, i.e., for which period the listed floods represent all floods above the threshold. For each catchment, the perception threshold was set to the lowest observed historical flood value in the historical periods as recommended in Prosdocimi (2017). For Bulken, Lalm, and Strandefjord we did not have any specific information about when the historical period starts. We therefore followed Prosdocimi (2017) and set the length of the historical period to be the time span from the first historical event to the end of the historical period plus the average time spacing between the historical events. Details for each catchment are given in Table 3. Labru we treated differently and we subjectively set the historical period to start in 1790 since the 1860 flood is the largest one since ‘Storofsen’ in 1789. Using the approach by Prosdocimi (2017) would set the historical period too short and therefore lead to overestimation of flood quantiles. For all catchments the historical period ends when the systematic records starts.
Experiment 1 consisted of two parts. The first part was to assess the sensitivity of estimated flood quantiles to the use of the historical flood information by estimating the flood quantiles both with and without the created historical information, and fixing the length of the historical period \( (h = n - s) \), in Equation (7). The second part was to assess the sensitivity of estimated flood quantiles to the estimated length of the historical period by (i) setting \( h \) to the known length \( (h = n - s) \), (ii) setting \( h \) to be the time span from the first historical event to the end of the historical period, and (iii) following Prodocimi (2017) and setting \( h \) to be the time span from the first historical event to the end of the historical period plus the average time spacing between the historical events. Note that \( h \) was estimated for each sub-sample \( Q^*_h \).

The resampling provided us with \( k \) estimated flood quantiles. In order to assess the stability of flood quantile estimates, the coefficient of variation of estimated flood quantiles for selected return periods was calculated. The reliability was assessed by calculating the mean absolute error (MAE) and mean relative error (MRE) between design flood estimates based on the whole resampled data series \( Q^* \) and design flood estimates based on subsets of \( Q^* \) (i.e., \( Q^*_s \) and \( Q^*_h \)) as described above.

**Experiment 2 (using systematic and historical data)**

In the second experiment, the real historical information (see Tables 2 and 3) was combined with the systematic observations. For the first part of this experiment, we wanted to explore how the length of the systematic flood data influences the flood quantile estimates. Therefore, we resampled, with replacement, the systematic flood data to obtain an independent data sample \( Q^*_s \) of length \( n \). Then, a subset \( Q^*_s \) of length \( s = 5–75 \) years was combined with the systematic information listed in Table 2 and used for estimating flood quantiles. For the second part of this experiment, we wanted to explore how the length of the historical period influences the flood quantile estimates. Therefore, we resampled, with replacement, the historical flood information that consists of the six flood values given in Table 1 and 319 zeros. This provided an independent data sample \( Q^*_h \) of length 325. Then a subset \( Q^*_{sh} \) of length \( h = 100–300 \) years was combined with the systematic data of length 123 years. For the estimation, we set the

---

**Table 3 | Length of the historical period for the four study catchments**

<table>
<thead>
<tr>
<th>Name</th>
<th>Historical period</th>
<th>Minimum length of historical period</th>
<th>Estimated length of historical period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulken</td>
<td>1604–1891</td>
<td>288</td>
<td>335</td>
</tr>
<tr>
<td>Lalm</td>
<td>1789–1906</td>
<td>118</td>
<td>176</td>
</tr>
<tr>
<td>Strandefjord</td>
<td>1789–1907</td>
<td>119</td>
<td>177</td>
</tr>
<tr>
<td>Labru</td>
<td>1860–1873</td>
<td>14</td>
<td>84</td>
</tr>
</tbody>
</table>

**Model evaluation**

In order to assess the added value of the historical information in design flood estimation, we carried out two model evaluation experiments. For both experiments, we used criteria describing reliability and stability as proposed by Renard et al. (2015). Reliability describes the ability of a fitted model to predict actual design flood levels, whereas stability describes the sensitivity to the underlying data. Stability is an important criterion for design flood estimation in order to keep the actual protection level constant. In particular, with dam safety which has to be re-assessed every 15 years, large changes in design floods estimates might be expensive for the dam owners. Both evaluation experiments were carried out for Bulken only, since this gauging station has the longest record of systematic data (123 years). For both experiments, we used a bootstrapping approach in order to create new data samples that were used for estimation. The original data sets were used as a reference when evaluating the reliability.

**Experiment 1 (using systematic data only)**

The systematic record of annual maximum floods \( Q \) was resampled \( k \) times with replacement to obtain an independent data sample \( Q_s \) of length \( n \). Then, a subset \( Q_s \) of length \( s = 5–75 \) years of the resampled record was used as systematic data in order to estimate the likelihood term \( l_s \) in Equation (3). To create historical information we used \( Q_h \) of length \( h = n - s \) that was data in \( Q^* \) not included in \( Q_s \). From \( Q_h \), we extracted all \( t \) floods above a certain threshold \( Q_0 \) in order to calculate the remaining likelihood terms \( (l_b \) or \( l_a ) \) given in Equations (4)–(7). Only samples \( Q^* \) with at least one flood in \( Q^*_h \) exceeding the threshold \( Q_0 \) were accepted.
length of the historical period in two ways: (i) we used the known length of the historical period, i.e., the length of the subset, and (ii) we followed Prosdocimi (2017) as described above. Note that \( h \) was estimated for each subsample \( Q_i \).

To assess the stability of the design flood estimates, the coefficient of variation of design flood estimates based on the \( k \) samples was used. In order to assess reliability, the distribution was fitted to the complete data set (i.e., the systematic flood data \( Q \) of length 123 years and the historical information with the six known floods in 335 years) as a ‘truth’. As in experiment 1, the MAE and MRE for the difference between the reference and the distribution fitted to the different sub-samples were calculated.

For both experiments, the evaluation of the reliability is challenging since we used the distribution fitted to the original data as truth, i.e., the truth is unknown. We assumed that the evaluation of reliability can be trusted for return periods shorter than the 123 years of data used to estimate the truth. The stability, on the other hand, can be analyzed for any return period. Using MAE and MRE to measure the reliability also gives information about stability, since a small MAE indicates a small spread in the estimates.

### Return level plot

To demonstrate the use of historical information in selected Norwegian catchments, the fitted distributions were plotted together with their 95% credibility intervals and their empirical distribution. Each plot includes models fitted to systematic data only, and to both systematic data and historical information. We applied the Cunnane plotting position (Cunnane 1978) where the exceedance probability of the flood, whereas the lowest threshold is exceeded 24 times and represents a five-year flood. The flood magnitudes estimated for the 20, 50, and 100 year return periods using the whole original record, were used as the ‘truth’ in Figures 5 and 6. The longest return period of 100 years is shorter than the 123 used for estimating the truth. In Figure 7, we show the CV for return periods of 20, 200, and 500 years. In Figure 8, boxplots of estimated return levels are shown for selected return periods for both thresholds. The boxplot of the reference is based on the 123 years of data resampled with replacement. For the systematic record, 30 years were used (the first 30 years of the resampled data), with the length of the historical period taken to be the remaining 93 years.

Results for the second part of experiment 1 in Figure 9 show the sensitivity of the evaluation criteria to the estimated length of the historical period. The historical data were created only for the highest threshold (500 m³/s), where \( i \) is the rank, \( l \) is the number of extraordinary floods, \( n \) is now the length of the period for which we have information about floods (note that \( n + h + s \)), \( s \) is the length of the systematic record, \( h \) is the length of the historic information, and \( e \) is the number of extraordinary floods in the systematic record (note that the number of historical floods \( t = l - e \)).

### RESULTS

#### Model evaluation

**Experiment 1**

The results for the first part of experiment 1 are shown in Figures 5–7 where the MAE (Figure 5), MRE (Figure 6), and coefficient of variation (Figure 7) are plotted as a function of the length of the systematic record. Note that for all these figures, the length of the historical period decreases as the length of the systematic record increases, thus the strength of the likelihood term \( l_s \) increases and \( l_b \) or \( l_s \) decreases as the length of the systematic record increases. Thresholds at 420 m³/s and 500 m³/s were used in order to evaluate the sensitivity to the ‘perception threshold’. The highest threshold is exceeded eight times and represents approximately a 15-year flood, whereas the lowest threshold is exceeded 24 times and represents a five-year flood. The flood magnitudes estimated for the 20, 50, and 100 year return periods using the whole original record, were used as the ‘truth’ in Figures 5 and 6. The longest return period of 100 years is shorter than the 123 used for estimating the truth. In Figure 7, we show the CV for return periods of 20, 200, and 500 years. In Figure 8, boxplots of estimated return levels are shown for selected return periods for both thresholds. The boxplot of the reference is based on the 123 years of data resampled with replacement. For the systematic record, 30 years were used (the first 30 years of the resampled data), with the length of the historical period taken to be the remaining 93 years.

Results for the second part of experiment 1 in Figure 9 show the sensitivity of the evaluation criteria to the estimated length of the historical period. The historical data were created only for the highest threshold (500 m³/s),
and the estimated distribution using all 123 years of AMS data was used as the ‘truth’.

Experiment 2

In experiment 2 we used the historical flood information at Bulken (see Tables 2 and 3) together with the systematic flood data (123 flood values). The outcome of experiment 2 is shown in Figures 10 and 11 where the MAE, MRE, and CV of flood quantiles are plotted as a function of the length of systematic record and the length of historical period, respectively. For both figures, the MAE and MRE used the estimated quantiles based on all the systematic data and the magnitude of the historical flood data as the
truth. In Figure 10, the smallest historical flood was used as the perception threshold (i.e., $x_0 = 560$ m$^3$/s), whereas in Figure 11 the smallest of the resampled floods was used as the perception threshold. In Figure 11, the length of the historical period was set in two ways: (i) by the length of the resampled data and (ii) estimated based on Prosdocimi (2017). In Figure 12, boxplots of estimated flood quantiles are shown for selected return periods. The length of the systematic record is 30 years and all historical data were used in the estimation.

**Return level plots**

The estimated return level together with a 95% confidence interval are plotted for the four study catchments. Figure 13 compares: (i) the use of systematic data only and (ii) the combination of systematic and historical data. The empirical distribution of all data (systematic and historical) is included.

**DISCUSSION**

The first objective of this paper was to assess the added value of using historical data for flood quantile estimation, and how the added value depends on data availability. The results from both experiments 1 and 2 show that using historical flood information improves both the reliability and the stability of the design flood estimates.
The reliability is the primary criterion for choosing between model fits. All results show a reduction in MAE when using historical information, and the reduction in MAE is greatest when the length of the systematic record is shortest (Figures 5 and 10). For experiment 1, there is a small difference between using systematic data and adding information about the number of historical floods, especially for the lowest threshold and the longest return periods. This shows that when only information about number of floods above a threshold is known, a high threshold provides more information than a low threshold. Historical flood information has the potential to improve design flood estimates, since floods mentioned in historical sources are, in most cases, exceptional. These results show also that for low perception thresholds, it is important to assess the historical flood magnitudes. The results from experiment 2 are somewhat contradictory to those from experiment 1. Now, there is only a small difference between using the number of historical floods and the magnitude of the historical floods, indicating that it is sufficient to know only the number of floods above a threshold. The reason is that the perception threshold is higher (560 m³/s) and represents approximately a 50-year event. A key to understanding this difference is found in the relative strengths of the likelihood terms $l_s$, $l_b$, and $l_{a1}$ in Equation (8). In experiment 2, the combined likelihood for the historical period ($l_b \cdot l_{a1}$) is less sensitive to which of the terms $l_{a1}$ and $l_{a3}$ are used, since $l_b$ contains 285 years of no floods and $l_{a1}$ and $l_{a3}$ contain six flood events.

The MRE reflects the bias in the estimation. From experiment 1, we see that using only systematic data has the
tendency to underestimate the flood quantiles when we have less than 20 years of systematic data, and using information about number of floods exceeding a threshold overestimates the flood quantiles when we have less than 50 years of data (Figure 6). For experiment 2, we also see that the MRE depends on length of systematic record. However, using information about number of floods exceeding a threshold provides the least biased results (Figure 10).

Finally, the results show that using information about the magnitude of historical floods results in the more stable design flood estimates, whereas using only systematic data result in the lowest stability (Figures 7 and 10). The reduction in CV is largest when the systematic sample size is the shortest. For experiment 1, the improvement in stability when using the number of floods in the historical period is low when the length of the systematic record is longer than around 25 years and the length of the historical period is shorter than 98 years (Figure 7). In experiment 2, we do not see this behavior (Figure 10). This behavior in experiment 1 is related to the relative strengths of the different likelihood terms in the estimation \((l_s, l_b, \text{ and } l_a)\), as discussed above for the results of MAE.

All three evaluation criteria show a small sensitivity to the length of the historical period in the range from 100 to 300 years (Figure 11) when all systematic data are used. As for Figure 10, there is a small difference between using historical flood sizes and using the number of historical floods. Also the sensitivity to the length of the systematic record is relatively small when all historical data are used. This demonstrates the strength of combining these two data sources.

Figure 10 | Mean absolute error (MAE), mean relative error (MRE), and coefficient of variation (CV) of estimated return levels as a function of the length of the systematic data period (5–123 years). The systematic data were resampled with replacement and combined with all the historical flood information.
The results discussed above are valid for this specific case study, but they confirm results from other studies. The use of historical information decreases the estimation uncertainty of flood quantiles (e.g., Stedinger & Cohn 1986; Brázdil et al. 2006; Payrastre et al. 2011; Viglione et al. 2013; Macdonald et al. 2014; Schendel & Thongwichian 2017). We therefore suggest that a general behavior for most sites would be that a smaller estimation bias (both MAE and MRE) and improved stability indicated as smaller CV is achieved when historical information is used, and that the improvement is the most pronounced when the amount of systematic data is short. We also suggest that for a relatively
low perception threshold, knowing the sizes of the historical floods provides the best results. For a high threshold, knowing the number of historical floods is sufficient in order to improve estimates, and confirm results from Payrastre et al. (2011) and Macdonald et al. (2014) that when the length of the systematic data is shorter than the return period of the perception threshold, the peak discharge estimates are not needed. Several previous studies have also shown that a high perception threshold leads to improved design flood estimates (Stedinger & Cohn 1986; Cohn & Stedinger 1987; Martins & Stedinger 2001). In particular, Payrastre et al. (2011) suggest that a perception threshold with a return period that is a bit longer than the length of the systematic record, is most effective in reducing estimation uncertainty. This way of using historical information is also less sensitive to biases and uncertainties in estimates of historical flood sizes.

The results show that the evaluation criteria are sensitive to how the length of the historical period is estimated (Figures 9 and 11). In experiment 1, we see that using the maximum distance estimator provides best results, whereas using the year of the earliest historical flood event as the starting point for the historical period contributes to an overestimation of return levels (Figure 9). This confirms results from Prosdocimi (2017). For experiment 2, we see that using the estimated length is even better than using the known length of the historical period (Figure 11). The maximum distance estimator suggested by Prosdocimi (2017) should therefore be a standard approach for estimating the historical period length.

The results in Figure 13 demonstrate that adding historical information, in particular helps to reduce estimation uncertainty. In particular, for Lalm, Strandefjord, and Labru with relatively short local flood records, historical information is useful. At Bulken the systematic record is long, and we see limited added value in using historical information. At Strandefjord and Labru the use of historical flood leads to a considerable increase in the estimated flood

![Figure 13](image-url)
quantiles, indicating that the relatively small record of systematic data has too few large floods.

Possible non-stationarities in floods were ignored in this study. Non-stationarity might be manifested as either quasi cyclic flood-rich and flood-poor periods (for European studies see, e.g., Mudelsee et al. 2004; Jacobiet et al. 2005; Brätzil et al. 2005; Glaser et al. 2010; Kundzewicz 2012; Swierczynski et al. 2013; Hall et al. 2014), or a systematic trend or sudden shift in flood sizes (e.g., Wilson et al. 2010; Machado et al. 2015). The quasi cyclic variation can be linked to external drivers like circulation indices (e.g., Machado et al. 2015), and the periods that are flood rich and poor depend on the region (Macdonald & Sangster 2017). It might, however, be challenging to establish significant clusters of flood-poor and flood-rich periods as shown for German data in Merz et al. (2016).

There are several studies investigating possible long-term trends in floods in Norway. The historical period covered starts in the ‘little ice age’ which culminated in the middle of the 18th century when most glaciers in Norway were at their maximum. The period covered by systematic data (1880 until today) is characterized by an increase in both temperature and precipitation (Hanssen-Bauer et al. 2015). No obvious trends are found in the magnitudes of annual maximum floods in Norway (Wilson et al. 2010). However, trends are found in flood generating processes characterized by (i) a transition from snowmelt floods to rain-dominated floods in many catchments (Vormoor et al. 2016), (ii) spring floods arrive earlier in the year (Wilson et al. 2014), and (iii) rain-dominated catchments have a tendency for more frequent floods (Vormoor et al. 2016).

These two forms of non-stationarity (quasi cyclic variations and long-term trends or shifts) are addressed in different ways for design flood estimation. Since design flood estimates are used for assessing average risk over the lifetime of a construction, it is desired that design flood estimates are stable over time and are not sensitive to quasi cyclic variations in flood sizes on annual to decadal time scales. The effect of cyclic variations can effectively be removed by a temporal extension of flood time series using historical information (Macdonald et al. 2014). The results in this paper demonstrate that the temporal extension of flood data make flood quantile estimates more stable and less sensitive to one or more outstanding floods, and thus make the estimates less sensitive to these quasi cyclic variations.

Temporal trends in floods sizes, however, need to be accounted for and should also be projected into the future in order to assess the future flood risk. In our case studies, we used only floods from catchments and time periods not influenced by river regulations and negligible changes in land use. The studies of trends in floods in Norway (cited above) indicate that the assumption of long-term stationarity of annual maximum floods is reasonable. The use of a stationary model for estimating design floods is therefore reasonable. A stationary model is also used as a reference for addressing non-stationarity in flood sizes caused by climate change. The standard approach is to establish a climate factor that describes the expected change in design flood estimates from a stationary reference (Lawrence 2016).

These results indicate that using historical information has the potential to improve design flood estimates and could be included in future national guidelines. According to the current Norwegian guidelines for design flood estimation, at least 30 years of data should be available for a local flood frequency analysis, and if less than 50 years of data are available, a two-parameter distribution is recommended (Midttømme et al. 2011; Castellanin et al. 2012). The results in this paper indicate that, provided sufficient historical information is available, even five to ten years of systematic data are sufficient for estimating design floods with a reliability and stability comparable to an estimate based on 50 years of systematic data.

In this study, we have not addressed the sensitivity of the result to biases in the historical information. In our case studies, the bias in the historical information is, to a large degree, related to our assumption of stability of river profile for Labru, Lalm, and Strandefjord, and to a rather simplistic quantification of historical flood sizes at Bulken. In the latter case, we do not have the necessary information to establish rating curves for the period before the modification of the river profile in the 19th century.

In this study, we have not used information about naturalized floods, i.e., estimates of flood sizes as if there were no reservoirs. The three catchments Lalm, Strandefjord, and Labru are all today heavily regulated, each with several reservoirs (see Table 1). The effect of river regulations is characterized by a non-uniform decrease in flood
magnitudes that influences both the mean, the spread, and the skewness of the annual maximum floods. Naturalized flood estimates (i.e., estimated flood sizes where the effect of reservoirs is removed) are available in some of these catchments. These data were not used in this study due to the large estimation uncertainty that is unique for each flood event. Naturalized floods are based on water balance calculations for each reservoir, and need time series of overflow, gate openings, power production, and reservoir levels. The flood estimates are uncertain on a daily time resolution since: (i) in some of the reservoirs, the water levels were measured at a weekly time resolution before 1990; (ii) reduction of flood peaks in natural lakes before the reservoirs were established is not known or not accounted for; (iii) the traveling time from the reservoirs to downstream reservoirs and to the gauging station is partly ignored.

CONCLUSIONS

In this paper, the use of historical information in flood frequency analysis was investigated. The historical information could be either the number of floods or the magnitude of all floods above a threshold for a specified period. We evaluated the added value of historical information using criteria measuring reliability and stability of the design flood estimates. We also demonstrated how historical information could be used in four selected catchments in Norway. Based on this study, the following conclusions were drawn:

- There is added value in using historical information. The added value is greatest when the magnitude of the historical floods is known, and when the length of the systematic record is short. If only the number of floods above a perception threshold is known, the added value is greatest when the perception threshold is relatively high. In our four study catchments it is shown that even the longest time series of 123 years benefits from the use of historical information. It is also recommended to follow Prosdocimi (2017) and set the length of the historical period to be the time span from the first historical event to the end of the historical period plus the average time spacing between the historical events.

- It is feasible to include historical information in the flood frequency estimation in many Norwegian catchments. A major challenge is to translate information about historical floods to flood levels and discharges, and in this study we selected – on purpose – catchments where information about flood levels is available. For other catchments, the use of historical information might therefore be even more challenging and detailed analyses might be required to assess flood magnitudes. In this study, we have identified 1789 and 1860 as years with exceptional flood magnitudes in eastern Norway, and these years should be considered for flood studies in other catchments in this region.

A major limitation in using this approach is the nature of the flood information available. Frequently, as in this study, water levels on flood stones or marks are all that is available. In such cases, the stationarity of the river profile is often assumed. In other cases, information about flood damage rather than flood levels, is all that is available. In Norway, flood damage information for large floods is typically sourced from either written documents or tax reduction records which indicate the farms that have suffered damage. Using such information is more time-consuming and requires detailed mapping combined with routing models in order to assess the flood magnitudes which have caused damage.

In order to make historical flood information easily accessible, NVE has now begun storing historical flood magnitudes in the national hydrological database and is developing a set of analysis tools available for use with such data. The registration of historical flood magnitudes will be encouraged in future assessments of design floods.

ACKNOWLEDGEMENTS

This work was jointly funded by the Research Council of Norway and Energy Norway (grant ES519956 FlomQ) and internal research funding at the Norwegian Water Resources and Energy Directorate. The data were extracted from the national hydrological database at the Norwegian Water Resources and Energy Directorate and are available upon request to the main author. The authors...
would like to thank the reviewers for their helpful and constructive comments and suggestions that greatly contributed to improving the manuscript.

REFERENCES


Hanssen-Bauer, I., Forland, E. J., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A., Nilsen, J. E. Ø., Sandven, S., Sando, A. B.,


Kristensen, I. 1911 Description of Numedalslågen. Vassdragsdirektoratet, Oslo.


Lawrence, D. 2016 Climate Change and Future Floods in Norway (Klimaendring og framtidige flommer i Norge) NVE Rapport; 2016:81, 66 pp., NVE, Oslo, Norway.


Ottes, J. 1985 Large Floods in the Begna River (Store flommer i Begna.). Along Begna, the association for regulation of the Begna catchment over a period of 75 years (Langs Begna, Foreningen til Begnavaassdragets Regulering 75 år). Harald Lyche & Co, Drammen, Norway.


Roald, L. 2013 Floods in Norway (Flom i Norge). Forlaget Tom & Tom, Oslo.


Swierczynski, T., Lauterbach, S., Dulski, P., Delgado, J., Merz, B. & Brauer, A. 2015 Mid- to late Holocene flood frequency changes in the northeastern Alps as recorded in varved sediments of Lake Mondsee (Upper Austria). Quaternary Science Reviews 80, 78–90.


Viglione, A. 2012 nsRFA: Non-Supervised Regional Frequency Analysis. R package version 0.7-11.


