AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE

10.1029/2017WR022460

Bayesian Regional Flood Frequency Analysis for Large Catchments

Thordis L. Thorarinsdottir[1], **Kristoffer H. Hellton**[1], **Gunnhildur H. Steinbakk**[1], **Lena Schlichting**[2], and **Kolbjørn Engeland**[2]

[1]Norwegian Computing Center, Oslo, Norway, [2]Hydrological Modeling Section, Norwegian Water Resources and Energy Directorate (NVE), Oslo, Norway

Key Points:

- The proposed regional model for flood estimation in Norway has significantly better predictive performance than the current framework
- Bayesian model provides direct quantification of flood estimation uncertainty
- Bayesian model averaging allows for efficient covariate selection

Correspondence to:
K. H. Hellton,
hellton@nr.no

Citation:
Thorarinsdottir, T. L., Hellton, K. H., Steinbakk, G. H., Schlichting, L., & Engeland, K. (2018). Bayesian regional flood frequency analysis for large catchments. *Water Resources Research*, *54*. https://doi.org/10.1029/2017WR022460

Received 22 DEC 2017
Accepted 21 AUG 2018
Accepted article online 27 AUG 2018

Abstract Regional flood frequency analysis is commonly applied in situations where there exists insufficient data at a location for a reliable estimation of flood quantiles. We develop a Bayesian hierarchical modeling framework for a regional analysis of data from 203 large catchments in Norway with the generalized extreme value distribution as the underlying model. Generalized linear models on the parameters of the generalized extreme value distribution are able to incorporate location-specific geographic and meteorological information and thereby accommodate these effects on the flood quantiles. A Bayesian model averaging component additionally assesses model uncertainty in the effect of the proposed covariates. The resulting regional model is seen to give substantially better predictive performance than the regional model currently used in Norway.

1. Introduction

Flood frequency analysis (FFA) is a statistical data-based approach to determine the magnitude of a flood event with a certain return period. If sufficient data are available at a single site, an extreme value distribution is usually fitted to the observed data (at-site FFA or local model). However, ungauged sites or sites suffering from incomplete data require the use of data from nearby or comparable gauged stations. We refer to this as regional flood frequency analysis (RFFA).

The motivation for this study is the need to update the current guidelines for estimation of the design flood in ungauged catchments in Norway (Castellarin et al., 2012; Midttømme et al., 2011). The current guidelines are based on recommendations that are more than 20 years old (Sælthun et al., 1997). With increased data availability (20 more years) and the development of new Bayesian inference methods, including for engineering applications (e.g., Ball et al., 2016), there is a significant potential for improvement to the current RFFA methods.

The classical approach for RFFA is the index flood method; see, for example, Dalrymple (1960) and Hosking and Wallis (1997). This method assumes that the flood frequency curve for all sites in a region follows the same distribution up to a scaling factor. The index flood method hence consists of three independent steps: (1) identification of homogeneous regions or similar stations, (2) estimation of the index flood, and (3) derivation of the growth curve that gives factors for scaling an index flood to a suitable return level. If no appropriate data are available, the index flood is derived by regional regression analysis or based on nearby measurement stations (scaled with respect to catchment area), and a regional growth curve is applied. An overview of European procedures for RFFA is given in Castellarin et al. (2012). For step (1), typical approaches are to use fixed geographical regions, station similarity, or focused pooling (region of influence). In Austria, an interpolation approach is used. For step (2), linear regression, possibly combined with transformation of independent and/or dependent variables is used. For step (3), maximum likelihood, ordinary moment estimation, or estimation based on L-moments are commonly used.

The regional approach currently used in Norway is based on (1) fixed geographical regions, (2) linear regression on transformed variables, and (3) a unique growth curve for each region. The growth curve of each region is an average of all growth curves within a region based on an L-moment estimator. This approach does not account for parameter uncertainty, neither in the regression equation for the index flood estimation nor in the growth curves. The model uncertainty originates from the model selection in the regression analysis for the index flood model as well as the selection of a parametric distribution for the growth curve, whereas the parameter uncertainty originates from a limited sample size and measurement errors (e.g., Steinbakk et al., 2016).

A Bayesian approach accounts for these uncertainties and easily provides the predictive distribution of design floods (e.g., Fawcett & Walshaw, 2016; Kochanek et al., 2014). The potential for increasing the reliability of design flood estimates by including/improving the knowledge basis from which flood estimates are derived has increased the popularity of Bayesian methods. In addition, such methods are well suited for combining sources of information, such as historical information and expert judgments (e.g., Parent & Bernier, 2002; Reis & Stedinger, 2005). In the most recent update of the guidelines for FFA in Australia, the Bayesian approach is recommended (Ball et al., 2016). In a recent study on RFFA in small catchments in Norway, a Bayesian index flood approach is recommended (Glad et al., 2014).

Bayesian hierarchical models taking into account spatial and temporal structures have been used to describe extreme values in meteorology including extreme precipitation (Cooley et al., 2007; Dyrrdal et al., 2015; Renard et al., 2013), wind, and storm surges (Fawcett & Walshaw, 2016). These Bayesian hierarchical models have a data layer described by an extreme value distribution at each site which depends on some unknown parameters, typically location, scale, and shape. The spatial dependency is, in most cases, modeled by letting each parameter in the extreme value distribution depend on geographical, meteorological, and other location-specific characteristics. For some applications, a spatial dependency is based on distance, for example, the Austrian RFFA procedure described in Castellarin et al. (2012) and the regional approaches for precipitation presented in Renard et al. (2013) and Dyrrdal et al. (2015). In a recent review of statistical modeling approaches for spatial extremes, Davison et al. (2012) conclude that models of this type *allow a realistic and flexible spatial structure in the marginal distributions and thus enable a good assessment of the variation of return levels across space*. If, on the other hand, the aim is to model the spatial structure of individual extreme events rather than return levels, max-stable process models seem more appropriate (Asadi et al., 2015; Davison et al., 2012).

The main objective of this study is to establish a hierarchical Bayesian model that can be used for estimating design floods in ungauged catchments. The following subobjectives were identified as follows:

1. Estimate the regional model and identify the most important predictors;
2. Evaluate the predictive performance of the regional model;
3. Compare the regional model to the existing model of Sælthun et al. (1997).

In order to achieve these aims, we build upon the work of Dyrrdal et al. (2015) and use a hierarchical Bayesian model to spatially describe the parameters of a generalized extreme value (GEV) distribution for annual maximum daily discharge in 203 catchments in Norway. The method simplifies the three-step procedure described above in that it does not require the identification of homogeneous regions or similar stations. Furthermore, the joint estimation of all parameters of the marginal distribution, rather than separate assessments of the index flood and the growth curve, provides a coherent framework for assessing the full estimation uncertainty.

For comparison, we also fit a local extreme value distribution to the discharge series at individual sites, using a three-parameter GEV distribution if more than 50 years of data are available and the simpler two-parameter Gumbel distribution if between 20 and 50 years of data are available, following the current guidelines for statistical FFA in Norway (Midttømme et al., 2011).

The remainder of the paper is organized as follows: Sections 2 and 3 present the data and the regional GEV model, detailing the Bayesian hierarchical framework and Markov sampling, respectively. Section 4 gives an overview of the regional model currently used in Norway, and section 5 outlines the setup for model validation. Section 6 presents first the resulting model and discusses model validation in terms of reliability and stability, then explores individual return levels for certain stations, and ends with a comparison of the model to the current regional and local models. Section 6 contains some concluding discussion, and details of the Bayesian inference algorithm are given in Appendix A.

2. Data

The flood data consist of annual maximum floods from 203 streamflow stations of the Norwegian hydrological database *Hydra II* with at least 20 years of quality-controlled data for periods with minimal influence

10.1029/2017WR022460

Table 1
Overview of Covariate Information Used in the Regional Bayesian Model

Explanatory variables	Description
Longitude	
Latitude	
Effective lake	Percent of effective lake
Average fraction of rain	Average relative contribution of rain (vs. snowmelt) in the floods
Catchment area	Total area of catchment, also including parts outside Norway
Inflow	Average inflow per year
Average rain in April	Average over the period 1960–1990
Average rain in August	Average over the period 1960–1990
Snow melting in March	Average over the period 1960–1990
Catchment gradient	Difference in height meters between the 20th and 90th percentiles of the gradient profile, standardized by the total catchment length
Exposed bedrock	Percent of mountainous area
Relative catchment area	Total area divided by catchment length

from river regulations (see Engeland et al., 2016, for details). For all gauging stations, we extracted a set of catchment properties as listed in Table 1. Climatological temperature and precipitation information are derived from the 1×1-km daily data product SeNorge available at www.senorge.com. Histograms for record length, catchment areas, lake percentage, mean annual temperature and precipitation, and the contribution of rain to floods are shown in Figure 1. Figure 2 shows the spatial distribution of mean annual precipitation and temperature, mean annual maximum floods, and rain contribution.

The catchment areas range between 50 and 18,110 km^2 with a median of 247 km^2. The presence of lakes influences flood sizes, and 49.1% of the catchments have more than 1% of the catchment area (effectively) covered by lakes. For these catchments, the median effective lake percentage is 2.73%. The mean annual precipitation ranges from 291 to 3,571 mm with a median of 732 mm. We see a strong west-east gradient in the spatial distribution with the highest precipitation on the west coast. The mean annual temperature ranges from −4.6 to 6.0 °C with a median of 0.1 °C. The temperature is influenced by elevation as well as latitude in that it decreases with both elevation and latitude. The relative contribution of rain is estimated by calculating the ratio of accumulated rain and snowmelt in a time window prior to each flood and then averaging these ratios for all floods; see Engeland et al. (2016) for details. Rainfall gives the major contribution to floods in most coastal catchments, whereas snowmelt is important in inland, northern, and high-altitude catchments. The typical flood season for catchment dominated by snowmelt floods is spring and early summer, while no clear seasonal patterns are seen for the catchments dominated by rain floods.

The flood records and the associated catchment properties (catchment area, record length, mean annual runoff, and several other catchment descriptors) are available as supporting information (see Acknowledgments).

3. Methods

Extreme value theory provides a framework for modeling the tail of probability distributions. Let $V_1, \ldots, V_n$ denote continuous, univariate random variables assumed to be independent and identically distributed. If the normalized distribution of the maximum of the random variables, $\max\{V_1, \ldots, V_n\}$, converges as $n \to \infty$, then it converges to a GEV distribution (Fisher & Tippett, 1928; Jenkinson, 1955). For this reason, a GEV distribution is commonly used to model block maxima (the maxima over equally sized blocks of data) such as the annual maximum. See Coles (2001) for an introduction to the statistical application of extreme value theory.

In Norway, the three-parameter GEV distribution or a special case thereof, the two-parameter Gumbel distribution, is recommended for analyzing long data series from individual stations as these models have been found to provide the best fit for Norwegian data (Castellarin et al., 2012; Midttømme et al., 2011). We have thus

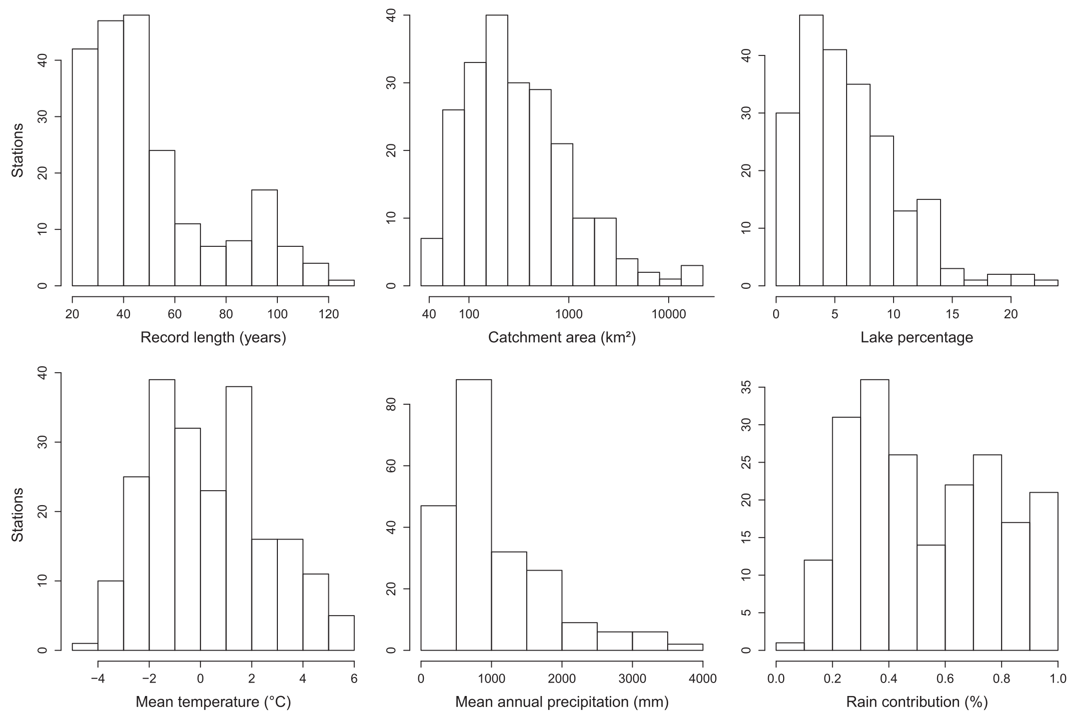

Figure 1. Histograms for record length (years), catchment areas (km^2), lake percentage (%), mean annual temperature (°C), mean annual precipitation (mm), and the estimated rain contribution to floods (%).

chosen to base our regional model on the three-parameter GEV distribution. Alternative models may provide a better fit in other regions. An overview of methods that are used for operational FFA in Europe is given in Castellarin et al. (2012).

3.1. Regional GEV Model

3.1.1. Model Formulation

The data are given by series of annual maximum floods. Denote by Y_{ts} the maximum flood in year $t \in \{1, \ldots, n_s\}$ at station $s \in \mathcal{S}$, the set of all stations in the data set, where n_s is the number of annual floods observed at station s. We assume the floods to be independent and identically distributed over time with site-specific covariates

$$Y_{ts} \sim \text{GEV}(\mu_s, \kappa_s, \xi_s), \quad t \in \{1, \ldots, n_s\}; s \in \mathcal{S}.$$

The three-parameter GEV distribution is here parameterized in terms of the location $\mu_s \in \mathbb{R}$, inverse scale $\kappa_s \in \mathbb{R}_+$, and shape $\xi_s \in \mathbb{R}$. The distribution is usually parameterized with the scale parameter $\sigma_s^2 = 1/\kappa_s$, rather than the inverse scale (Coles, 2001), but the current parametrization is common in the Bayesian setting; see, for example, Dyrrdal et al. (2015). The density is given by

$$pr_{\text{GEV}}(y_{ts}|\mu_s, \kappa_s, \xi_s) = \kappa_s h(y_{ts})^{-(\xi_s+1)/\xi_s} \exp\left\{-h(y_{ts})^{-\xi_s^{-1}}\right\}, \tag{1}$$

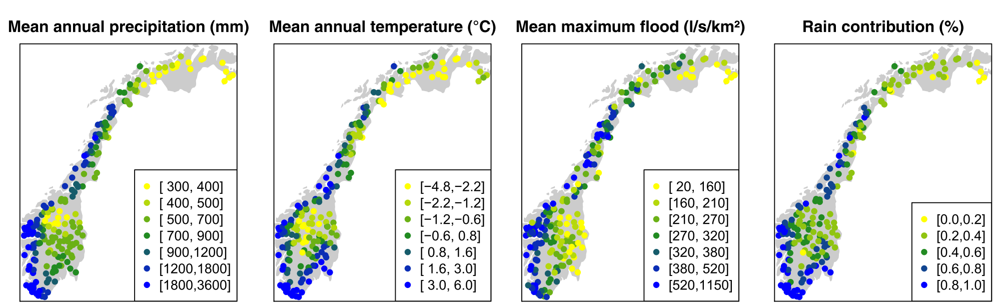

Figure 2. Spatial distribution of mean annual precipitation (mm), mean annual temperature (°C), mean annual maximum floods (l/s/km^2) and the rain contribution to floods (%).

for $h(y_{ts}) > 0$ with

$$h(y_{ts}) = 1 + \xi_s \kappa_s (y_{ts} - \mu_s)$$

when $\xi \neq 0$. For $\xi = 0$ the density is given by the Gumbel distribution

$$pr(y_{ts}|\mu_s, \kappa_s) = \kappa_s h(y_{ts}) \exp\left\{-h(y_{ts})\right\}$$

for $h(y_{ts}) > 0$ with

$$h(y_{ts}) = \exp(-\kappa_s(y_{ts} - \mu_s)).$$

For the purposes of dam safety, we are interested in estimates of certain high quantiles of the resulting GEV distribution at site s. The tail behavior is driven by the value of the shape parameter ξ_s, providing vital information on the statistical properties of the variable of interest, and is, concurrently, difficult to estimate due to the involved parametric form of the density in (1) as a function of ξ_s. To estimate the quantile p of the resulting GEV distribution function, we employ the GEV quantile function

$$z_s^p = \mu_s - \frac{1}{\kappa_s \xi_s} \left\{1 - [-\log(p)]^{-\xi_s}\right\}. \tag{2}$$

That is, z_s^p is the return level associated with the return period $1 - 1/p$ at site s.

Note that the model formulation in (1) assumes stationarity in time, ignoring, for example, potential effects of climate change. This follows Wilson et al. (2010) who found no systematic trends over time when analyzing annual maximum flood magnitudes in the Nordic countries.

3.1.2. Bayesian Hierarchical Framework

The model in (1) assumes a set of site-specific parameters (μ_s, κ_s, and ξ_s) at each station $s \in \mathcal{S}$. The spatial variability is the result of a number of factors related to the variation in terrain and climate which we aim to capture through the covariates $\boldsymbol{x}_s$ listed in Table 1. Each of the parameters μ_s, κ_s, and ξ_s is specified by a linear model, for example, for the location parameter

$$\mu_s = \boldsymbol{x}_s^\top \theta^\mu, \tag{3}$$

where the first covariate in the vector $\boldsymbol{x}_s$ is a constant equal to 1. Here we assume that $\theta^\mu \in \boldsymbol{\Theta}_{M^\mu}^\mu$ for a fixed model M^μ in which some of the elements of θ^μ are assumed to take real values, while others may be restricted to be equal to zero. The constraint $\theta_i^\mu = 0$ implies that the ith covariate does not influence the location parameter under the model M^μ. In addition to perform inference over the parameter vector θ^μ, we thus also perform inference over the set of possible models M^μ through Bayesian model averaging as described in Dyrrdal et al. (2015, section 3.3). That is, the final posterior predictive distribution is given by an average over distributions with different parameter estimates that also represent different models through some parameter values being equal to zero. We can then assess the posterior marginal inclusion probabilities of each covariate; that is, the proportion of the posterior sample for which the regression parameter associated with each covariate is nonzero.

The linear model in (3) assumes that the variability in the GEV parameters across is fully captured by the variability in the covariates, but in practice, there may be additional heterogeneity not directly captured by $\boldsymbol{x}_s$. To account for this, we include a site-specific random factor for each of the parameter τ_s^μ resulting in the model

$$\mu_s = \boldsymbol{x}_s^\top \theta^\mu + \tau_s^\mu,$$

where the random terms τ_s^μ are independent and given by a zero mean Gaussian prior distribution. Our full model can thus be written as follows:

$$\begin{aligned} Y_{ts} &\sim \text{GEV}(\mu_s, \kappa_s, \xi_s) \\ \mu_s &= \boldsymbol{x}_s^\top \theta^\mu + \tau_s^\mu \\ \eta_s &= \boldsymbol{x}_s^\top \theta^\kappa + \tau_s^\kappa \\ \xi_s &= \boldsymbol{x}_s^\top \theta^\xi + \tau_s^\xi \\ \tau_s^\nu &\sim N(0, 1/\alpha^\nu), \quad \nu \in \{\mu, \kappa, \xi\}, \end{aligned} \tag{4}$$

with $\eta_s = \log(\kappa_s)$. We aim to use uninformative priors, with a gamma distribution for the precision α^ν and independent standard normal priors for θ^ν for $\nu \in \{\mu, \kappa, \xi\}$.

The model in (4) is a slight simplification of the model discussed by Dyrrdal et al. (2015) as it assumes the random factors τ_s^ν for $\nu \in \{\mu, \kappa, \xi\}$ to be independent across the stations s; see also the discussion in section 7. The inference over the parameters $\left\{\theta^\nu, \alpha^\nu, \{\tau_s^\nu\}_{s\in\mathcal{S}}\right\}_{\nu\in\{\mu,\kappa,\xi\}}$ can thus be performed in a Bayesian fashion as described in Dyrrdal et al. (2015) using an appropriate modification of the associated package `SpatGEVBMA` (Lenkoski, 2014), available in `R` (R Core Team, 2016). The model in Dyrrdal et al. (2015) assumes an identity link on the precision parameter κ. The extension to a logarithmic link requires the calculation of new proposal distributions for the Markov chain Monte Carlo (MCMC) algorithm. This is described in Appendix A.

3.1.3. Posterior Return Levels

We run a Markov chain to return a collection of R samples

$$\left\{\theta^\nu, \alpha^\nu, \{\tau_s^\nu\}_{s\in\mathcal{S}}\right\}_{\nu\in\{\mu,\kappa,\xi\}}^{[r]}, \quad r = 1, \ldots, R, \tag{5}$$

where R is typically in the range of 50,000 to 100,000, with a suitable number of burn-in samples removed, that is, the first 10,000 to 20,000 samples. This yields a Markov sample of the parameter set $\{\mu_s^{[r]}, \kappa_s^{[r]}, \xi_s^{[r]}\}$, including both fixed and random effects. The sample of the corresponding GEV distributions directly yields, by using the GEV quantile function in (2), a sample of quantiles

$$\{(z_s^p)^{[1]}, \ldots, (z_s^p)^{[R]}\}.$$

This sample approximates, $pr(z_s^p|\{\boldsymbol{y}_s\}_{s\in\mathcal{S}})$, the posterior distribution of the pth return level at the site $s \in \mathcal{S}$. Here $\boldsymbol{y}_s$ denotes the vector of maximum floods at site s for all years t for which measurements are available. Given this sample, it is straightforward to derive approximations for the posterior mean and median with pointwise posterior credible intervals for the pth return level.

The Markov sample also constitutes a mixture distribution based on the GEV density function in (1)

$$pr(y_s|\{\boldsymbol{y}_s\}_{s\in\mathcal{S}}) = \frac{1}{R}\sum_{r=1}^{R} pr_{\mathrm{GEV}}(y_s|\mu_s^{[r]}, \kappa_s^{[r]}, \xi_s^{[r]}), \tag{6}$$

approximating the posterior predictive distribution of a future observation y_{st}. Such mixture distributions do not have an explicit quantile function, and the return level is found by simulating a number of observations from each mixture component, using the empirical quantile of observations pooled over all mixture components as an approximation. Due to the large number of Markov samples, only around 10 to 100 simulated observations from each component are needed to achieve a good approximation of the true return level. Equation (6) implies Bayesian model averaging, given that the predictive distribution is averaged of different models.

Now assume that we are interested in estimating the pth return level at a new site $s_0 \notin \mathcal{S}$ not used to estimate the model but for which the covariates $\boldsymbol{x}_{s_0}$ listed in Table 1 are available. The pth return level of the posterior predictive distribution for site s_0 is given by the empirical pth quantile found when combining simulated observations from all mixture components, based on the fixed effects $\boldsymbol{x}_{s_0}^\top(\theta^\mu)^{[r]}$, $\kappa_{s_0}^{[r]}$, and $\xi_{s_0}^{[r]}$. The quantile of the predictive distribution will be close, but not identical, to (the approximation of) the median of the posterior distribution of the pth return level.

The uncertainty is quantified by the pointwise 80% posterior credible intervals of the quantiles corresponding to a posterior sample of the three GEV parameters, given as follows:

1. sample $(\tau_{s_0}^\mu)^{[r]} \sim N(0, 1/(\alpha^\mu)^{[r]})$,
2. set $\mu_{s_0}^{[r]} = \boldsymbol{x}_{s_0}^\top(\theta^\mu)^{[r]} + (\tau_s^\mu)^{[r]}$,

and similar for $\kappa_{s_0}^{[r]}$ and $\xi_{s_0}^{[r]}$. The additional sampling of random effects gives a better (out-of-sample) calibration. Note that in general there is a higher level of uncertainty for a new site s_0, than a site $s \in \mathcal{S}$ used to estimate the model, due to the independent sampling step in (i). This additional sampling of the random effects ensures that the uncertainty is not underestimated.

3.2. Local GEV Model

We compare the regional model to a local nonhierarchical GEV model for on-site analysis without any spatial structure. In this setting, we assume that the annual maxima are described by equation (1) with independent parameters, μ_s, κ_s, and ξ_s, not sharing informative covariates. The unknown parameters are estimated by

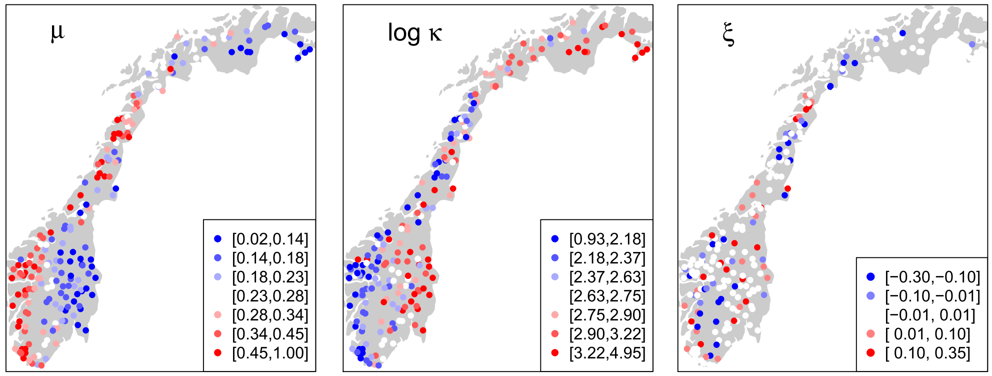

Figure 3. Spatial distribution of the median parameter estimates from the local generalized extreme value model.

MCMC within a Bayesian framework such that μ_s, κ_s, and ξ_s are updated in turn; see Steinbakk et al. (2016) for details on the prior distributions and their updates. For stations with data series less than 50 years, the shape parameter ξ_s is assumed to be zero and a Gumbel distribution is fitted instead of the GEV distribution according to FFA recommendations for Norway (Castellarin et al., 2012; Midttømme et al., 2011). Figure 3 shows the spatial distribution of the estimated parameters from the local GEV model. The map of the shape parameter ξ_s reveals that a considerable proportion of sites is estimated using a Gumbel distribution due to short data series.

4. Current Regional Model for Norway

The Norwegian Water Resources and Energy Directorate (NVE) is responsible for determining guidelines for FFA in Norway. The current regional framework for estimating flood return levels was established by Sælthun et al. (1997); see also Castellarin et al. (2012). The return values are found by first estimating the index flood based on geographical and hydrological parameters and second, using a growth curve to scale up the index flood to a certain design flood. The two steps are determined by different procedures. Both the model for index flood and the growth curve are based on initially dividing sites into distinct flood types (spring, autumn, glacier, or all-year floods) with geographical regions, four different zones for spring floods and three zones of autumn floods. Each subdivision then has a separate regression model and relevant covariates. An overview of the range of covariates, with descriptions, used in the separate regression models is given in Table 2.

One difficulty with the current regional model is the lack of a rigorous definition of the flood regions. It can be difficult to determine the appropriate region of a new site, and the estimated return values vary with the choice of region. Hence, it will be highly beneficial to model the regional or spatial characteristics of floods as a continuum in a more complex regression model. In addition, the growth curve is only available for certain predetermined return periods.

Table 2
Overview of Key Parameters for Computing Index Flood in the Current System of the Norwegian Water Resources and Energy Directorate

Parameter name	Unit
Catchment area	km^2
Mean specific annual runoff	l/s/km^2
Mean annual precipitation	mm
Effective lake	%
Exposed bedrock	%
Catchment length	km
Gradient of the main river	m/km

10.1029/2017WR022460

Table 3
Inclusion Probability (%) for the Covariates for Models for Location Parameter μ_s, Scale Parameter κ_s, and Shape Parameter ξ_s

	μ_s	κ_s	ξ_s
Constant	100	100	100
Longitude	53	99	6
Latitude	84	100	8
Percent of effective lake	98	100	2
Inflow	6	11	9
Average fraction of rain	3	22	58
Catchment area	2	5	12
Average rain in April	42	75	5
Average rain in August	100	100	5
Snow melting in March	8	16	4
Catchment gradient	45	12	2
Percent of bedrock	22	8	2
Relative area to length	13	95	11

5. Model Validation

To validate the models, we follow Renard et al. (2013) assessing reliability, or calibration, and stability. Reliability describes the consistency between validation data (data not used for calibration) and FFA predictions. A reliable, or well-calibrated, model should yield an estimated distribution close to the unknown true distribution of the data. Stability, on the other hand, quantifies the ability of the model to yield similar estimates when calibration data change.

We assess the predictive power of the regional model through a cross-validation study, such that reliability is assessed through the consistency between predictions and holdout data. Due to the heavy computational burden of the Markov sampling in the hierarchical model, a smaller number of sites, 27 stations, were selected for validation by experts to represent the range of different sites in Norway. We employ a leave-one-out cross-validation scheme for the 27 stations, where each station, out of the total of 203, in turn is left out of the model fitting, using the remaining 202 other stations. The distribution of the random effects gives the main difference between in-sample and out-of-sample predictions, as the in-sample parameter estimates allow for, and usually have, correlated random effects, while the random effects are independently drawn for the out-of-sample estimates.

5.1. Reliability

The main reliability assessment tool is the probability integral transform (PIT), displayed graphically by histograms and probability-probability (PP) plots. If observations follow the estimated distribution, the PIT values

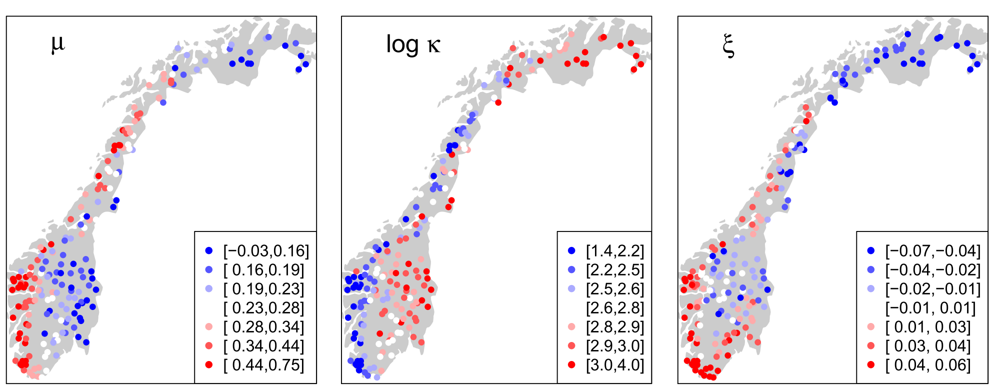

Figure 4. Spatial distribution of the median fixed effects per site, for 203 stations, for the three generalized extreme value parameters, $\mu_s^{[r]}$, $\kappa_s^{[r]}$, and $\xi_s^{[r]}$.

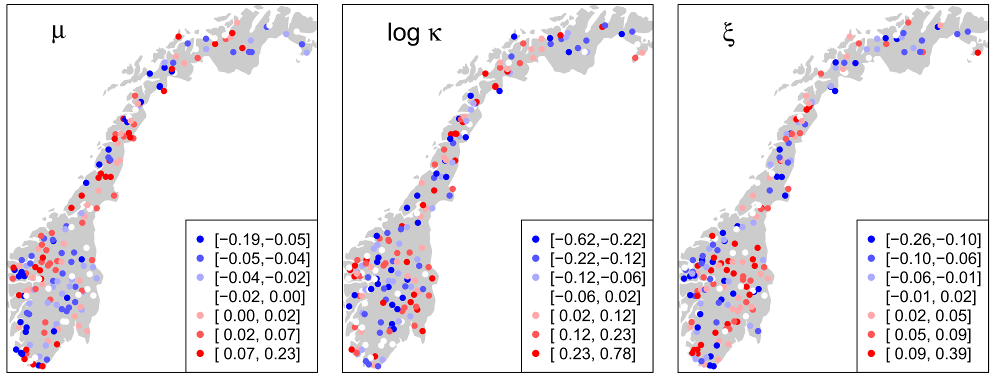

Figure 5. Spatial distribution of the median random effects per site, for 203 stations, for the three generalized extreme value parameters, $\mu_s^{[r]}$, $\kappa_s^{[r]}$, and $\xi_s^{[r]}$.

will be uniformly distributed (Dawid, 1984)

$$\hat{F}(y_{st}) \sim U([0,1]).$$

Histograms of PIT values can be assessed at a local level if the data series are long, whereas histograms with too few observations do not allow for a useful graphical assessment. Instead, one should assess regional average reliability by combining PIT values from several or all locations in a single histogram. The reliability of individual stations is assessed by PP plots displaying the observed empirical distribution of PIT values against the theoretical uniform distribution. The values should to the largest degree follow a one-to-one line, and deviation will indicate bad reliability such as overestimation or underestimation compared to observed values.

We assess the reliability in the tail of the predictive distribution using proper scoring rules, in particular the quantile score; see, for example, Friederichs and Hense (2007), Gneiting and Raftery (2007), and references therein. If denoting the predictive distribution by F and the realized observation by y, the quantile score is given by

$$s_Q(F, y|\tau) = (y - F^{-1}(\tau))(\tau - \mathbb{1}\{y \leq F^{-1}(\tau)\}),$$

for a specific probability level $\tau \in (0,1)$. Alternatively, one could use the weighted continuous ranked probability score integrating over all quantiles greater than some threshold, say the 50-year return level (Gneiting & Ranjan, 2011). But as the current regional model used in Norway is only (easily) available for certain predetermined return periods, models are compared using the quantile scores.

While histograms and PP plots of PIT values only assess the reliability of the predictions, scoring rules such as the quantile score simultaneously assess several aspects of the predictive distributions; see, for example, Stephenson et al. (2008) and Bentzien and Friederichs (2014). That is, by conditioning (stratifying) on the predicted probabilities, the scores may be decomposed into the sum of three components: reliability, resolution, and uncertainty. The resolution is related to the information content of the prediction model. It describes

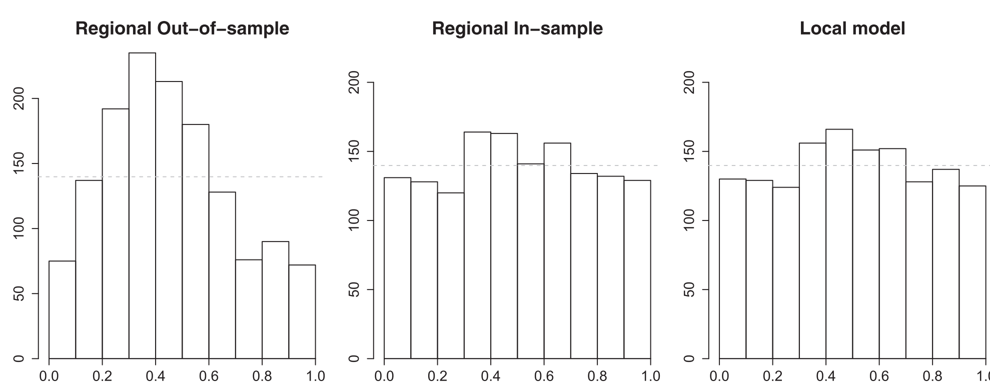

Figure 6. Histograms of probability integral transform values for all observations from the 27 cross-validation stations, a total of 1,305 observations. The regional out-of-sample model shows an excess of values away to 0 and 1.

10.1029/2017WR022460

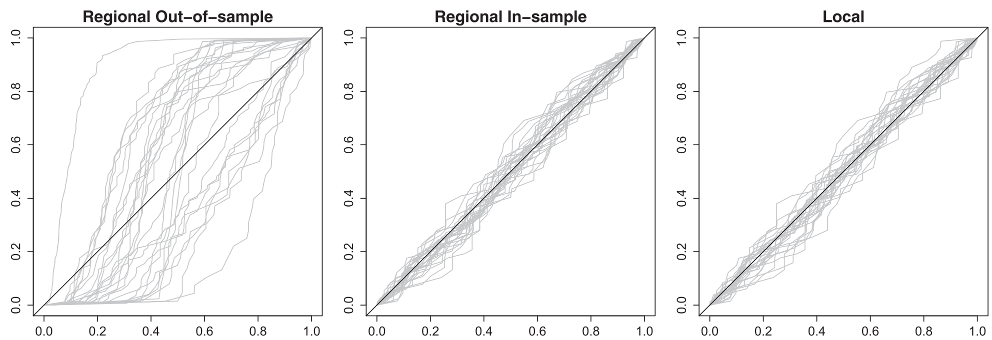

Figure 7. Probability-probability plots of the probability integral transform values for the 27 cross-validation stations. For the regional out-of-sample model some stations are highly overestimated, while others are underestimated.

the variability of the observations under different forecasts and indicates whether the prediction model can discriminate between different outcomes of an observation. In our setting, a prediction model with a good resolution is able to discriminate between locations with different flood characteristics, while a prediction model with no resolution would issue the same predictive distribution at all locations. The uncertainty component refers to the variability in the observations and is thus identical for competing models when assessed under the same data set.

5.2. Stability

Within a comparison framework, Renard et al. (2013) advised to first assess reliability and then use stability to further discriminate between models if several models are equally reliable. The stability quantifies to which degree the statistical model yields similar predictive distributions when trained on different, but identically distributed, data sets. This is a property solely of the statistical model; thus, arbitrarily large return periods can be assessed. A general procedure is as follows: First, the statistical model is fitted using all available data, yielding an estimated predictive distribution $\hat{G}$, followed by estimation on a set of leave-one-out scenarios, yielding estimated predictive distributions $\hat{F}_1, \dots \hat{F}_n$. Then each $\hat{F}_i$ is compared to $\hat{G}$ using some measure of divergence, giving an average or maximum divergence for $i = 1, \dots, n$. We will assess the stability of our model through the variability of parameter estimates, in particular the intercept term, over the leave-one-out cross-validation scheme.

6. Results

This section shows the results from the aforementioned models to predict annual flood maxima in Norway. We first assess the reliability of the Bayesian regional model following the cross-validation study and then present in-depth results for certain stations, selected to showcase the range of individual site behavior.

6.1. Explaining the Model

Relevant covariates for the regional model were first explored by assessing relationships between covariates and the estimated μ_s, κ_s, and ξ_s from the local model. This revealed that some covariates needed to

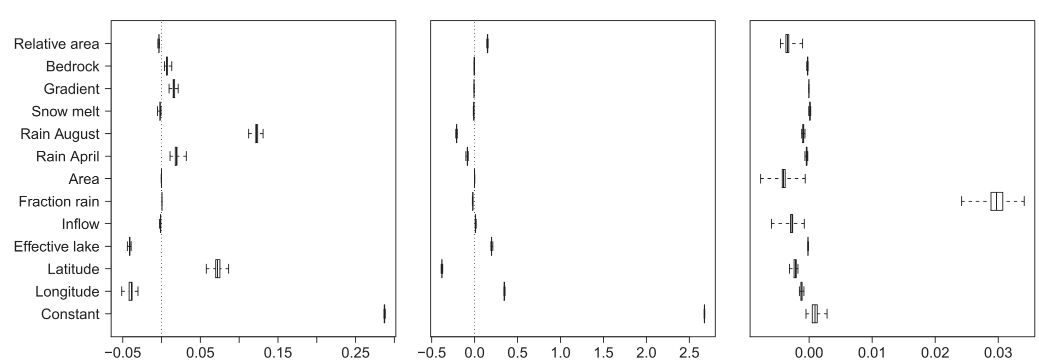

Figure 8. Assessment of stability: Box plots displaying the variability of the posterior mean regression coefficients over the 27 cross-validation models, shown for each of the 13 covariates behind μ, $\log \kappa$, and ξ.

10.1029/2017WR022460

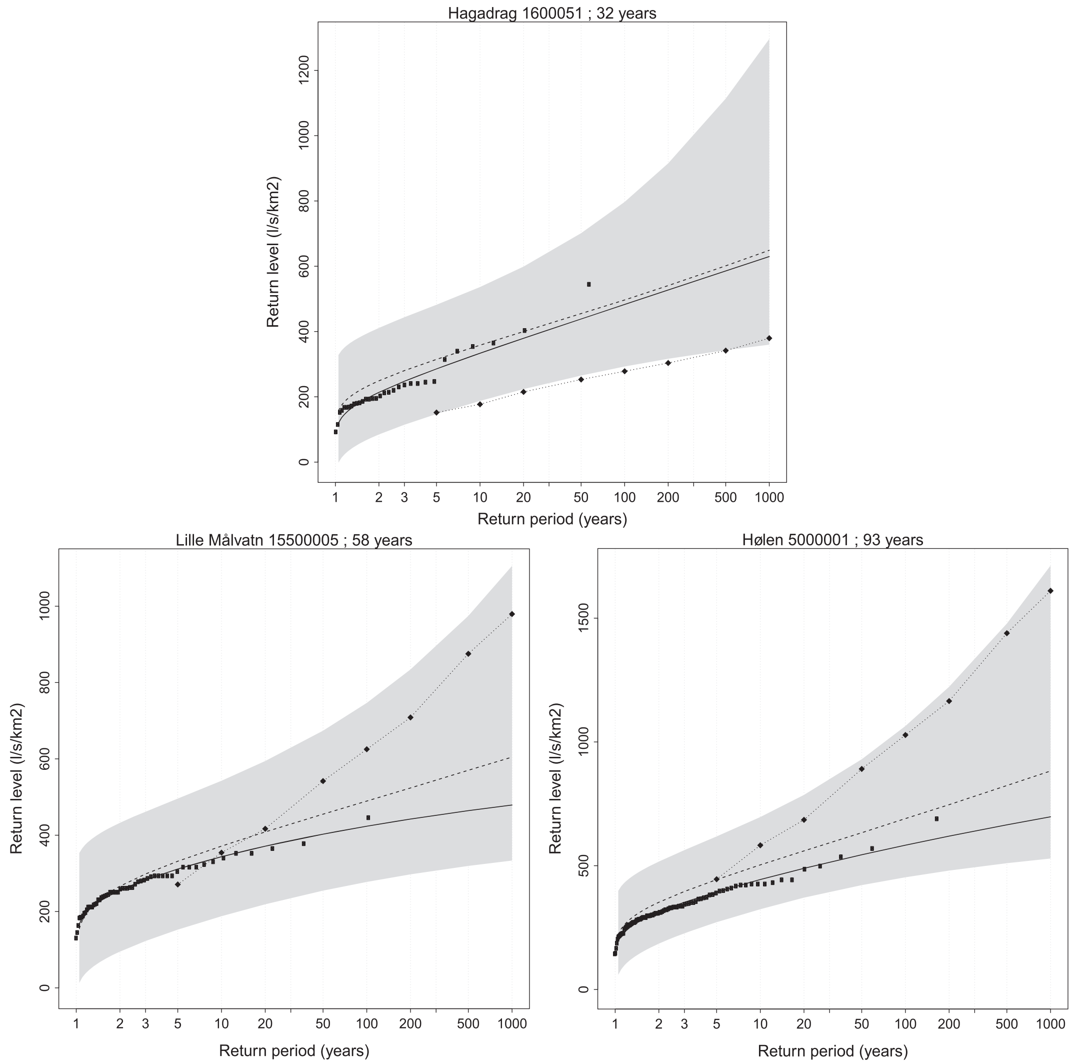

Figure 9. Estimated return levels for three stations (32, 58, and 93 years of observations, respectively) showing good agreement between the regional and local models. Black lines: Local Bayesian model. Dashed lines: Posterior predictive distribution. Gray area: 80% credibility interval for the posterior quantile distribution. Dotted line with squares: Standard regional model. Black dots: Data.

be transformed or be combined into variables. To decide on the specific set of covariates, we used stepwise linear regression optimizing Akaike information criterion (Siotani et al., 1985), with the index (mean level) flood for each station as the response. The resulting best model contained 13 covariates, seen in Table 1, and the selected variables overlap to a large degree with the covariates used by the current regional mode; for details see Table 2. New covariates not considered in the current framework are the average fraction of rain versus snowmelt, the meteorological variables (i.e., rain in April and August and snowmelt in March), and in particular, longitude and latitude. The regional model was run using 100,000 MCMC iterations and 20,000 burn-in samples, with the posterior marginal inclusion probability of each covariate given in Table 3.

It is seen that spatial location of a station, in terms of longitude and latitude, is crucial for all GEV parameters, thereby reflecting the importance of flood regions and the spatial similarity between flood sites. Maps of the median estimate of the fixed effects for μ_s, κ_s, and ξ_s are shown in Figure 4. There are clear spatial structures with an east to west and north to south gradient in both the location and the scaling parameter, while the shape parameter mainly displays a difference between coast and inland. From Figure 3 one can see that the estimated fixed effects for μ_S and κ_S agree with the distribution of the parameters from the local model, while the pattern of ξ_S is not observed in the local model. These results highlight that pooling of information across catchment sites via the linear regression model is highly beneficial.

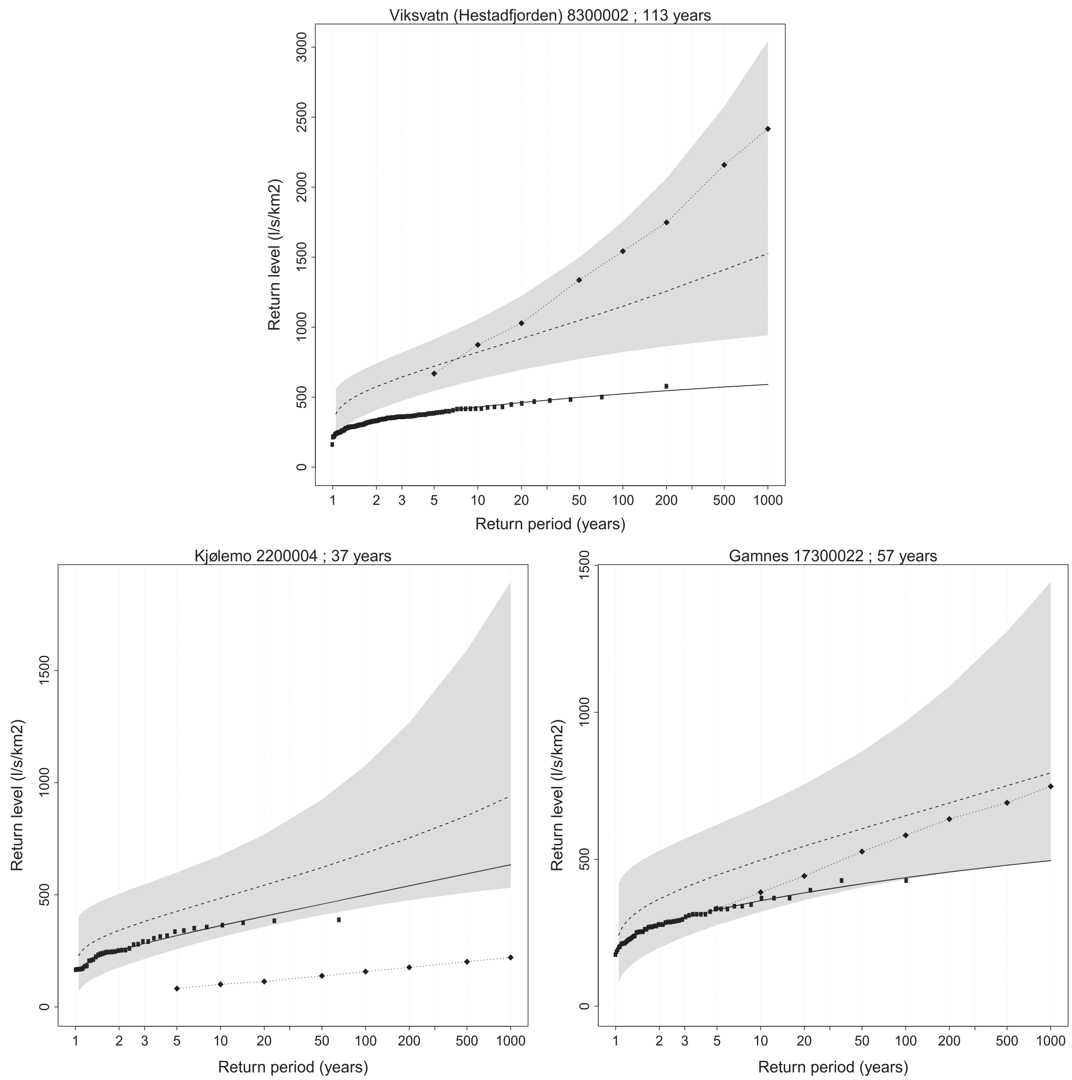

Figure 10. Estimated return levels for three stations (37, 57, and 113 years of observations, respectively) with a bad agreement between the regional and local models. Black lines: Local Bayesian model. Dashed lines: Posterior predictive distribution. Gray area: 80% credibility interval for the posterior quantile distribution. Dotted line with squares: Standard regional model. Black dots: Data.

Further, both the parameters μ_s and κ_s are explained by the percentage of effective lake, as a dominant lake in the catchment area can dampen both the mean flood level and the yearly variations, and by the average rain in April and August, as more rain will increase both average flood levels and variability. The derived covariate, catchment area divided by the basin length, is seen to significantly affect the scaling κ_s, suggesting that *long* catchments (elongated along the length) experience less variability possibly due to a dampening effect. The percent of bedrock and the catchment gradient, on the other hand, influence mainly the location parameter

Table 4
Quantile Scores for Return Periods T = 10, 50, *and* 100 *Comparing the Regional Model With the Local and the Regional Model of Sælthun et al. (1997) Currently in Use at Norwegian Water Resources and Energy Directorate (NVE)*

Model	T = 10		T = 50		T = 100	
NVE model	6.71	[6.23, 7.20]	4.04	[3.63, 4.46]	3.17	[2.80, 3.55]
Regional model	2.94	[2.72, 3.17]	1.01	[0.87, 1.17]	0.62	[0.51, 0.75]
Local model	1.95	[1.80, 2.11]	0.63	[0.57, 0.70]	0.36	[0.32, 0.40]

Note. The 90% uncertainty intervals are obtained by bootstrapping.

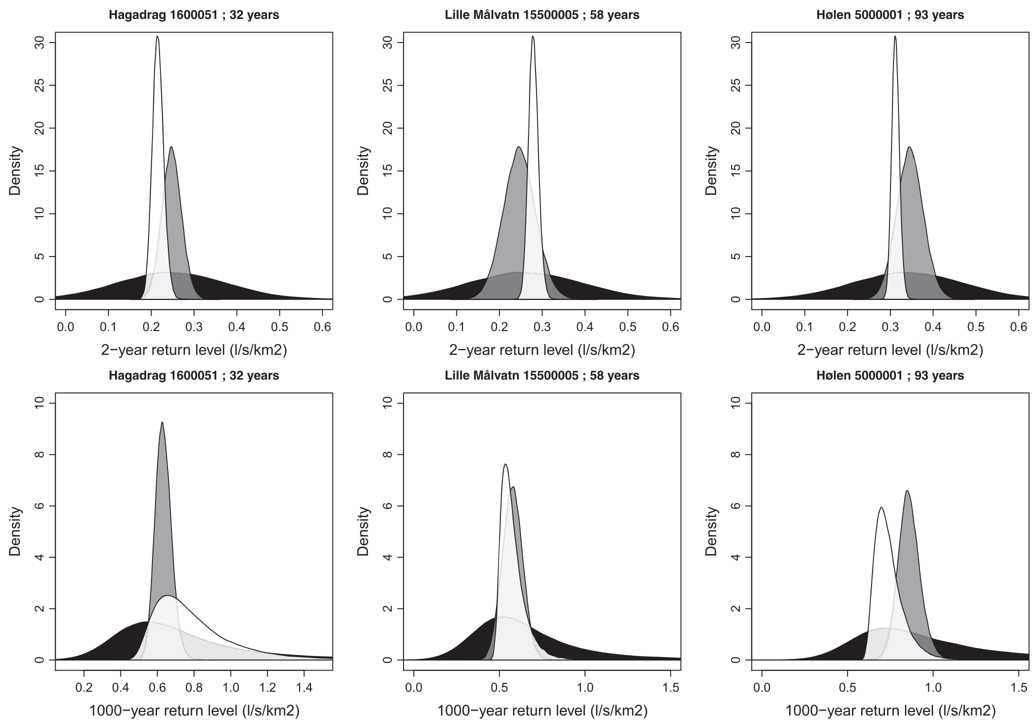

Figure 11. Full Bayesian distributions for the 2-year return level (the median flood) and 1,000-year return level for the three stations (32, 58, and 93 years of observations, respectively) showing good agreement between the regional and local models. Black: fixed effects with simulated random effects. Gray: fixed effects only. White: fixed effects with estimated random effects.

μ_s, as a larger degree of mountainous terrain within a catchment and a steeper gradient will increase the average flood level. Lastly, the shape parameter ξ_s is mainly explained by the average fraction of rain in the flood and to a smaller degree the catchment area, the relative area to length, and the inflow. In areas with a smaller fraction of rain compared to snowmelt, the annual maximum flood is more often a spring flood caused by snowmelt, which will to a larger degree be limited by an upper threshold. This characteristic would be accounted for by a negative shape parameter.

Figure 5 shows the median estimates of random effects for the three parameters, per station, and could reveal whether additional spatial effects need to be accounted for. Overall, the random effects seem to be spatially independent, apart from scattered, but small, clusters present in all three parameters which could be due to river networks or regional catchment areas.

We validate the model following section 5 and start by comparing histograms of aggregated PIT values for the out-of-sample and in-sample regional models and the local model. Figure 6 shows that the local and in-sample regional models are well calibrated over the 27 selected stations, while the out-of-sample model is somewhat overdispersive. The PP plots in Figure 7 show that the regional model will severely overestimate and underestimate the return levels of some out-of-sample stations, while the predictive distribution of most stations is seen to be well calibrated. We assess the stability of the model through the variability of the fixed effects in each of the three parameters. Figure 8 shows box plots of the mean regression coefficients over the 27 leave-one-out cross-validation models, for each of the 13 selected covariates and the 3 parameters. All estimates are seen to be very stable, in particular for μ and κ, and only the coefficient estimates of the rain contribution and area for ξ are slightly less stable.

6.2. Model Validation

To compare the fit of the different models with the observed data, we explore the results of six selected sites. Figure 9 shows the (out-of-sample) estimated return level plotted against the return period (on a log scale), for three sites with short ($n = 32$), medium ($n = 58$), and long ($n = 93$) data series. The figure shows the observed data (black dots), the median estimate of the local model (black lines), predictive distribution (dashed lines), and pointwise 80% credible bands (gray area) of the return level for the regional model and

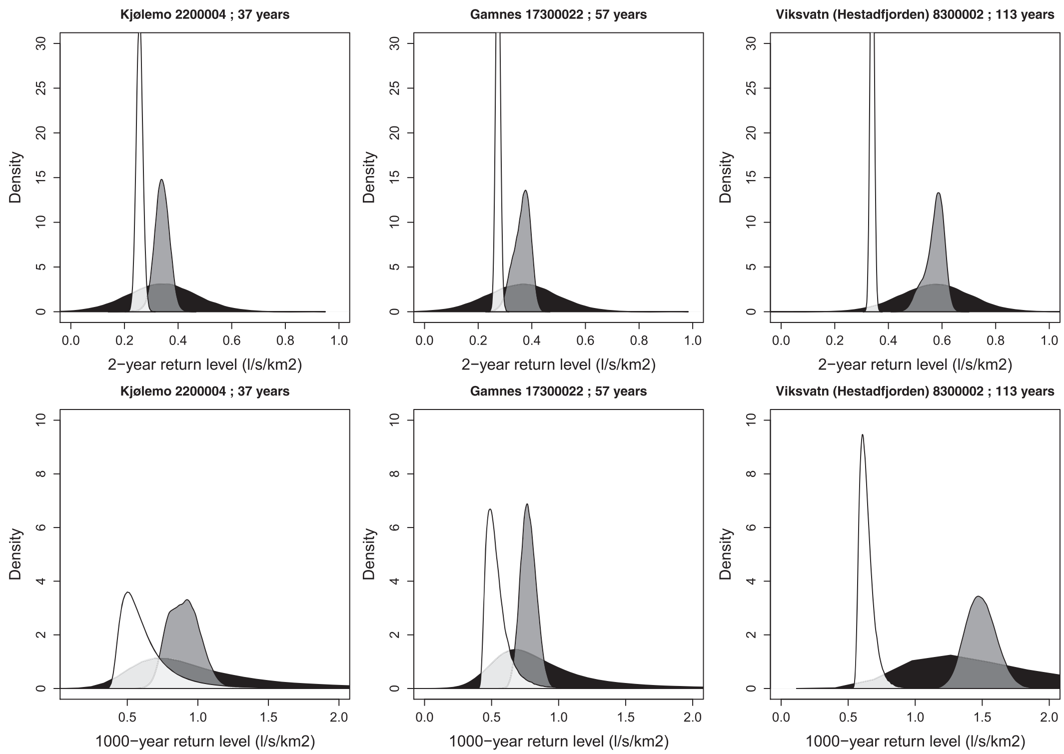

Figure 12. Full Bayesian distributions for the 2-year return level (the median flood) and 1,000-year return level for the three stations (37, 57, and 113 years of observations, respectively) showing bad agreement between the regional and local models. Black: fixed effects with simulated random effects. Gray: fixed effects only. White: fixed effects with estimated random effects.

the current regional model (in dotted lines with squares). The displayed stations have been selected to reflect a good fit by the regional model to the observed data, such that the regional predictive distribution gives a 1,000-year return level in agreement with the local model. It is seen that for these three selected sites the new regional model greatly improves on the regional model currently used in Norway. In addition, it is also clear that the current regional model stays within the 80% credible bands of the new regional model.

Figure 10 shows the same out-of-sample return value plots for three sites with a short ($n = 37$), medium ($n = 57$), and long ($n = 113$) data series but selected to reflect a bad agreement between the regional model and observed data. In all three sites the regional model overestimates the return levels compared to the local model, but in all three cases our regional model gives a better fit than the current regional model. Table 4 displays the average quantile scores for the out-of-sample predictions for the 27 selected stations, comparing the regional model to the local model and the regional model currently used by NVE. For all return periods, $T = 10, 50, 100$, the new regional model performs significantly better than the current regional model but not as well as the local model.

The total uncertainty of the model consists of two parts: the estimation uncertainty of the regression coefficients and remaining deviation from the linear model, captured by the random effects. The difference between the sources of uncertainty is illustrated in Figures 11 and 12, showing distributions of the 2-year (the median flood) and the 1,000-year return values over the fixed effects (gray), the fixed effect with simulated random effects (black), the out-of-sample prediction, and the fixed effects with estimated random effects (white), the in-sample prediction. The three distributions are shown for the same six stations selected in Figures 9 and 10. The in-sample distributions will be equivalent to the local analysis and hence have the smallest uncertainty, but it cannot be calculated for unobserved stations. The distribution over the fixed effects shows the uncertainty attributed to parameter uncertainty and tends to become smaller with longer data series, as clearly seen in the lower panels of Figure 12. As the distribution of the fixed effects often deviates from the in-sample distribution, the total uncertainty for a new, unobserved station is best captured by simulating the random effects in addition to the fixed effects, yielding the wider out-of-sample distribution.

10.1029/2017WR022460

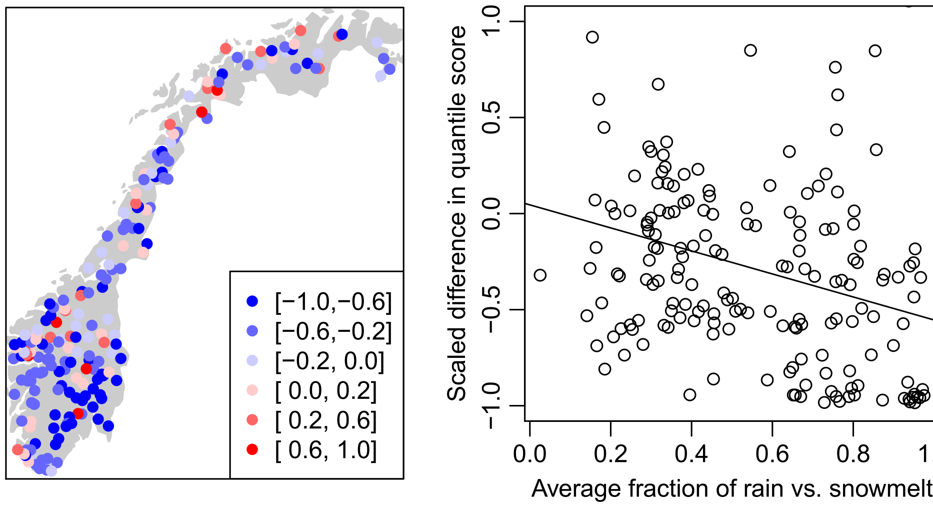

Figure 13. (left) Scaled difference in quantile score between the new regional model and the current NVE model. (right) The scaled difference in quantile score plotted against average proportion of rain versus snowmelt.

To illustrate how the regional model improves on the current NVE model, we calculate the difference in quantile scores for $T=100$ scaled by the score of the NVE model. This yields a relative change, where a value of −0.2 corresponds to a 20% improvement of the quantile score for the new regional model. The left panel of Figure 13 displays the spatial distribution of the scaled quantile score difference and shows that the regional model improves on the current NVE model to a larger degree in eastern part of Southern Norway. The quantile score difference correlates with a number of covariates, the percentage of agriculture, mountain and forest, and the mean temperatures in March, April, and May, and in particular the average fraction of rain. The right panel of Figure 13 shows the scaled difference plotted against the average fraction of rain, revealing a downward trend with a stronger relative improvement for stations with a high fraction of rain contributing to floods. These are all typical characteristics for stations situated in eastern part of Norway.

To explore the difference between the regional model and the local model, we calculate the same difference in quantile scores for $T=100$ but scaled by the score of the local model. As previously, a negative value indicates an improvement by the new regional model over the local model. The left panel of Figure 14 shows the spatial distribution of the scaled difference in quantile score and reveals no clear spatial pattern as to where the local model improves over the regional model. The local model is better than the regional model over the

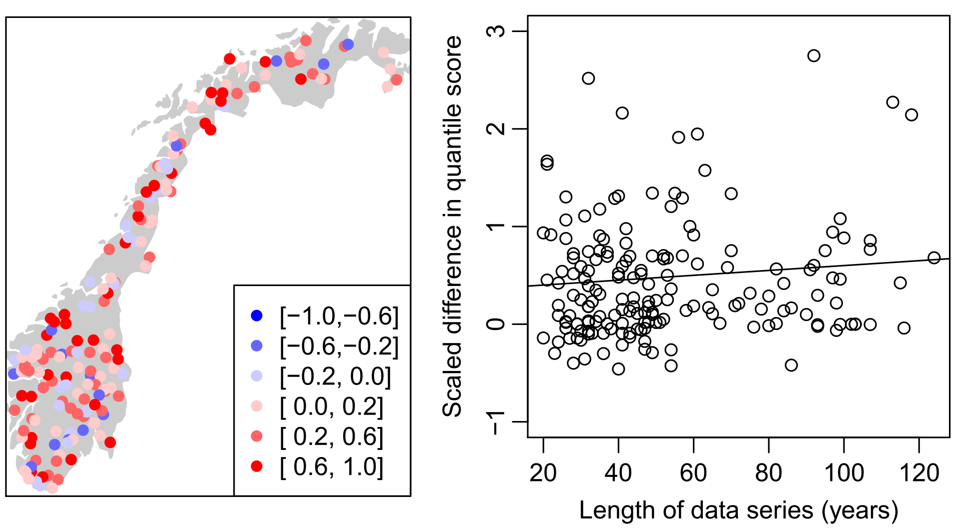

Figure 14. (left) Scaled difference in quantile score between the new regional model and the local model. (right) The scaled difference in quantile score plotted against length of the observed flood series.

whole of Norway. None of the recorded covariates are significantly correlated with quantile scores difference, suggesting that we have in fact found all information in available covariates, and further improvements of the regional model must require registration of previously unobserved and new covariates. However, the quantile score difference is weakly correlated with the length of the flood data series, as seen when plotted against the number of years, the right panel of Figure 14. The figure reveals a weak trend suggesting that the regional model improves on the local model for stations with short data series, with 20 to 40 years of observed floods.

7. Discussion

We have, in accordance with our main objective, developed a regional model for extreme flood estimation to be used when little or no data are available at a catchment site. The model is a Bayesian hierarchical framework with site-specific GEV parameters based on geographical, hydrological, and meteorological covariates, such that information is shared across catchment sites. We have identified 12 important predictors of this type to describe the local variation in the model parameters. By evaluating quantile scores for different return periods, it is seen that the Bayesian regional model gives better predictive performance than the current regional model used by NVE in Norway. Recent work by Yan and Moradkhani (2016) also supports that methods utilizing Bayesian hierarchical models and model averaging are beneficial for analyzing extreme flood data when emphasizing the quantification of extreme flood uncertainty.

While the new regional model significantly improves upon the current NVE model, a reliability analysis at 27 out-of-sample stations reveals that the model is, overall, somewhat overdispersive. A further analysis of the performance at individual stations indicates that the selected set of covariates may not be able to pick up the location-specific effects for all locations. Among the selected stations, the regional model tends to more often provide a higher estimates of flood sizes for high return periods than the local model. The Viksvatn station by the Gaular river in Sogn is one example of a catchment site where the regional model highly overestimates the flood distribution compared to the local model which is based on 113 years of observed annual floods, as seen in Figure 10. This is also the station with the worst reliability results in Figure 7. In other unrelated analyses, we have found that it appears that recorded level of rain at catchment sites in the Sogn area is consistently estimated too high. As the average rain in August is one of the most influential predictors of the overall flood level, any overestimation of this covariate will have an extreme impact on predicted return levels. This highlights the importance of good data quality for achieving reliable return level predictions. More generally, we find that poor reliability of the model at individual stations may potentially be explained by outliers in the relationships between important covariates and the GEV parameter values.

The covariate describing the average contribution of rain and snowmelt in the floods is found to be the best predictor of the shape parameter ξ_s. However, the exact calculation of this covariate requires some observed flood data. While the covariate may be estimated solely based on available meteorological data products such as the SeNorge data product, it remains to be assessed how much additional uncertainty is introduced. The covariates inflow and catchment area have relatively low posterior inclusion probabilities with the highest inclusion probability for each covariate 11% and 12%, respectively. However, these covariates represent the second and third most important covariates for the shape parameter ξ_s, and as shown in Figure 8, the posterior mean regression coefficients are different from zero across all the runs in the cross-validation study.

As there is some degree of correlation between the covariates, individual marginal posterior inclusion probabilities may be lower than they would be in a simpler model with fewer covariates. For this reason, we have excluded some covariates that, on their own, would explain significant portion of the spatial variability in the GEV parameters. This includes, for instance, average rain in other springs months besides April. In such cases, we have selected the covariate with the highest degree of correlation with the GEV parameters.

The general modeling approach proposed here models the spatial variation in the flood extremes using a latent variable approach. That is, we use covariate information and a stochastic process to model the spatial behavior of the marginal parameters which allows us to capture complex local variation in the return levels. As this local variation in the return levels is the object of interest for our application, we find this to be the appropriate approach in our setting in accordance with the conclusions of Davison et al. (2012). We have here employed a stochastic process without a spatial structure in contrast to, for example, some of the approaches described in Castellarin et al. (2012). The reason for this is twofold. First, an analysis of the posterior estimates of the random effects in the current model does not reveal any spatial dependencies based on the Euclidean distance between the station locations, indicating that such dependencies are already accounted for through the covariates if they exist in the data. Second, it is not clear that the Euclidean distance is the most appropriate distance measure for our data due to the complex orography of Norway. Rather, distances based on river network structures might be more appropriate (Asadi et al., 2015). However, we consider such approaches outside of the scope of the current study.

Appendix A: Hierarchical Model With a log Link on the Precision

This section discusses an extension to the MCMC algorithm described in Dyrrdal et al. (2015) where the regression equation for the precision parameter is defined with $\log\kappa$ as the response variable. In general, assume that we want to update a parameter ν in our model, where ν is the current value. We draw a new value ν' from a proposal distribution $pr(\nu'|\nu,\cdot)$ and accept the proposal with probability $\min\{r,1\}$ where

$$r = \frac{pr(\boldsymbol{y}|\nu',\cdot)pr(\nu'|\cdot)pr(\nu|\nu',\cdot)}{pr(\boldsymbol{y}|\nu,\cdot)pr(\nu|\cdot)pr(\nu'|\nu,\cdot)}.$$

Here $pr(\boldsymbol{y}|\nu,\cdot)$ denotes the likelihood of the full data set $\boldsymbol{y}$ which depends on ν and potentially other parameters which are kept fixed throughout, and $pr(\nu|\cdot)$ is the prior distribution of ν which similarly might depend on the other parts of the model. Given the complexity of the model, it is vital to design efficient proposal distributions which return good proposals and are robust in that they do not require fine-tuning for each individual data set.

For designing the proposal distribution, we employ a Gaussian approximation (Rue & Held, 2005, Chapter 4.4). Assume that the posterior distribution of the parameter ν' is written on the form

$$pr(\nu'|\cdot) \propto \exp\left(f(\nu')\right),$$

for some function f. A quadratic Taylor expansion of the log-posterior $f(\nu')$ around the value ν gives

$$\begin{aligned} f(\nu') &\approx f(\nu) + f'(\nu)(\nu'-\nu) + \frac{1}{2}f''(\nu)(\nu'-\nu)^2 \\ &= a + b\nu' - \frac{1}{2}c(\nu')^2, \end{aligned}$$

where $b = f'(\nu) - f''(\nu)\nu$ and $c = -f''(\nu)$. The posterior distribution $pr(\nu'|\cdot)$ may now be approximated by

$$\widetilde{pr}(\nu'|\cdot) \propto \exp\left(-\frac{1}{2}c(\nu')^2 + b\nu'\right),$$

the density of the Gaussian distribution $\mathcal{N}(b/c, c^{-1})$. We thus choose $\mathcal{N}(b/c, c^{-1})$ as our proposal distribution. This implies that in order to update the model in Dyrrdal et al. (2015) to include a logarithmic link for the precision κ, we need to calculate the first two derivatives of the log likelihood function $\log pr(y_{ts}|\tau_s^\kappa,\cdot)$ with respect to the random effect τ_s^κ.

A1. The Case $\xi \neq 0$

We have $\kappa_s = \exp(\eta_s)$, where $\eta_s = \boldsymbol{x}_s^\top\theta^\kappa + \tau_s^\kappa$ and $\tau_s^\kappa \sim \mathcal{N}(0, \alpha_\eta^{-1})$. Now, fix $\hat{\eta}_s = \boldsymbol{x}_s^\top\theta^\kappa$ and set $\epsilon_{ts} = y_{ts} - \mu_{ts}$. We then have

$$\log pr(y_{ts}|\tau_s^\kappa,\cdot) = \hat{\eta}_s + \tau_s^\kappa - \frac{\xi_s+1}{\xi_s}\log(h) - h^{-\xi_s^{-1}},$$

where

$$h = 1 + \xi_s\epsilon_{ts}\exp(\hat{\eta}_s + \tau_s^\kappa).$$

It follows that

$$\frac{\partial h}{\partial \tau_s^\kappa} = h - 1,$$

$$\frac{\partial}{\partial \tau_s^\kappa}\log pr(y_{ts}|\tau_s^\kappa,\cdot) = 1 - \frac{\xi_s+1}{\xi_s}\frac{h-1}{h} + \xi_s^{-1}h^{-\xi_s^{-1}} - \xi_s^{-1}h^{-\xi_s^{-1}-1}$$

and

$$\frac{\partial^2}{(\partial\tau_s^\kappa)^2}\log pr(y_{ts}|\tau_s^\kappa,\cdot) = -\frac{\xi_s+1}{\xi_s}\frac{1}{h^2} - \frac{h^{-\xi_s^{-1}}}{\xi_s^2} + \frac{\xi_s+2}{\xi_s^2}h^{-\xi_s^{-1}-1} - \frac{\xi_s+1}{\xi_s^2}h^{-\xi_s^{-1}-2}.$$

A2. The Case $\xi = 0$

Using the same notation as above, we have

$$\log pr(y_{ts}|\tau_s^\kappa, \cdot) = \hat{\eta}_s + \tau_s^\kappa + \log(h) - h,$$

where

$$h = \exp\left\{-\exp(\hat{\eta}_s + \tau_s^\kappa)\epsilon_{ts}\right\}.$$

Note that

$$\frac{\partial h}{\partial \tau_s^\kappa} = h\log(h).$$

Thus,

$$\frac{\partial}{\partial \tau_s^\kappa}\log pr(y_{ts}|\tau_s^\kappa, \cdot) = 1 + \log(h) - h\log(h),$$
$$\frac{\partial^2}{(\partial \tau_s^\kappa)^2}\log pr(y_{ts}|\tau_s^\kappa, \cdot) = \log(h) - h\log(h)^2 - h\log(h).$$

Acknowledgments

This work was jointly funded by the Research Council of Norway and Energy Norway through grant 235710 (FlomQ). The flood and hydrological data were extracted from the National Hydrological Database at the Norwegian Water Resources and Energy Directorate and are listed in the supporting information. Climatological data, derived from the data product SeNorge, are available at www.senorge.com. The source code for estimating the flood frequency distributions is implemented in the statistical programming language R (http://www.R-project.org) as a part of the R package `SpatGEVBMA` which is available on GitHub at http://github.com/NorskRegnesentral/SpatGEVBMA. All data, `R` code, and use of `R` packages, with working scripts, are available at the GitHub webpage of the Norwegian Computing Center (https://github.com/NorskRegnesentral/BRFFA).

References

Asadi, P., Davison, A. C., & Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, *9*(4), 2023–2050.

Ball, J., Babister, M., Nathan, R., Weeks, W., Weinmann, E., Retallick, M., & Testoni, I. (2016). *Australian rainfall and runoff: A guide to flood estimation*. Canberra: Commonwealth of Australia.

Bentzien, S., & Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1924–1934.

Castellarin, A., Kohnová, S., Gaál, L., Fleig, A., Salinas, J. L., Toumazis, A., et al. (2012). *Review of applied statistical methods for flood frequency analysis in Europe*. Lancaster: NERC/Centre for Ecology & Hydrology.

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

Cooley, D., Nychka, D., & Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, *102*(479), 824–840.

Dalrymple, D. (1960). Flood-frequency analysis, Geological survey water-supply (Tech. Rep. Paper 1543-A). Washington, DC: United States Government Printing Office. https://pubs.usgs.gov/wsp/1543a/report.pdf

Davison, A. C., Padoan, S. A., & Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, *27*(2), 161–186.

Dawid, A. P. (1984). Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series A*, *147*, 278–292.

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L., & Stordal, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, *26*, 89–106.

Engeland, K., Schlichting, L., Randen, F., Nordtun, K. S., Reitan, T., Wang, T., et al. (2016). Utvalg og kvalitetssikring av flomdata for flomfrekvensanalyser (Tech. Rep. 85-2016). Oslo: NVE.

Fawcett, L., & Walshaw, D. (2016). Sea-surge and wind speed extremes: Optimal estimation strategies for planners and engineers. *Stochastic Environmental Research and Risk Assessment*, *30*(2), 463–480.

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

Friederichs, P., & Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, *135*(6), 2365–2378.

Glad, P. A., Reitan, T., & Stenius, S. (2014). Regionalt formelverk for flomberegning i små nedbørfelt (Tech. Rep. 62-2014). Oslo: NVE.

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

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional Frequency Analysis*. UK: Cambridge University Press.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, *81*, 158–171.

Kochanek, K., Renard, B., Arnaud, P., Aubert, Y., Lang, M., Cipriani, T., & Sauquet, E. (2014). A data-based comparison of flood frequency analysis methods used in France. *Natural Hazards and Earth System Sciences*, *14*, 295–308.

Lenkoski, A. (2014). SpatGEVBMA: Hierarchical spatial generalized extreme value (GEV) modeling with Bayesian Model Averaging (BMA). R package version 1.0.

Midttømme, G., Pettersson, L., Holmqvist, E., Nøtsund, Ø., Hisdal, H., & Sivertsgård, R. (2011). *Guidelines for Flood Estimation (In Norwegian)*. Oslo: NVE. retningslinjer nr. 4/2011.

Parent, E., & Bernier, J. (2002). Bayesian POT modelling for historical data. *Journal of Hydrology*, *274*, 95–108.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reis, D., & Stedinger, J. (2005). Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology*, *313*(1-2), 97–116.

Renard, B., Kochanek, K., Lang, M., Garavaglia, F., Paquet, E., Neppel, L., et al. (2013). Data-based comparison of frequency analysis methods: A general framework. *Water Resources Research*, *49*, 825–843. https://doi.org/10.1002/wrcr.20087

Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: CRC Press.

Sælthun, N., Tveito, O., Bønsnes, T., & Roald, L. (1997). Regional flomfrekvensanalyse for norske vassdrag (Tech. Rep. 14-97). Oslo: NVE.

Siotani, T., Fujikoshi, Y., & Hayakawa, T. (1985). *Modern Multivariate Statistical Analysis, A Graduate Course and Handbook*. New York: American Sciences Press.

Steinbakk, G. H., Thorarinsdottir, T. L., Reitan, T., Schlichting, L., Hølleland, S., & Engeland, K. (2016). Propagation of rating curve uncertainty in design flood estimation. *Water Resources Research*, *52*, 6897–6915. https://doi.org/10.1002/2015WR018516

Stephenson, D. B., Coelho, C. A. S., & Jolliffe, I. T. (2008). Two extra components in the brier score decomposition. *Weather and Forecasting*, *23*, 752–757.

Wilson, D., Hisdal, H., & Lawrence, D. (2010). Has streamflow changed in the Nordic countries? – Recent trends and comparisons to hydrological projections. *Journal of Hydrology*, *394*, 334–346.

Yan, H., & Moradkhani, H. (2016). Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling. *Natural Hazards*, *91*, 203–225. https://doi.org/10.1007/s11069-015-2070-6