Nat. Hazards Earth Syst. Sci., 15, 2653–2667, 2015
www.nat-hazards-earth-syst-sci.net/15/2653/2015/
doi:10.5194/nhess-15-2653-2015

Natural Hazards and Earth System Sciences
Open Access

Evaluation of a compound distribution based on weather pattern subsampling for extreme rainfall in Norway

J. Blanchet[1,2], **J. Touati**[1,2], **D. Lawrence**[3], **F. Garavaglia**[4], **and E. Paquet**[4]

[1]Univ. Grenoble Alpes, LTHE, 38000 Grenoble, France
[2]CNRS, LTHE, 38000 Grenoble, France
[3]Norwegian Water Resources and Energy Directorate (NVE), P.O. Box 5091, Majorstua, Oslo, Norway
[4]EDF – DTG, 21 Avenue de l'Europe, BP 41, 38040 Grenoble CEDEX 9, France

Correspondence to: J. Blanchet (juliette.blanchet@ujf-grenoble.fr)

Received: 16 April 2015 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: 3 June 2015
Revised: 13 October 2015 – Accepted: 13 November 2015 – Published: 10 December 2015

Abstract. Simulation methods for design flood analyses require estimates of extreme precipitation for simulating maximum discharges. This article evaluates the multi-exponential weather pattern (MEWP) model, a compound model based on weather pattern classification, seasonal splitting and exponential distributions, for its suitability for use in Norway. The MEWP model is the probabilistic rainfall model used in the SCHADEX method for extreme flood estimation. Regional scores of evaluation are used in a split sample framework to compare the MEWP distribution with more general heavy-tailed distributions, in this case the Multi Generalized Pareto Weather Pattern (MGPWP) distribution. The analysis shows the clear benefit obtained from seasonal and weather pattern-based subsampling for extreme value estimation. The MEWP distribution is found to have an overall better performance as compared with the MGPWP, which tends to overfit the data and lacks robustness. Finally, we take advantage of the split sample framework to present evidence for an increase in extreme rainfall in the southwestern part of Norway during the period 1979–2009, relative to 1948–1978.

1 Introduction

Flood estimation is important for design and safety assessments, flood risk management and spatial planning. It aims to assess the probability of occurrence of large events, e.g., discharges with return periods of 100 to 10 000 years. Estimation of events with such low probability is particularly arduous. It can only be based on a few data points representing the most extreme events in a time series of a limited length. Thus extrapolation to long return periods is usually needed. In dam safety analyses, for example, return period estimations of 10^3 to 10^4 years are often used (Paquet et al., 2013). Methods for deriving such estimations can be classified into two main groups: statistical flood frequency analysis and precipitation–runoff modeling. Statistical flood frequency analysis is based on the analysis of an observed streamflow record for which the return periods of the highest events are modeled using extreme value theory, and magnitudes with longer return periods are estimated using the fitted statistical model. A drawback of this method is that it relies on local or regional streamflow data and is likely to be very sensitive to the density of observations (for the regional case) and to the type of distribution chosen (Klemes, 2000a, b). Furthermore, heavy rainfall is a major factor driving the occurrence of flooding, even in areas where snowmelt also plays a significant role, such as in Norway. Rainfall series are generally more abundant, often have longer periods of record, and they usually show stronger regional consistency. This observation is one of the main motivations of the GRADEX method (Guillot, 1993) which uses the distribution of rainfall to extrapolate the distribution of discharge. This has further led to the development of rainfall–runoff simulation methods for extreme flood estimation. The idea is to extend the database of streamflow by converting rainfall into surface runoff using a model of the catchment response. Input rainfall may be either observed or synthetic events with an estimated probability of occurrence (*event-based method*) or, either historical or synthetic rainfall records for gener-

Published by Copernicus Publications on behalf of the European Geosciences Union.

ating a continuous streamflow series (*continuous simulation approach*).

In Norway, a simple event-based rainfall–runoff model, PQRUT, has been used since the 1980s as a simulation method for dam safety analyses for which the magnitude of low frequency events (e.g., 500-, 1000-year peak inflow) and the probable maximum flood are required. Recently, a semi-continuous model, SCHADEX (Paquet et al., 2013) has been tested as an alternative approach for obtaining such estimates. SCHADEX has been developed and applied in France by Electricité de France (EDF) for dam spillway design since 2006. It has also recently been applied in different regions of the world (in France, Austria, Canada and Norway) (e.g., Brigode et al., 2014) and has been more extensively evaluated for three catchments in Norway in Lawrence et al. (2014). Of particular interest in Norway is the need for a method which takes the combined probability of extreme rainfall and snowmelt into account, for which SCHADEX is well suited in comparison with event-based approaches. It is expected that the SCHADEX method should give results more similar to those obtained with statistical flood frequency analysis based on observed discharge series, and this was found in two of the three catchments considered by Lawrence et al. (2014). However, a global evaluation of the SCHADEX method covering the range of conditions found in Norway has yet to be achieved and is a necessary precursor to the wider implementation of the method in standard practise. This article aims to make the first step towards such an evaluation. More specifically, we evaluate the rainfall probabilistic component of SCHADEX: the so-called multi-exponential weather pattern (MEWP) distribution (Garavaglia et al., 2010), a compound distribution based on season and weather pattern subsampling, for the whole of Norway. This approach is in contrast with the recent analyses of extreme precipitation in Norway based on annual maximum series and the application of a generalized extreme value distribution undertaken by Dyrrdal et al. (2014). In our work we analyze over threshold values for rainfall, rather than using a block maxima approach. Our goal is to evaluate the performance of MEWP at the national scale and to decide whether it should be preferred to simpler, and perhaps more classical, seasonal and nonseasonal distributions, or, further, whether its generalization towards heavy-tailed distributions should be considered. A brief analysis of trends in extreme precipitation is also performed based on the split samples used in the evaluation.

2 Data

Daily data from 368 precipitation stations in Norway were extracted from the European Climate Assessment and Dataset (ECA&D), a database of daily meteorological stations across Europe. From these 368 stations, 192 stations with at least 50 years of record with less than 10 % missing data per year over the period 1948–2009 were selected for further analyses. Years with more than 10 % missing data are entirely replaced by 'NA', representing missing values. Figure 1 shows the location and altitude of the 192 stations. Station altitude ranges from sea level to approximately 1000 m a.s.l., i.e., none of the stations lie at the higher altitudes in the mountainous regions. All the stations above 500 m a.s.l., however, are found in the central southern inland region adjacent to zones of higher altitude. The network is denser in southern Norway, particularly along the coast, reflecting the higher population densities in this zone. This implies that southern Norway will have more weight in the model evaluation but we view this as preferable to deleting a number of stations to create a more spatially uniform network density. The mean number of observed years is 56 (maximum 62, minimum 50).

As already stated in Sect. 1, the main topic of this study is the evaluation of MEWP, the rainfall probabilistic model used in SCHADEX. SCHADEX aims to describe the distribution of floods by a stochastic simulation process which combines heavy rainfall events and catchment saturation states, including simulated snowmelt. In SCHADEX, heavy rainfall events are considered as 3-day centered precipitation events, being composed of a central rainfall and two adjacent rainfalls which are lower than the central one (Paquet et al., 2013). The value for central rainfall is simulated using a fitted MEWP distribution for the extreme rainfall (Garavaglia et al., 2011), and the 2 adjacent days are simulated conditionally, using contingency tables to account for the dependence of the magnitude of the rainfall on the day before and after the peak rainfall. Given that MEWP is a probabilistic model for heavy "central" rainfall, rather than for all daily rainfall values, a pre-processing of the data was required to select the central rainfall values exceeding the precipitation received on both the preceding and following days by 1 mm or more at each station. By doing this we obviously reduce the number of data available for analysis. In Norway about one-quarter of the days of record represent central rainfall values, and this is, on average, about one-half of the days with precipitation. However one advantage of this pre-processing is that central rainfall values at a given location can be expected to be independent since they are always separated by at least 1 day. For extreme values, this independence can be quantitatively assessed by computing the so-called extremal coefficients (Coles, 2001; Ferro and Segers, 2003) for the daily and central samples and comparing their respective values for each station. Extremal coefficients lie between 0 and 1 and the closer to 1, the less dependent the extremes. The inverse of the extremal coefficient can be more easily interpreted as the mean size of clusters at extreme level, i.e., roughly speaking, the mean number of consecutive values that are extreme. Using the estimation method of Ferro and Segers (2003) with a threshold equal to the 90 % quantile of daily rainfall, we find that extremal coefficients for daily rainfall are about 0.6, whereas those for central rainfall are about 0.8 (representing

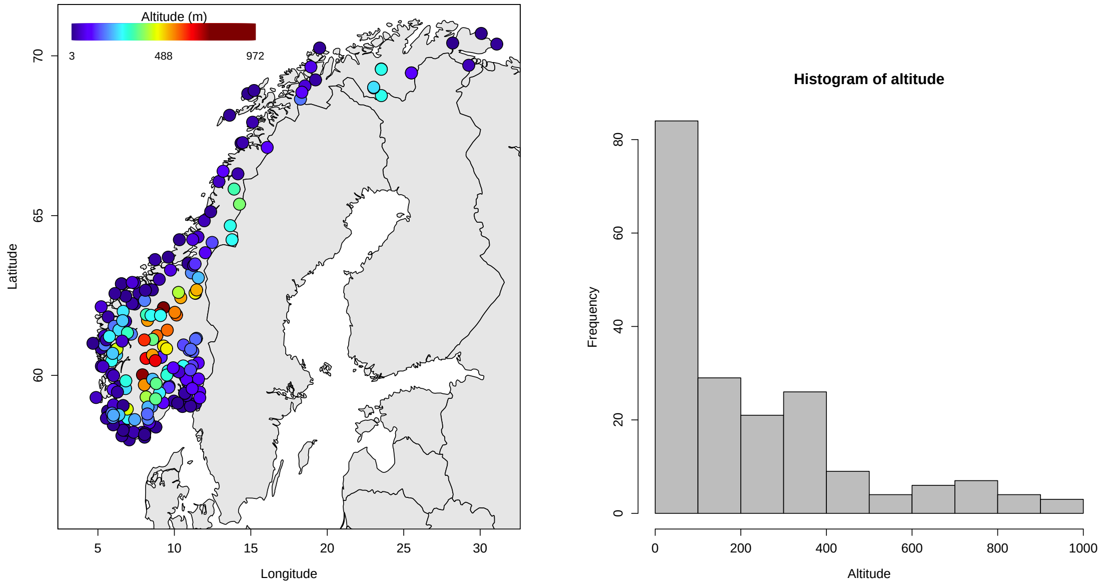

Figure 1. Left: location and altitude (m a.s.l.) of the stations. Right: histogram of altitude (m a.s.l.).

a mean cluster size of about 1.25 days). The central rainfall values can therefore be considered to be close to the case of complete independence.

3 Model and method

3.1 Modeling

3.1.1 Exponential and GPD models

Let X be the random variable of central rainfall at some location in Norway. We are interested in the distribution of extreme values, i.e., of $\Pr(X \leq x)$ when x is large. Let us consider a (high) level α and write q_α the α-quantile of X, i.e., such that $\alpha = \Pr(X \leq q_\alpha)$. Then, for all x exceeding q_α, we have the decomposition

$$F(x) \equiv \Pr(X \leq x) = \alpha + (1-\alpha)\Pr(X \leq x | X \geq q_\alpha). \qquad (1)$$

Extreme value theory (EVT) ensures that if the central rainfall values are independent and identically distributed and for large enough α, $\Pr(X \leq x | X \geq q_\alpha)$ can be approximated by the distribution

$$G(x;\sigma_\alpha,\xi) = \begin{cases} 1-\left(1+\frac{\xi(x-q_\alpha)}{\sigma_\alpha}\right)^{-1/\xi}, & \text{if } \xi \neq 0, \qquad (2) \\ 1-\exp\left(-\frac{(x-q_\alpha)}{\sigma_\alpha}\right), & \text{if } \xi = 0, \qquad (3) \end{cases}$$

for all $x \geq q_\alpha$, provided in Eq. (2) that $x \leq q_\alpha - \sigma_\alpha/\xi$ if $\xi < 0$. Parameter ξ in Eq. (2) is independent of α; this is the shape parameter which models the heaviness of the tail of the distribution. Parameter $\sigma_\alpha > 0$ in Eqs. (2) and (3) depends upon α and is called the scale parameter. Equations (2) and (3) imply that excesses $(X - q_\alpha | X \geq q_\alpha)$ follow the generalized Pareto distribution (GPD) in Eq. (2) and the exponential distribution (EXP) with rate $1/\sigma_\alpha$ in Eq. (3). Models (Eqs. 2 and 3) have been widely used worldwide for modeling rainfall extremes. A good review is provided in the introduction of Serinaldi and Kilsby (2014). Equations (2) and (3) combined with Eq. (1) give the approximation of the distribution of X for all $x \geq q_\alpha$:

$$F(x) \approx \alpha + (1-\alpha)\, G(x;\sigma_\alpha,\xi), \qquad (4)$$

where $\alpha = \Pr(X \leq q_\alpha)$.

3.1.2 MEWP and MGPWP models

In the previous section, we implicitly assumed that central rainfall, X, is identically distributed throughout the year. This assumption may be questioned. Indeed, different climatological processes trigger precipitation, leading to the occurrence of rainfall of different natures and intensities (e.g., convective vs. stratiform precipitation). Furthermore, rainfall occurrence and intensities often vary with season, reflecting both variations in temperature and in storm tracks, for example. For this reason, Garavaglia et al. (2010) proposed the use of subsampling based on seasons and weather patterns (WP). Each day of the record period is assigned to a WP. If S seasons and K WP are considered, then days are classified into $S \times K$ subclasses. The law of total probability gives, for all x,

$$F(x) = \sum_{s=1}^{S}\sum_{k=1}^{K} \Pr(X \leq x | \text{season} = s, \text{WP} = k)\, p_{s,k}, \qquad (5)$$

where $p_{s,k}$ is the probability that a given day is in season s and in WP k (thus $\sum_s \sum_k p_{s,k} = 1$). The central rainfall values occurring in season s and WP k can be assumed to be identically distributed (Garavaglia et al., 2010). Thus the extreme value theory described in Sect. 3.1.1 can be applied to $F_{s,k}(x) = \Pr(X \le x | \text{season} = s, \text{WP} = k)$. Let us consider a high level α (taken for simplicity constant for all $F_{s,k}$) and $q_{\alpha,s,k}$, the α quantile of $F_{s,k}$. Application of Eq. (4) to $F_{s,k}$ gives the approximation for $x \ge q_{\alpha,s,k}$,

$$F_{s,k}(x) \approx \alpha + (1-\alpha) G(x; \sigma_{\alpha,s,k}, \xi_{s,k}), \tag{6}$$

where $G(x; \sigma_{\alpha,s,k}, \xi_{s,k})$ is given by Eqs. (2) and (3), where q_α, σ_α and ξ are respectively replaced by $q_{\alpha,s,k}$, $\sigma_{\alpha,s,k}$ and $\xi_{s,k}$. Thus, Eqs. (5) and (6) give, for all $x \ge q_\alpha^+ = \max_{s,k} q_{\alpha,s,k}$, the approximation of the distribution of X:

$$F(x) \approx \alpha + (1-\alpha) \sum_{s=1}^{S} \sum_{k=1}^{K} G(x; \sigma_{\alpha,s,k}, \xi_{s,k})\, p_{s,k}. \tag{7}$$

MEWP and MGPWP (Multi Generalized Pareto Weather Pattern) models are both defined by Eq. (7) with different choices of $\xi_{s,k}$ (Garavaglia et al., 2010, 2011): in MEWP all $\xi_{s,k}$ are set to 0 – in which case G is the EXP distribution – while in MGPWP, $\xi_{s,k}$ is free to vary in the positive range. We exclude cases $\xi_{s,k} < 0$ because they give bounded GPD distributions with an upper bound at $q_{\alpha,s,k} - \sigma_{\alpha,s,k}/\xi_{s,k}$, which is usually unrealistically low for rainfall. Using the GPD with $\xi_{s,k} > 0$ in Eq. (7) allows models with heavier tails than with the EXP distribution, which is light-tailed. Theoretically, other heavy-tailed distribution could be used for G in Eq. (7) but the GPD is justified by EVT and it provides a natural generalization of MEWP by allowing the $\xi_{s,t}$ to vary freely (within the positive range). Both models, MEWP and MGPWP, will be evaluated on the Norwegian data. To keep track of the level α and of the fact that S seasons and K WP are used in Eq. (7), we will respectively write these two models as MEWP(α, S, K) and MGPWP(α, S, K). Likewise, we write EXP(α) and GPD(α) to represent the basic cases of Eq. (4) when neither season nor WP are considered, corresponding to cases MEWP(α, 1, 1) and MGPWP(α, 1, 1).

3.2 Model estimation

Use of the EXP, GPD, MEWP and MGPWP models requires the choice of high enough thresholds such that EVT can be applied. Selection of an adequate threshold gives rise to a bias-variance tradeoff: the higher the threshold, the better the approximation of the tail of F (smaller bias), but at the same time, the higher the variance of the estimated parameters because a smaller number of exceedances are available. Graphical tools for threshold selection, such as mean residual life plots (Coles, 2001), are usually difficult to interpret in practice. Therefore, the common practice is to fix a high enough level α and to set thresholds $q_{\alpha,s,k}$ to the empirical α quantile of rainfall occurring in season s and WP k.

Given α (and therefore q_α), the parameters that must be estimated for the EXP and GPD models (Eq. 4) are those of G in Eqs. (2) and (3). Estimation is made by the method of L-moments (Hosking, 1990):

$$\hat{\xi} = (\lambda_1 - q_\alpha)/\lambda_2 - 2, \;\; \hat{\sigma}_\alpha = (1-\hat{\xi})(\lambda_1 - q_\alpha), \text{for GPD}(\alpha),$$
$$\hat{\sigma}_\alpha = \lambda_1 - q_\alpha, \text{for EXP}(\alpha),$$

where λ_1 and λ_2 are the sample L-moments of order 1 and 2 for the central rainfall exceeding q_α, which are independent; see Sect. 2. In the GPD case, if $\hat{\xi} < 0$, then $\hat{\xi} = 0$ is imposed (i.e., the EXP distribution) to exclude bounded distributions. It should be noted that the choice of the L-moments method only affects the GPD case since for the EXP case, the commonly used L-moments, moments and maximum likelihood estimators coincide. For the GPD case, a separate analysis (not shown) reveals that the choice of the estimation method does not actually affect the regional evaluation very much because slight differences in estimation that occur at the local scale are smoothed out at the regional scale.

Parameters $\xi_{s,k}$ and $\sigma_{\alpha,s,k}$ in G of Eq. (7) for MEWP and MGPWP are estimated likewise by the L-moments method, using the observed central rainfall of season s and WP k exceeding $q_{\alpha,s,k}$. Probability $p_{s,k}$ is estimated as the empirical proportion of days in season s and WP k. Estimation of F is then obtained for all $x > q_\alpha^+$ with Eq. (7).

3.3 Computation of return levels

The T-year return level r_T is the level expected to be exceeded on average once every T years. It satisfies the relationship $F(r_T) = 1 - 1/(T\zeta)$, where ζ is the mean number of central rainfall events per year. When F is EXP(α) or GPD(α), estimation of r_T is obtained explicitly as

$$\hat{r}_T = \begin{cases} q_\alpha + \hat{\sigma}_\alpha \log\{(1-\alpha)T\zeta\} & \text{for EXP}(\alpha) \\ q_\alpha + \hat{\sigma}_\alpha \{[(1-\alpha)T\zeta]^{\hat{\xi}} - 1\}/\hat{\xi} & \text{for GPD}(\alpha) \end{cases}, \tag{8}$$

where $\hat{\sigma}_\alpha$ and $\hat{\xi}$ are the parameter estimates of F of Sect. 3.2. For the MEWP and MGPWP models, there is not an explicit formulation for $\hat{r}_T$ and it is obtained numerically by solving $F(\hat{r}_T) = 1 - 1/(T\zeta)$ in Eq. (7). Equation 8 shows that in GPD(α) model, $\hat{r}_T$ is mainly influenced by the value of $\hat{\xi}$. For the MGPWP model, practice shows that for reasonable to large T (typically $T > 50$ years), $\hat{r}_T$ is mainly influenced by the largest $\hat{\xi}_{s,k}$.

3.4 Model evaluation

The goal of this evaluation is to assess which model performs better at the regional scale, i.e., for a set of N stations taken as a whole, rather than individually. We follow the split sample evaluation proposed in Garavaglia et al. (2011) and Renard et al. (2013). We divide the data for each station i into two subsamples, $C_i^{(1)}$ and $C_i^{(2)}$, and fit a given competing model on each of the subsamples, giving two estimated distributions $\hat{F}_i^{(1)}$, estimated on $C_i^{(1)}$, and $\hat{F}_i^{(2)}$, estimated on $C_i^{(2)}$. Our goal is to test the consistency between validation data and predictions of the estimates, and the accuracy and stability of the estimates when calibration data change. For this, three scores are computed, assessing respectively stability (SPAN) and reliability (AREA(FF) and AREA(N_T)) of the fits. These scores were proposed and used in Garavaglia et al. (2011) and Renard et al. (2013).

The SPAN criterion evaluates the stability of the return level estimation, when using data for each of the two subsamples. More precisely, for a given return period T and station i,

$$\mathrm{SPAN}_{T,i} = \frac{|\hat{r}^{(1)}_{T,i} - \hat{r}^{(2)}_{T,i}|}{1/2\{\hat{r}^{(1)}_{T,i} + \hat{r}^{(2)}_{T,i}\}},$$

where $\hat{r}^{(1)}_{T,i}$, e.g., is the T-year return level for the distribution F (see Sect. 3.3) estimated on subsample $\mathrm{C}^{(2)}_i$ of station i. $\mathrm{SPAN}_{T,i}$ is the relative absolute difference in T-year return levels estimated on the two subsamples. It ranges between 0 and 2; the closer to 0, the more stable the estimations for station i. For the set of N stations, we obtain a vector of SPAN_T of length N with a distribution which should remain reasonably close to zero. A rough summary of this information is obtained by computing the mean of the N values of $\mathrm{SPAN}_{T,i}$, $i = 1, \ldots, N$:

$$\mathrm{MEAN}(\mathrm{SPAN}_T) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{SPAN}_{T,i}. \tag{9}$$

For competing models, the closer the mean is to 0, the more stable the model is.

The FF criterion is used to estimate the reliability in estimating the probability of occurrence of the maximum of independent variables. Let $(X_1, \ldots, X_n)$ be a set of n independent and identically distributed rainfall values with distribution F and $Z = \max_{j=1}^{n} X_j$. Then $\Pr(Z \le x) = \{\Pr(X \le x)\}^n = \{F(x)\}^n$ and, thus, the distribution of Z is F^n. Therefore $FF = \{F(Z)\}^n$ follows the uniform distribution on $(0, 1)$. Now write $\hat{F}_{1,i}$ and $\hat{F}_{2,i}$, where the estimation of F for station i is obtained respectively for subsamples $\mathrm{C}^{(1)}_i$ and $\mathrm{C}^{(2)}_i$. If $\hat{F}_{1,i}$ and $\hat{F}_{2,i}$ are good estimations of F, then $FF^{(1)}_i = \{\hat{F}^{(1)}_i(Z)\}^n$ and $FF^{(2)}_i = \{\hat{F}^{(2)}_i(Z)\}^n$ should approximately follow the uniform distribution, Unif(0, 1). Now let $n^{(1)}_i$ (resp. $n^{(2)}_i$) be the number of central (thus independent) rainfall values in subsamples $\mathrm{C}^{(1)}_i$ (resp. $\mathrm{C}^{(2)}_i$) and $z^{(1)}_i$ (resp. $z^{(2)}_i$) the corresponding observed maximum, then

$$\begin{aligned} ff^{(12)}_i &= [\hat{F}^{(2)}_i(z^{(1)}_i)]^{n^{(1)}_i} \\ ff^{(21)}_i &= [\hat{F}^{(1)}_i(z^{(2)}_i)]^{n^{(2)}_i} \end{aligned}$$

should both be realizations of the uniform distribution. For the set of N stations, this gives two uniform samples $ff^{(12)}$ and $ff^{(21)}$ of size N each. Hypothesis testing for assessing if the uniform assumption is valid is challenging because the ff_i are not independent from site to site, due to the spatial dependence between data. Thus Renard et al. (2013) proposed to base comparison on the graphical analysis of cumulative distribution functions (CDFs), by inspecting how much the CDF of the ff diverge from the 1 : 1 line, corresponding to the CDF of uniform variates on $(0, 1)$. A quantitative assessment of this divergence is provided by computing the area between both CDFs. However, we find such evaluation confusing because the value of the area depends on where, between 0 and 1, the divergence is located. An illustration of this is given in Fig. 2 for three simulated series of length 200 (which is about the number of stations). In case 0, the ff are all drawn from Unif(0, 1) (reference case). In cases 1 and 2, 80 % of the ff are drawn from Unif(0, 1) and 20 % are drawn from Unif(0, 0.1) in case 1 and from Unif(0.5, 0.6) in case 2. Departure of ff from the uniform case is sometimes not easy to interpret. However case 1 corresponds usually to a tendency towards an overestimation of the largest observation, while case 2 corresponds to a tendency towards overfitting the largest observation. In the CDF plot (upper left), the area value is as expected the lowest for case 0. However case 2 gives surprisingly also a very good score, whereas that of case 1 is 3 times as large. Therefore these criteria would falsely indicate a better performance (i.e., smaller area value) of case 2 (overfitting) as compared to case 1 (overestimation), although they both contain 20 % data diverging from the uniform on $(0, 1)$. As an alternative, we prefer to base evaluation on divergence between densities rather than CDFs. A reasonable estimate of this latter is obtained by computing the empirical histogram of the ff with 10 equal bins between 0 and 1, and comparing it with the uniform density between 0 and 1 (which equals 1). For a more quantitative assessment, we compute the area between both densities as follows:

$$\mathrm{AREA}(FF) = \frac{1}{18}\sum_{\ell=1}^{10} \left| 10\frac{\mathrm{Card}\{ff_i \in \mathrm{bin}(\ell), i = 1, \ldots, N\}}{N} - 1 \right|, \tag{10}$$

where Card{...} denotes the cardinality of the set. The term inside the absolute value in Eq. (10) is the difference between densities in the ℓth bin. The division by 18 forces the score to lie in the range $(0, 1)$ with lower values indicating better fits (the worst case being all values lying in the same bin). Illustration of this computation is shown in Fig. 2 on the aforementioned simulated data (upper right and lower panels). The score for case 0 is again the lowest, however the value is larger than when comparing CDFs due to the discretization into bins. As expected, the criteria now give similar scores for cases 1 and 2, unlike the method based on CDFs. This leads us to base comparison on the new AREA score (Eq. 10), giving preference to lower scores but keeping in mind that a score of 0.1 is already a good score since this is the mean AREA value we obtain when simulating uniforms on $(0, 1)$. Returning to ff values of cross-validation, $ff^{(12)}$ and

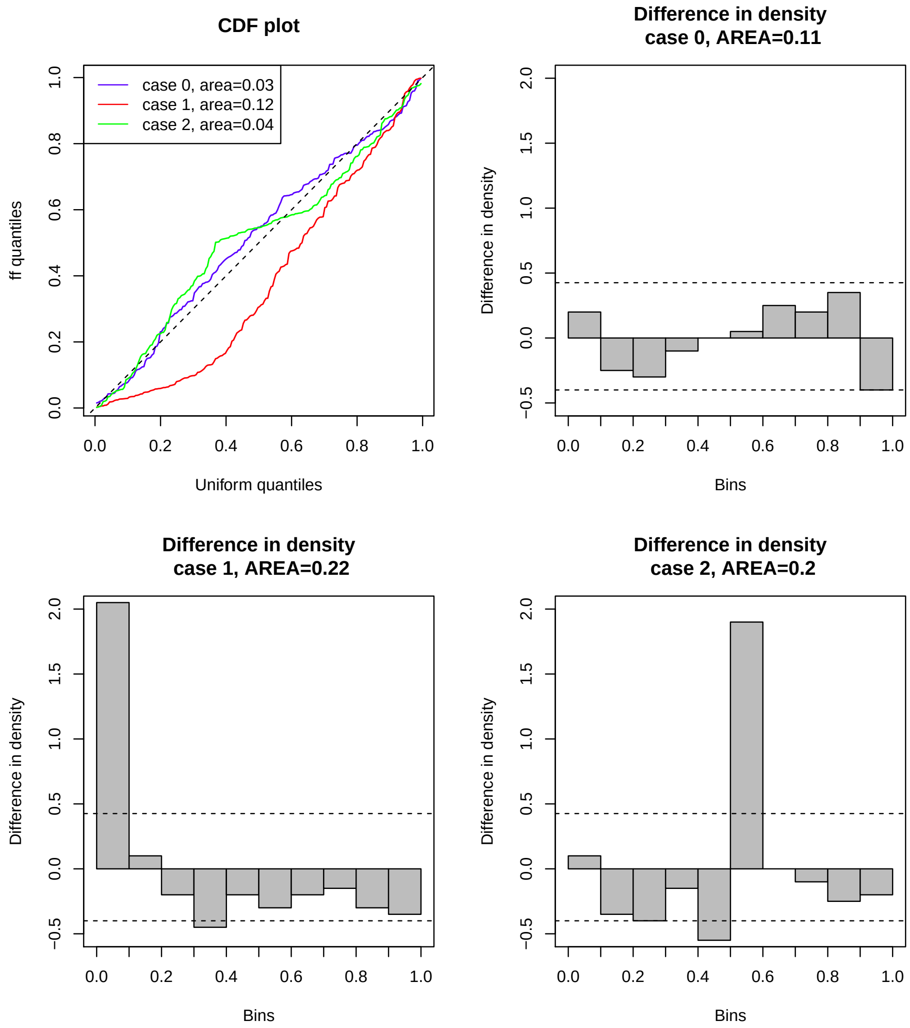

Figure 2. Graphical tools for model evaluation based on *FF* scores, for three simulated series of length 200. The CDF case (upper left) is the method of Renard et al. (2013). The density case (upper right and lower panels) is the alternative method comparing densities (Eq. 10). The dotted horizontal lines show the 95 % confidence interval for uniform variates on $(0,1)$ of length 200, based on 1000 simulations.

$ff^{(21)}$, this gives us two scores of model evaluation, namely AREA($FF^{(12)}$) and AREA($FF^{(21)}$).

The N_T criterion assesses reliability of the fit, as FF, but focuses on prescribed quantiles rather than on the overall maximum. Let $(X_1,\ldots,X_n)$ be a set of n independent and identically distributed rainfall values with distribution F, and let N_T be the random variable equal to the number of exceedances of the T-year return level, i.e., $N_T = \mathrm{Card}\{X_j; F(X_j) > 1-1/(\zeta T)\}$, where ζ is the mean number of observations per year. Since every event $\{F(X_j) > 1-1/(\zeta T)\}$ occurs with probability $1/(\zeta T)$, N_T follows a binomial distribution with parameters $(n, 1/(\zeta T))$. Let H_T be the corresponding cumulative distribution function, i.e., such that $H_T(k) = \Pr(N_T \le k)$, $k = 0,\ldots,n$ and $H(-1) = 0$. Because H_T is not continuous, the probability-transformed indices $H_T(N_T)$ are not uniform. Thus, Renard et al. (2013) propose to consider the random variable $\widetilde{N}_T$ such that

$$\widetilde{N}_T | N_T = k \sim \mathrm{Unif}\{H_T(k-1), H_T(k)\},$$

and show that $\widetilde{N}_T$ is uniform on $(0,1)$. Now, consider the estimates $\hat{F}_i^{(1)}$ and $\hat{F}_i^{(2)}$ for a given station i and

$$\begin{aligned} n_{T,i}^{(12)} &= \mathrm{Card}\{x_{i,j} \in C_i^{(1)}; \hat{F}_i^{(2)}(x_{i,j}) > 1-1/(\zeta_i T)\}, \\ n_{T,i}^{(21)} &= \mathrm{Card}\{x_{i,j} \in C_i^{(2)}; \hat{F}_i^{(1)}(x_{i,j}) > 1-1/(\zeta_i T)\}, \end{aligned}$$

where ζ_i is the mean number of central rainfall events per year at station i. If $F_i^{(1)}$ and $F_i^{(2)}$ are exact estimates for F, then $n_{T,i}^{(12)}$ (resp. $n_{T,i}^{(21)}$) should be realizations of a binomial with parameters $n_i^{(1)}$ (resp. $n_i^{(2)}$) and $1/(\zeta_i T)$. Let $H_{T,i}^{(1)}$ and $H_{T,i}^{(2)}$ be the corresponding binomial cumulative distribution functions and let $\widetilde{n}_{T,i}^{(jk)}$, $j,k = 1,2$, be uniform simulations between $H^{(k)}_{T,i}(n^{(jk)}_{T,i}-1)$ and $H^{(k)}_{T,i}(n^{(jk)}_{T,i})$. Then $\widetilde{n}^{(jk)}_{T,i}$ are realizations of the uniform distribution (Renard et al., 2013). For i ranging over the set of N stations, we thus obtain two vectors of size N of uniform samples, so that we can write $\widetilde{n}^{(12)}_{T}$, $\widetilde{n}^{(21)}_{T}$. Scores are calculated as for FF by comparing the empirical densities of $\widetilde{n}^{(jk)}_{T}$, $j,k=1,2$ to the theoretical uniform density, giving the two scores AREA($N^{(jk)}_T$).

4 Application of MEWP and MGPWP in Norway

4.1 Models considered

We wish to evaluate and compare the performance of EXP, GPD, MEWP and MGPWP for estimating central rainfall values across Norway. To apply the split sample procedure described in Sect. 3.4 for each station i, we randomly divide years into two subsamples such that 50 % of the observed years are in sample $\mathrm{C}^{(1)}_i$ and the remaining 50 % are in sample $\mathrm{C}^{(2)}_i$. This split sample procedure is applied to each station independently (meaning that years of $\mathrm{C}^{(1)}_i$ and $\mathrm{C}^{(1)}_{i'}$ are very unlikely to all be equal for $i \neq i'$). This creates two new data sets, each comprising 192 stations with a maximum of 31 years of observations.

As is always the case for extreme value analysis, threshold choice is uncertain. We therefore considered a large set of thresholds with α between 0.50 and 0.97. The evaluation scores are then used to select both the best model and the best threshold(s). Choice of α as low as 0.50 may at first glance appear to be very low for studying extremes, but one has to remember that the data series are already preprocessed to include only central rainfall values. Days with central rainfall will tend to have higher intensities than a randomly selected day with rainfall, as by construction, the central rainfall series excludes the previous and following days with lower rainfall intensities (see Sect. 2). A threshold level of 0.50 corresponds actually to a level of about 0.75 for the daily (nonzero) rainfall values.

The estimation scheme can be summarized as follows. For each of the considered α values, we fit six models with the exponential distribution:

- EXP(α), which is a particular case of MEWP, where S is one season and K is one weather pattern;
- MEWP($\alpha, 1, K$), i.e., a combination of K WP distributions, where $K=4$ or 8 (see below);
- MEWP($\alpha, 2, 1$), i.e., a combination of two seasonal distributions. Choice of the seasons is explained below;
- MEWP($\alpha, 2, K$), i.e., a combination of seasonal and WP distributions, with $K=4$ or 8;

and the six corresponding models with the GPD distribution. This gives in total 12 fits $\hat{F}^{(1)}_i$ and 12 fits $\hat{F}^{(2)}_i$, for each station i and each level α.

For the cases involving the use of WP, we employ the weather-type (WT) classification described in Fleig (2011), following the "bottom-up" method presented in Garavaglia et al. (2010). Details of this scheme are also reported in Lawrence et al. (2014) and can be briefly summarized as follows: ascending hierarchical classification is first performed on the rain fields for days with rain, as described by 175 stations in Norway and the surrounding region. The average synoptic pattern (WT) associated with each rain-field class is then identified from an atmospheric pressure data set constructed from geopotential height data centered over Norway. Finally, every day of the period considered (1948–2009) is assigned to a WT using the proximity of its geopotential height data to one described by a WT. In the first instance (Fleig, 2011), eight distinct WTs were defined, seven corresponding to days with rain and one representing dry days. For the first application of SCHADEX in Norway (Lawrence et al., 2014), a grouping of the eight weather types into four weather patterns (WP) was made to improve the robustness of the MEWP models (Fig. 3) by increasing the number of values in the subsamples. In this paper we, however, use the term "weather patterns" (WP) to refer to both sets of classifications, i.e., having four or eight classes; and both the use of the full set of eight classes or the grouped set of four classes are evaluated.

In cases where subsampling is also undertaken by season, we impose a restriction of S being two seasons, representing the season-at-risk and the season-not-at-risk. Furthermore, we impose the season-at-risk to be composed of 2 to 4 consecutive months (the remaining months falling in the season-not-at-risk). The optimum choice of the months composing the season-at-risk is made following the procedure of Penot (2014), which is applied to each station and model separately, using the whole series (i.e., without splitting into $\mathrm{C}^{(1)}$ or $\mathrm{C}^{(2)}$). The principle is to find the season-at-risk for which the estimated model fits at best the months with the highest risk (of extreme rainfall intensities). In detail, the procedure is as follows:

- Step 1: compute the 12 mean monthly maxima of central rainfall.
- Step 2: set $M=2$.
- Step 3: compute the mean of these values over moving windows of size M months.
- Step 4: select the M consecutive months corresponding to the highest of these values. These M months define the season-at-risk. The remaining months define the season-not-at-risk.
- Step 5: fit the considered model (e.g., MEWP(0.5, 2, 8)) with this seasonal definition.
- Step 6: compare the monthly fits to the monthly empirical distributions. This comparison is made with the

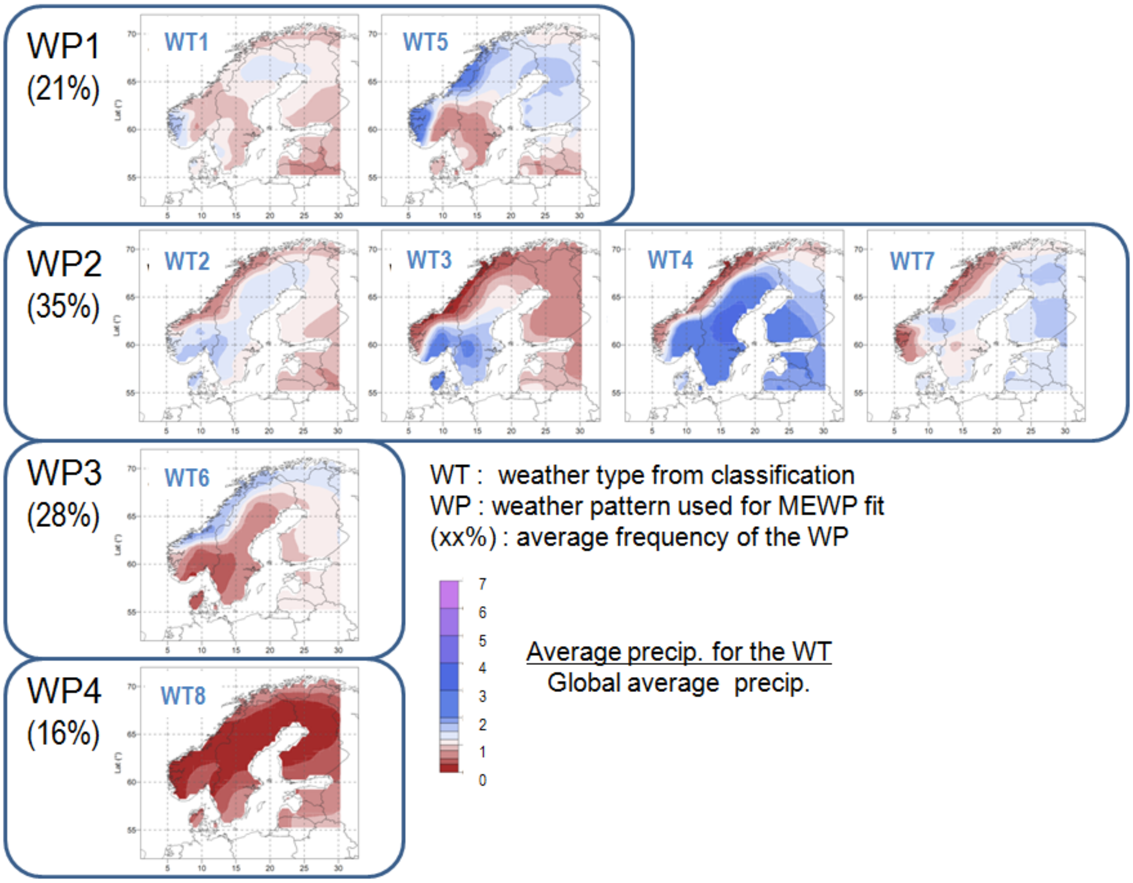
WP1
(21%)
WT1
WT5
WP2
(35%)
WT2
WT3
WT4
WT7
WP3
(28%)
WT6
WP4
(16%)
WT8
WT : weather type from classification
WP : weather pattern used for MEWP fit
(xx%) : average frequency of the WP
7
6
5
4
3
2
1
0
Average precip. for the WT
Global average precip.

Figure 3. Weather pattern classification with four classes (denoted WT1 to WT4 above) and eight classes (WP1 to WP8 above). This is Fig. 5 of Lawrence et al. (2014). Case with four classes is obtained by combining the eight classes into four. The last class of each classification (respectively WT4 and WP8) represent dry days.

KGE score (Kling–Gupta efficiency, Gupta et al., 2009), which is computed for a given month, m, as

$$\mathrm{KGE}_m = \left\{\mathrm{corr}(\widetilde{F}_m, \hat{F}_m) - 1\right\}^2 + \left\{\mathrm{std}\left(\frac{\widetilde{F}_m}{\hat{F}_m}\right) - 1\right\}^2 + \left\{\mathrm{mean}\left(\frac{\widetilde{F}_m}{\hat{F}_m}\right) - 1\right\}^2,$$

where $\widetilde{F}_m$ and $\hat{F}_m$ are respectively the empirical and fitted distributions for month m. It should be noted that the KGE criterion is not the only score which could be used here, and was not necessarily developed for scoring distributions. However, the final result (i.e., the seasonal split selected) is not particularly sensitive to the score used.

- Step 7: compute a global KGE score as a weighted mean of these 12 KGE scores, with weights proportional to the mean monthly maxima, in order to force the model to have the best fits for the months with the highest risk.
- Step 8: set $M = 3$ and apply steps 3 to 7.
- Step 9: set $M = 4$ and apply steps 3 to 7.
- Step 10: compare the three global KGE scores obtained respectively for $M = 2, 3, 4$. Select the seasonal definition corresponding to the lowest of these scores.

This procedure is applied for each station and each model separately. This implies that, for a given station, the choice of season may vary among models. However, it was found that changes in the definition of the season-at-risk for a given station are very minimal (i.e., a few percent difference, and always pertain to the intermediate months that could well be classified into either of the two periods). We believe that these differences have very little influence on the evaluation of the model fits. For illustration, Fig. 4 shows the length of the season-at-risk and the first month of this season for the 192 Norwegian stations when using MEWP(0.5, 2, 8) (which is found to be the best model; see Sect. 4.2.1). Interestingly, the local definition of the seasons define four regions with an intense season in fall in the western part of Norway and an intense season in late summer–early fall

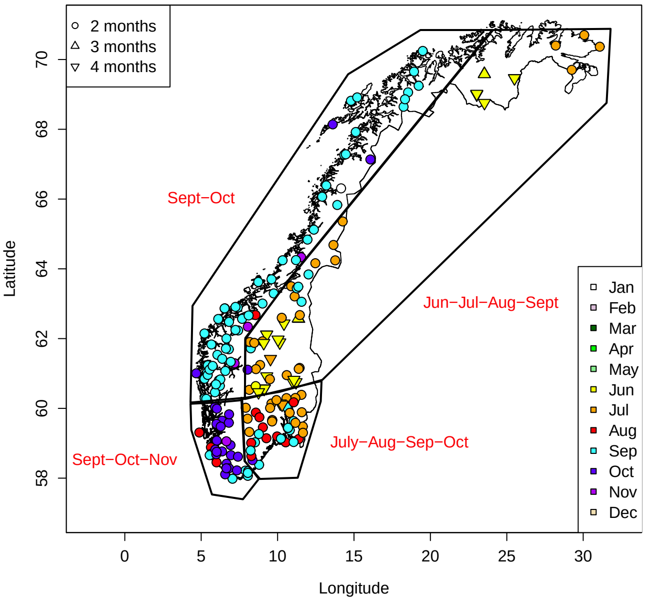

Figure 4. Length of the season-at-risk (shapes) and first month of the season (color code in the inset) for each station, with model MEWP$(0.5, 2, 8)$. The local definition of seasons is used in Sect. 4.2.1, while the regional definition, with four regions, is used in Sect. 4.2.2.

in the eastern part. Furthermore, the intense season starts 1 month earlier in the eastern part than in the western part. The distinction between a heavy rainfall season beginning in the fall in western Norway vs. late summer in eastern Norway is associated with the two different mechanisms leading to heavy precipitation in each of these regions. In western Norway, heavy precipitation is most commonly derived from frontal activity leading to storms arriving from the southwest. The eastern part of Norway is in the lee of the mountainous area in the central zone of southern Norway, and is, therefore, somewhat sheltered from this storm activity. The heaviest precipitation in the eastern region generally occurs due to convective activity producing intense rain showers, often during the late summer months. It can also be noted that the spatial pattern of the precipitation seasons shows a good correspondence with previously published maps of precipitation regions in Norway (see e.g., Hanssen-Bauer and Førland, 2000, Fig. 1) and with the occurrence of days with precipitation over 10 mm (see Tveito et al., 2001, Fig. 2.5). The regional seasons will be used in Sect. 4.2.2 to check the sensitivity of MEWP with respect to slight changes in the definition of the season-at-risk.

4.2 Results

4.2.1 Model evaluation and selection

The SPAN, FF and N_T scores presented in Sect. 3.4 are used to assess the quality of the estimations. We use the three scores because they give complementary answers. Taken together, they allow a global evaluation of both the reliability and the stability of the fits. Different return periods T are also considered for SPAN and N_T in order to evaluate different parts of the tail of the distribution. With large T we assess the very tail of the distribution while with small T we assess the bulk of the distribution.

Scores are reported in Figs. 5 and 6 for the 12 models, using threshold values equal to the 0.5-, 0.7- and 0.9-quantile of the central rainfalls. Keep in mind that all scores lie in the range $(0, 1)$ and the closer to 0, the better the score. For each model and threshold, we depict three MEAN(SPAN$_T$) scores for $T = 20, 100$ and 1000 years, the value of AREA($FF^{(12)}$) and the three AREA($N_T^{(12)}$) values for $T = 5, 10$ and 20 years. Values of AREA($FF^{(21)}$) and AREA($N_T^{(21)}$) are not shown as they are very similar. For the SPAN scores, it may seem highly questionable to extrapolate return levels up to 1000 years given that estimation is based on about 30 years of data. This is actually the level required by engineering practices and regulatory rules (if not higher) in many countries for risk assessment associated with dam safety. For example, in France 1000- or even 10 000-year return periods are used to design dam spillways (Paquet et al., 2013), and the 1000-year return period is also used as the design flood level for the higher risk classes of dams in Norway, whilst the probable maximum flood is used to assess the safety of these dams with respect to the potential for dam failure (Midttømme et al., 2011).

Figure 5 shows that for the exponential models, there is a clear benefit obtained from the use of seasonal splitting (case EXP vs. $(S, K) = (2, 1)$) and WP splitting (case EXP vs. $(S, K) = (1, 4)$ and $(1, 8)$), and the combination of both seasonal and WP splitting performs even better (see cases $(2, 4)$ and $(2, 8)$). Indeed, subsampling by season and WP creates groups of rainfall values that are more likely to be identically distributed and therefore more easily fitted than groups of rainfall values derived from different parent populations. Using eight rather that four WPs also slightly improves the N_T scores, but the improvement is somewhat marginal when compared with the gain derived from sampling by season and WP.

Figure 5 surprisingly shows that for MEWP distributions, scores of N_T improve when T increases, meaning that the bulk of the distribution is actually less well fitted than the tail. This may be due to the lack of flexibility of the exponential distribution. Using the more flexible GPD distribution (in the GPD and MGPWP models of Fig. 6) indeed tends to improve N_5 and N_{10}. However, it clearly also degrades the FF scores. Keep in mind that FF is based on the maximum observed value (see Sect. 3.4) and, thus, permits an assessment of the quality of the fit of the very tail of the distribution. Therefore, although the bulk of the distribution tends to be better fitted with MGPWP distributions (N_5 and N_{10}), the very tail (FF) is overfitted, usually giving poorer FF scores.

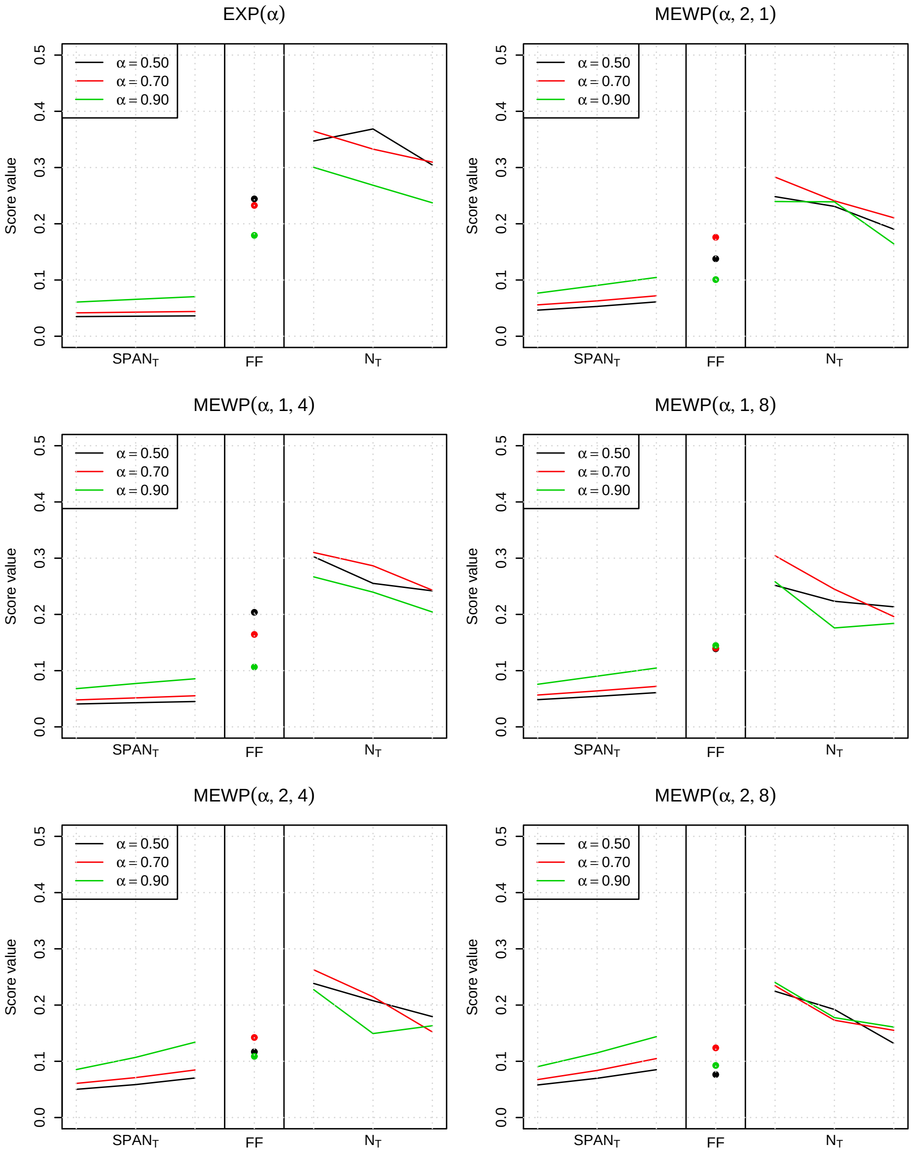

Figure 5. Scores of evaluation for MEWP models, for $\alpha = 0.5$, 0.7 and 0.9. Better scores have values closer to 0. Scores of SPAN$_T$, for $T = 20-$, $100-$ and 1000-year return periods, are the mean scores of Eq. (9), while scores of *FF* and N_T, $T = 5, 10$ and 20 years, are based on the density areas (Eq. 10).

Figure 6 also shows a clear loss in stability (indicated by the SPAN scores) when using the MGPWP distribution. Figure 7 illustrates this issue by comparing the 100-year and 1000-year return levels estimated on C$^{(1)}$ and C$^{(2)}$ with the four MEWP models and the four MGPWP models, with a level $\alpha = 0.5$. This shows a difference of up to 100 mm day^{-1} with MGPWP models for the 100-year return level and up to 300 mm day^{-1} for the 1000 year-return level, whereas the MEWP models are much more stable. This lack of robustness is due to the difficulty in estimating the shape parameter ξ of the GPD distribution, which has a large influence on the extrapolation to long return periods (see also page 528 of Garavaglia et al. (2011) or the upper right of page 350 of Serinaldi and Kilsby (2014)). Figure 8, on the left hand side, compares the values of ξ estimated on C$^{(1)}$ and C$^{(2)}$ by all MGPWP models. Values between -0.5 and 0.5 are mainly found, but differences between the two estimates vary in a similar range. Positive values, even when not very large (typically $\xi > 0.1$) lead to unrealistic return levels at extrapolation, with e.g., up to 600 mm day^{-1} for the 1000-year return level in the MGPWP case versus 270 mm day^{-1} in the MEWP case (see Fig. 7). Figure 8, right, shows that estimates of ξ based on fewer than 1000 observations are highly variable. Similar variability in the shape of the GPD is found in Serinaldi and Kilsby (2014) for a worldwide data set. Cases with fewer than 1000 observations occur more of-

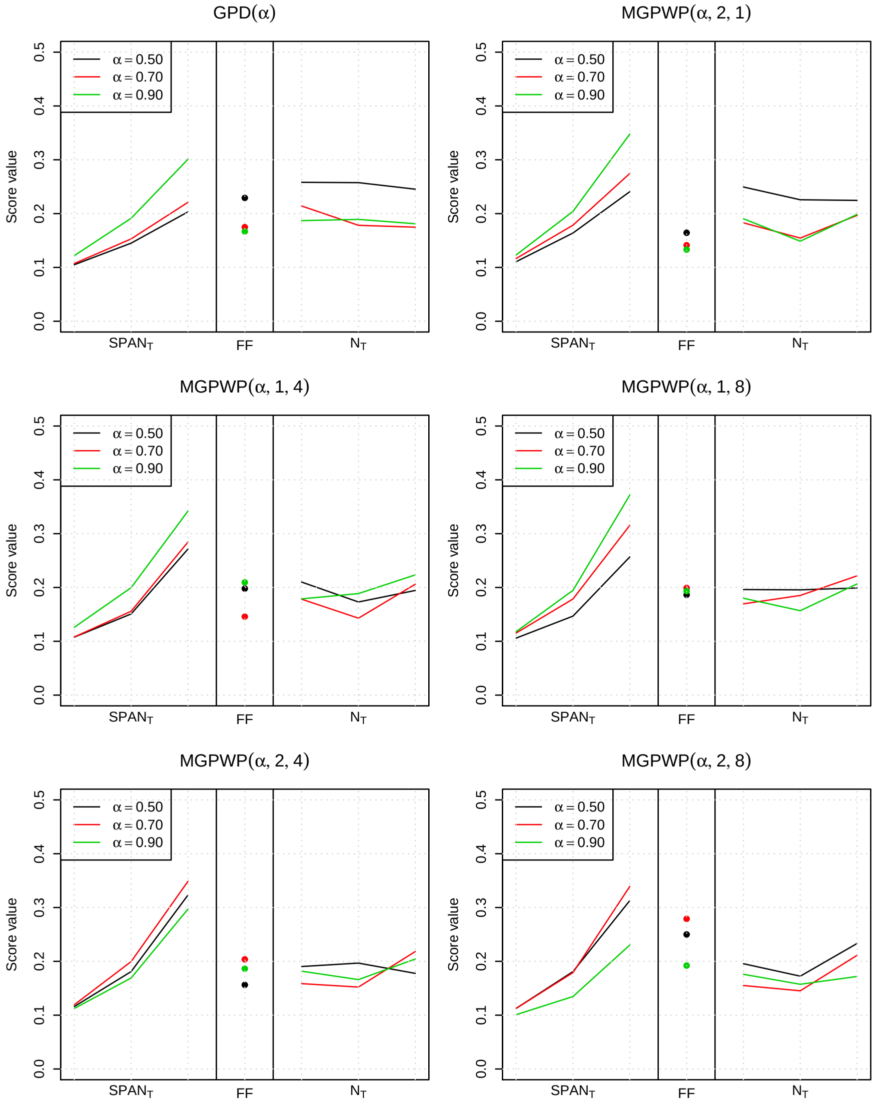

Figure 6. Same as Fig. 5 for MGPWP models.

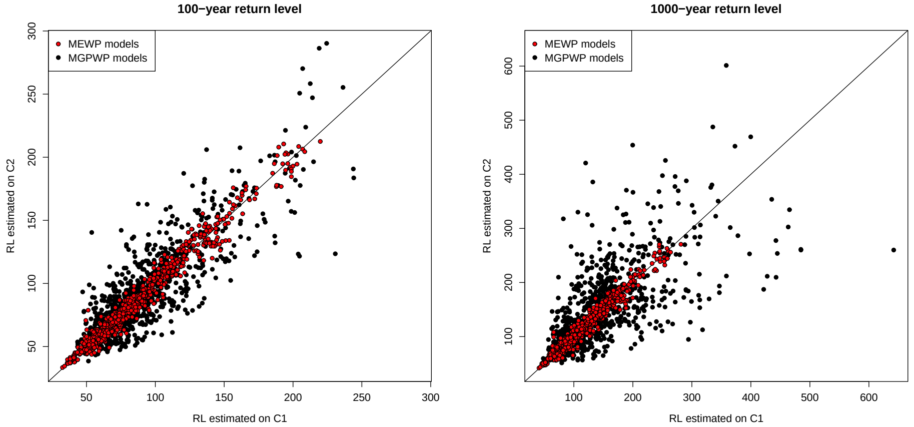

Figure 7. Comparison of the 100-year and 1000-year return levels (in mm) estimated on $C^{(1)}$ and $C^{(2)}$, for the four MEWP models (in red) and the four MGPWP models (in black), with a level $\alpha = 0.5$ (one point per station).

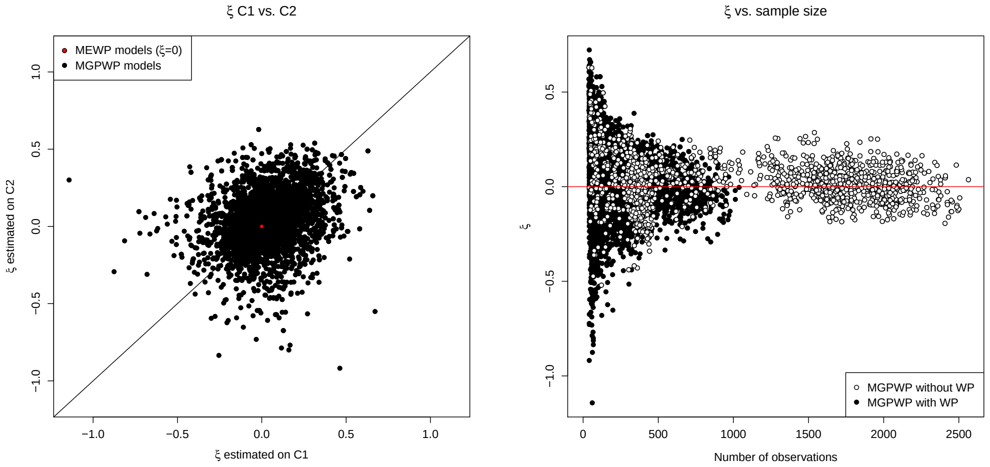

Figure 8. Left: estimated ξs on $C^{(1)}$ and $C^{(2)}$ for the four MGPWP models, with $\alpha = 0.5$ (one point per station). MEWP models correspond to $\xi = 0$ (red points). Right: same ξs as a function of the sample size with WP (black points) and without WP (white points) (one point per station and period).

ten when WP are considered, due to the additional subsampling which produces smaller data sets. However, the SPAN values of Fig. 6 show that even for the GPD and MGPWP with $K = 1$, robustness is very poor. This lack of robustness is an important limitation of their value and suitability for practical applications.

Regarding the choice of threshold, MEWP distributions give relatively stable scores for α between 0.5 and 0.7 (see Fig. 5) but there is a loss in stability as α increases over 0.9 (see green curves of SPAN scores in Fig. 5). For MEWP$(\alpha, 2, 8)$, which gives the best scores overall, the case $\alpha = 0.5$ usually seems to be slightly better. Therefore we select the model MEWP$(0.5, 2, 8)$ for further consideration.

It is interesting at this point to compare large return levels obtained with the selected MEWP$(0.5, 2, 8)$ with those obtained for the other MEWP models with the same α. Figure 9 makes this comparison for the 100-year return levels. It appears that the other MEWP models tend to give lower return levels (i.e., positive values of the difference). This underestimation is more marked for the EXP model (mean underestimation of about 5 mm of the 100-year return level), and decreases when seasons (MEWP$(0.5, 2, 1)$) and WP (MEWP$(0.5, 2, 4)$) are used. Therefore, the use of more WPs helps to better model the heaviness of the tail.

4.2.2 Use of regional seasons

We have already mentioned in Sect. 4.1 that the local definition of the seasons displays a regional pattern, with a season-at-risk in late summer in the two eastern regions and in fall in the two western ones, as illustrated in Fig. 4. We test here the use of this regional definition of the seasons by fitting new MEWP$(0.5, 2, 8)$ models and comparing the overall scores to those of the local definition of Sect. 4.2.1. As shown in Table 1, scores of the two definitions are fairly similar, particularly in light of the differences obtained between the models of Fig. 5. Robustness (SPAN) is slightly improved with the regional definition. However the fact that scores of both *FF* and N_{20} are slightly better (i.e., smaller) when seasons are defined locally gives evidence of a better fit of the very tail with the local definition, and therefore probably a better extrapolation of return levels. Therefore, if one would want to select one and only one definition, we would be tempted to recommend the local one. However, if using MEWP at ungauged sites is of interest, the regional definition of the seasons of Fig. 4 provides a reasonable alternative.

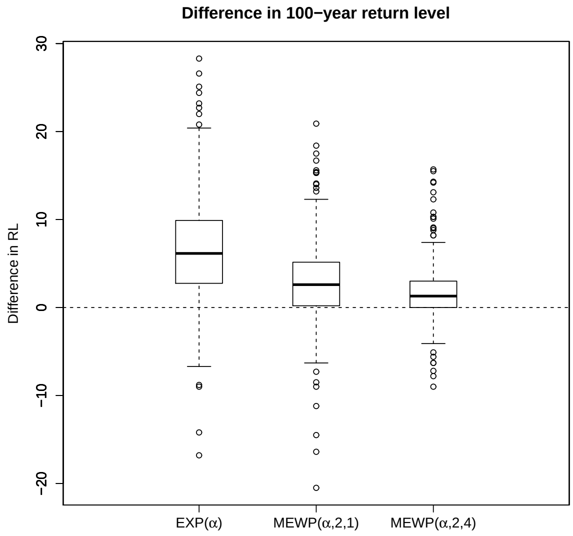

Figure 9. Box plot of the difference (in mm) between the 100-year return levels of MEWP$(\alpha, 2, 8)$ and the three other EXP-based models, for $\alpha = 0.5$ (one point per station and period).

Table 1. Scores of evaluation for the local and regional definition of the seasons. Better scores have values closer to 0. Scores of SPAN$_T$, for $T = 20, 100$ and 1000 years, are the mean scores of (Eq. 9), while scores of FF and N_T, $T = 5, 10$ and 20 years, are based on the density areas (Eq. 10).

	SPAN$_{20}$	SPAN$_{100}$	SPAN$_{1000}$	$FF^{(12)}$	$N_5^{(12)}$	$N_{10}^{(12)}$	$N_{20}^{(12)}$
Local seasons	0.058	0.070	0.085	0.076	0.209	0.163	0.130
Regional seasons	0.053	0.062	0.074	0.080	0.202	0.185	0.158

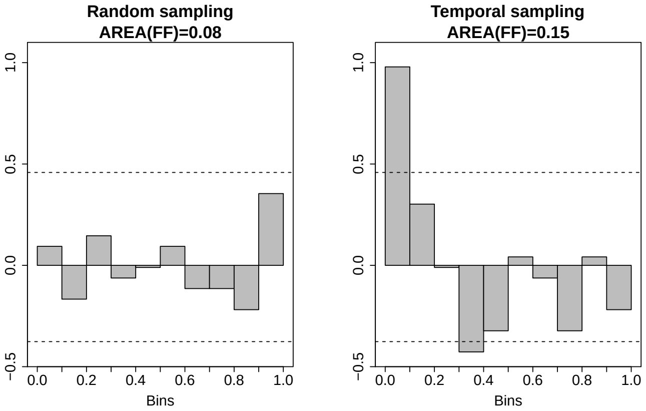

Figure 10. Divergence in density between $ff^{(12)}$ and the uniform case, under random sampling (left) and temporal sampling (right), with corresponding scores AREA(FF). The closer the bars are to 0, the better the fit is. The dotted horizontal lines show 95 % confidence interval for uniform variates.

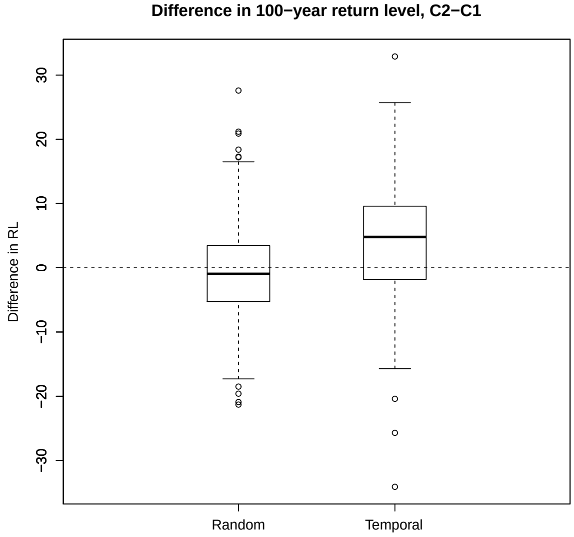

Figure 11. Box plot of the difference in 100-year return level estimated for C$^{(1)}$ and C$^{(2)}$ with MEWP$(0.5, 2, 8)$ under random sampling (left) and temporal sampling (right) (one point per station).

4.2.3 Evidence of trend

The split sample procedure can be used to give insight about potential change in extreme rainfall in Norway over the period represented by the rainfall time series. For this we split the observed years of each station into two subsamples: C$^{(1)}$ contains all years between 1948 and 1978 and C$^{(2)}$ contains all the remaining years, between 1979 and 2009. So, in contrast with the previous analysis, all stations are assigned the same C$^{(1)}$ and C$^{(2)}$ and these are temporal instead of being random. Remember that $ff^{(12)}$ assesses how well the maximum of C$^{(1)}$ is fitted by the distribution estimated on C$^{(2)}$, namely $\hat{F}^{(2)}$ (see Eq.). Therefore a parallel comparison of the density of the values of $ff_i^{(12)}$, for $i = 1, \ldots, 192$, for this temporal sampling compared to the random one of Sect. 4.2.1 can give insight into increases or decreases in extreme rainfall in Norway between the two periods. The density of these values is shown in Fig. 10.

We see that $ff^{(12)}$ tends to have too many small values with respect to the uniform density under the temporal sampling, whereas it was fairly uniform under the random sampling of Sect. 4.2.1 (a complementary analysis, not shown, revealed that very similar densities are obtained with other random splitting approaches). We conclude that $\hat{F}^{(2)}$ tends to overestimate the probability of occurrence of the maximum of C$^{(1)}$ under the temporal sampling. Broadly speaking this means that the maximum of C$^{(1)}$ tends to be too small with respect to that of C$^{(2)}$. This indicates that extremes during the second-half of the observed period (1979–2009) tend to be higher than those of the first half (1948–1978). This is confirmed by a comparison of return levels obtained on both periods, as shown in Fig. 11. For the random sampling case, return levels are almost equal for C$^{(1)}$ and C$^{(2)}$ whereas in the temporal sampling case, 100-year return level is about 5 mm higher in C$^{(2)}$, with 10 % of the stations showing an increase higher than 10 mm (vs. 3 % in the random case). As shown in Fig. 12, these 10 % stations lie mainly in the southwestern region, between Bergen and Stavanger, which is one of the most rainy areas in Norway, with 100-year return levels higher than 100 mm (Fig. 12, left). This brief analysis gives evidence for an increase in extreme rainfall intensities which may already be evident in observations for the southwestern region in Norway. This evaluation does not take the place of a full, detailed trend analysis per se, but rather should be taken as a motivation for such an analysis of trends. Our evaluation relies in particular on a somewhat arbitrary splitting of the years in the middle of the observation period. Assessment of possible trends, including when such trends started

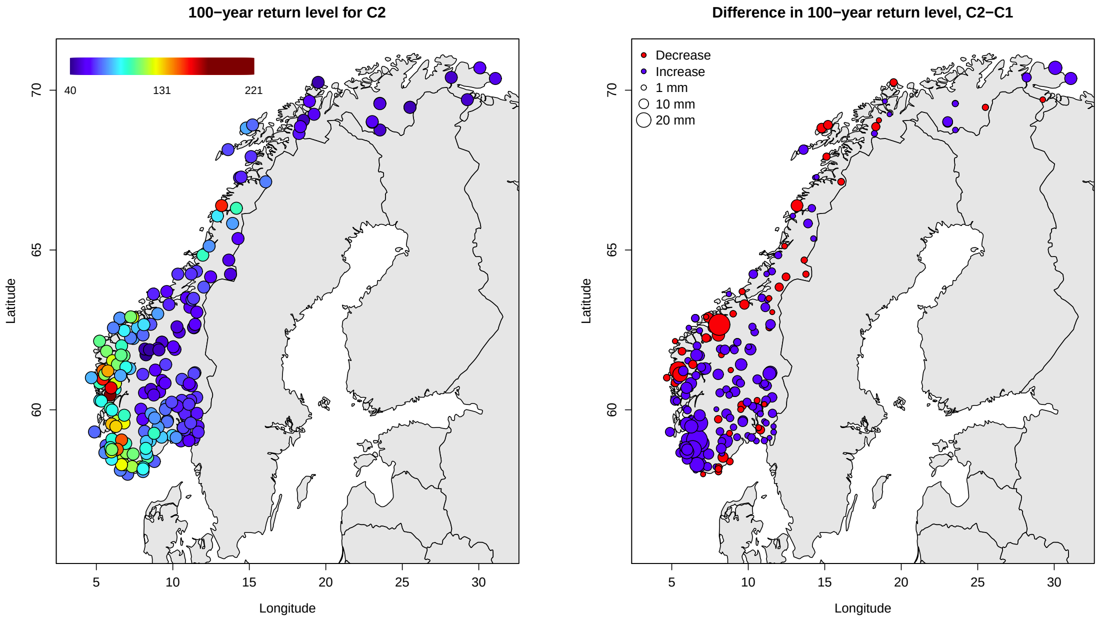

Figure 12. Left: map of 100-year return level estimated on $C^{(2)}$ (1979–2009) with $\text{MEWP}(0.5, 2, 8)$. Right: difference in 100-year estimated on $C^{(1)}$ and $C^{(2)}$.

and their consistency over time is beyond the scope of this paper, but may be of interest in future studies.

5 Conclusions

This article evaluates a compound model based on weather pattern classification, seasonal splitting and exponential distributions, the so-called MEWP model, for its suitability for use in Norway. The MEWP model is the rainfall probabilistic model used within the SCHADEX method, which is currently being tested in Norway as an alternative simulation method for flood estimation. We show in particular the benefit gained by subsampling the heavy rainfall data according to season and weather pattern. Our results also indicate that models based on the exponential distribution perform better than those based on the more flexible generalized Pareto distribution, which tends to overfit the data and lacks robustness. We have also demonstrated that a regional definition of seasons in MEWP is possible. Finally, we give evidence for an increase in extreme rainfall intensities in Norway in recent years, particularly in the southwestern region.

Our analysis has also shown that the GPD distribution better models the bulk of the distribution of extremes, but fails to robustly estimate the tail, and therefore fails in extrapolation to large return levels. The reason for this failure is twofold: firstly, the lack of data for estimating such a flexible distribution when using a local approach; secondly, the inherent nature of the GPD, which is a heavy-tailed distribution when the shape parameter is positive, and can therefore tend to give unrealistic return levels for very long return periods. To address this issue, a regional approach allowing the use of neighboring stations to infer MEWP distributions at local sites is of interest. Finally, there are also other, more flexible, distributions which may be more robust than the GPD distribution and could be used within the MEWP approach. This also represents an important topic for future work.

This study is the first extensive evaluation of MEWP in Norway. It has also been applied successfully in France (Garavaglia et al., 2011; Neppel et al., 2014), Austria and West Canada (Brigode et al., 2014). MEWP is a general model imposing no specific hypotheses on the data, so its application in other regions of the world is absolutely worth considering. The only limitation is that a classification into weather patterns suitable for evaluating extreme precipitation is needed as a precursor to such an analysis, but this is already available in several regions around the world (see Brigode et al., 2014)

Acknowledgements. Deborah Lawrence acknowledges support from the project “FLOMQ – Robust framework for estimating extreme floods in Norway” supported by the ENERGIX program of the Norwegian Research Council, as well as internal research funds from NVE which have made the collaboration with LTHE possible. The authors acknowledge support from the COST Action ES0901 FloodFreq which has made the collaboration between NVE and EDF possible, including a Short Term Scientific Mission for Anne Fleig to EDF.

Edited by: P. Tarolli
Reviewed by: two anonymous referees

References

Brigode, P., Bernardara, P., Paquet, E., Gailhard, J., Garavaglia, F., Merz, R., Mićović, Z., Lawrence, D., and Ribstein, P.: Sensitivity analysis of SCHADEX extreme flood estimations to observed hydrometeorological variability, Water Resour. Res., 50, 353–370, doi:10.1002/2013WR013687, 2014.

Coles, S.: An introduction to statistical modeling of extreme values, Springer Series in Statistics, Springer-Verlag, London, 208 pp., 2001.

Dyrrdal, A. V., Skaugen, T., Stordal, F., and Førland, E. J.: Estimating extreme areal precipitation in Norway from a gridded dataset, Hydrolog. Sci. J., doi:10.1080/02626667.2014.947289, 2014.

Ferro, C. A. T. and Segers, J.: Inference for Clusters of Extreme Values, J. Roy. Stat. Soc. B, 65, 545–556, 2003.

Fleig, A.: Scientific Report of the Short Term Scientific Mission - Anne Fleig visiting Électricité de France, Grenoble, Tech. rep., NVE, available at: http://www.costfloodfreq.eu/component/k2/item/download/6_8e45d035c2e09839e0c43e63ed0cdc81, 2011.

Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garçon, R., and Bernardara, P.: Introducing a rainfall compound distribution model based on weather patterns sub-sampling, Hydrol. Earth Syst. Sci., 14, 951–964, doi:10.5194/hess-14-951-2010, 2010.

Garavaglia, F., Lang, M., Paquet, E., Gailhard, J., Garçon, R., and Renard, B.: Reliability and robustness of rainfall compound distribution model based on weather pattern sub-sampling, Hydrol. Earth Syst. Sci., 15, 519–532, doi:10.5194/hess-15-519-2011, 2011.

Guillot, P.: The arguments of the gradex method: a logical support to assess extreme floods, Proceedings of the Yokohama Symposium, 213, 287–298, 1993.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, doi:http://dx.doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hanssen-Bauer, I. and Førland, E.: Temperature and precipitation variations in Norway 1900-1994 and their links to atmospheric circulation, Int. J. Climatol., 20, 1693–1708, doi:10.1002/1097-0088(20001130)20:14<1693::AID-JOC567>3.0.CO;2-7, 2000.

Hosking, J. R. M.: L-moments: analysis and estimation of distributions using linear combinations of order statistics, J. Roy. Stat. Soc. B, 52, 105–124, 1990.

Klemes, V.: Tall Tales about Tails of Hydrological Distributions. I, J. Hydrol. Eng., 5, 227–231, doi:10.1061/(ASCE)1084-0699(2000)5:3(227), 2000a.

Klemes, V.: Tall Tales about Tails of Hydrological Distributions. II, J. Hydrol. Eng., 5, 232–239, doi:10.1061/(ASCE)1084-0699(2000)5:3(232), 2000b.

Lawrence, D., Paquet, E., Gailhard, J., and Fleig, A. K.: Stochastic semi-continuous simulation for extreme flood estimation in catchments with combined rainfall-snowmelt flood regimes, Nat. Hazards Earth Syst. Sci., 14, 1283–1298, doi:10.5194/nhess-14-1283-2014, 2014.

Midttømme, G., Pettersson, L., Holmqvist, E., Nøtsund, Ø., Hisdal, H., and Sivertsgård, R.: Retningslinjer for flomberegninger (Guidelines for flood estimation), NVE Retningslinjer, 4/2011, 2011.

Neppel, L., Arnaud, P., Borchi, F., Carreau, J., Garavaglia, F., Lang, M., Paquet, E., Renard, B., Soubeyroux, J., and Veysseire, J.: Résultats du projet Extraflo sur la comparaison des méthodes d'estimation des pluies extrêmes en France, La Houille Blanche – Revue internationale de l'eau, 2, 14–19, doi:10.1051/lhb/2014011, 2014.

Paquet, E., Garavaglia, F., Garçon, R., and Gailhard, J.: The SCHADEX method: A semi-continuous rainfall-runoff simulation for extreme flood estimation, J. Hydrol., 495, 23–37, doi:http://dx.doi.org/10.1016/j.jhydrol.2013.04.045, 2013.

Penot, D.: Cartographie des événements hydrologiques extrêmes et estimation SCHADEX en sites non jaugés, PhD thesis, Université de Grenoble, 244 pp., 2014.

Renard, B., Kochanek, K., Lang, M., Garavaglia, F., Paquet, E., Neppel, L., Najib, K., Carreau, J., Arnaud, P., Aubert, Y., Borchi, F., Soubeyroux, J.-M., Jourdain, S., Veysseire, J.-M., Sauquet, E., Cipriani, T., and Auffray, A.: Data-based comparison of frequency analysis methods: A general framework, Water Resour. Res., 49, 825–843, doi:10.1002/wrcr.20087, 2013.

Serinaldi, F. and Kilsby, C. G.: Rainfall extremes: Toward reconciliation after the battle of distributions, Water Resour. Res., 50, 336–352, doi:10.1002/2013WR014211, 2014.

Tveito, O., Førland, E., Alexandersson, H., Drebs, A., Jónsson, T., Tuomenvirta, H., and Vaarby Laursen, E.: Nordic climate maps, Tech. Rep. 06/01 KLIMA, DNMI – Report, Norwegian Meteorological Institute, Oslo, Norway, 28 pp., 2001.