Evaluation framework for sub-daily rainfall extremes simulated by regional climate models

Hans Van de Vyver*, Bert Van Schaeybroeck, Rozemien De Troch, Lesley De Cruz and Rafiq Hamdi

Royal Meteorological Institute, Brussels, Belgium

Cecille Villanueva-Birriel†, Philippe Marbaix and Jean-Pascal van Ypersele

Earth and Life Institute, Georges Lemaître Centre for Earth and Climate Research, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Hendrik Wouters, Sam Vanden Broucke and Nicole P.M. van Lipzig

Department of Earth and Environmental Sciences, KU Leuven, Leuven, Belgium

Sébastien Doutreloup, Coraline Wyard‡, Chloé Scholzen§ and Xavier Fettweis

Laboratory of Climatology, Department of Geography, UR SPHERES, Université de Liège, Liège, Belgium

Steven Caluwaerts and Piet Termonia

Royal Meteorological Institute, Brussels, and Department of Physics and Astronomy, Ghent University, Ghent, Belgium

**Corresponding author*: Hans Van de Vyver, hvijver@meteo.be

[†]Current affiliation: National Weather Service, NOAA , San Juan, Puerto Rico

[‡]Remote Sensing and Geodata Unit, Institut Scientifique de Service Public, Liège, Belgium

[§]Department of Geosciences University of Oslo, Oslo, Norway

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

ABSTRACT

Sub-daily precipitation extremes are high-impact events that can result in flash floods, sewer system overload, or landslides. Several studies have reported an intensification of projected short-duration extreme rainfall in a warmer future climate. Traditionally, regional climate models (RCMs) are run at a coarse resolution using deep-convection parameterization for these extreme events. As computational resources are continuously ramping up, these models are run at convection-permitting resolution, thereby partly resolving the small-scale precipitation events explicitly. To date, a comprehensive evaluation of convection-permitting models is still missing. We propose an evaluation strategy for simulated sub-daily rainfall extremes that summarizes the overall RCM performance. More specifically, the following metrics are addressed: the seasonal/diurnal cycle, temperature and humidity dependency, temporal scaling and spatio-temporal clustering. The aim of this paper is: (i) to provide a statistical modeling framework for some of the metrics, based on extreme value analysis, (ii) to apply the evaluation metrics to a micro-ensemble of convection-permitting RCM simulations over Belgium, against high-frequency observations, and (iii) to investigate the added value of convection-permitting scales with respect to coarser 12-km resolution. We find that convection-permitting models improved precipitation extremes on shorter time scales (i.e, hourly or two-hourly), but not on 6h-24h time scales. Some metrics such as the diurnal cycle or the Clausius-Clapeyron rate are improved by convection-permitting models, whereas the seasonal cycle appears robust across spatial scales. On the other hand, the spatial dependence is poorly represented at both convection-permitting scales and coarser scales. Our framework provides perspectives for improving high-resolution atmospheric numerical modeling and datasets for hydrological applications.

1. Introduction

Extreme precipitation events have a large impact on society through damage caused by floods, dike breaches, landslides, or sewer system overload. The total cost of natural disasters in the European Environment Agency (EEA) member countries during 1980-2017 is estimated at EUR 557 billion (EEA 2019). Around 63 % of all the economic losses were the result of storms and rainfall-induced floods.

Sub-daily extreme rainfall, caused by convective events, is known to impact many sectors such as agriculture, forestry, tourism, health, and water management. As an example, sub-daily extreme rainfall can induce flash floods that may develop in time scales shorter than an hour in an urban area. In recent years, there has been a growing body of evidence that short-duration rainfall intensity has increased at global scale (Westra et al. 2014), and that future increases are to be expected (Ban et al. 2015; Prein et al. 2017; Martel et al. 2020).

Lately, the effect of climate change on short-duration extreme precipitation has been under strong investigation through the use of climate modeling. In fact, current Global Climate Models (GCMs) operate on a spatial resolution in the range of 50–100 km (e.g. CMIP6 - Coupled Model Intercomparison Project Phase 6), or even on a finer scale of 25 km (Vannière et al. 2019), and are therefore unable to explicitly resolve small-scale physical processes such as convection. Regional Climate Models (RCMs) allow to downscale the GCM results by increasing the spatial resolution in a small, limited area of interest. As a result, significant international efforts, such as the Coordinated Regional Downscaling Experiment (CORDEX), have been made to downscale GCMs using RCMs over various regions in the world (Giorgi et al. 2009). For the European domain, the simulations are conducted at resolutions of 0.44° $\sim$ 50km (EUR-44), and 0.11° $\sim$ 12.5km (EUR-11) (Jacob et al. 2014, 2020).

Several studies have proved that a clear added value of regional downscaling can be found for extreme precipitation (Prein et al. (2015) and references therein). For EURO-CORDEX, for instance, recent works found that the fine-gridded (12-km resolution) RCMs add value to the coarser-gridded (50-km) RCMs in terms of both mean and extreme precipitation for almost all seasons and all regions throughout Europe, especially due to improved orographic representation but in part also due to a better representation of convective processes (Torma et al. 2015; Prein et al. 2015, 2016).

While coarser resolution RCMs (grid spacing above 5 km) are found reliable for projecting seasonal changes and frequency in precipitation, simulations with a horizontal grid spacing below 5 km are generally considered to resolve (at least partly) convective phenomena, which is essential to capture rainfall extremes. Kendon et al. (2017) found that the explicit representation of the convective storms is necessary for capturing changes in the intensity and duration of summertime rain on daily and shorter time scales. Climate-change projections at these convection-permitting scales exist for different limited regions worldwide (e.g. Pan et al. 2011; De Troch et al. 2014; Argüeso et al. 2014; Kendon et al. 2014; Ban et al. 2015; Brisson et al. 2016b; Fosser et al. 2017; Reszler et al. 2018), but are mostly single-model efforts due to the huge computational costs. Coordinated ensemble simulations at convection-permitting scale have been very rare (Helsen et al. 2020; Met Office 2021), but have recently been set up over different areas in Europe with promising first results (Coppola et al. 2020). Large climate simulations at the 1-km scale are currently being considered but face data-handling issues (Schär et al. 2020).

Prior to climate change assessments, it is important to compare the runs with observations to find potential added value of high-resolution simulations and to gain confidence in climate projections. The evaluation of precipitation extremes in RCMs is traditionally based on a comparison of the extreme value statistics of observations and simulations. In a novel approach, Westra et al.

(2014) and Cortés-Hernández et al. (2016) formulated a range of statistical metrics to evaluate the capability of RCMs in capturing the dominant spatio-temporal characteristics of short-duration rainfall extremes. This includes the assessment of a correct simulation of the seasonal cycle, the diurnal cycle, the relationship between extreme intensity and temperature/humidity (e.g. ~CC-scaling), temporal scaling (IDF information), and the spatial organization of rainfall extremes. A positive evaluation of these metrics should increase confidence in RCM simulations, and could give more insights in its limitations with potential directions for model improvements.

In addition to model evaluation, the metrics can also support the relationship between climate science and impact studies, the latter of which is highly relevant to climate services and decision-making. For instance, a correct simulation of the diurnal cycle is required for quantifying outdoor thermal comfort conditions because the timing of a heavy rainfall event is important. Similarly, a correct simulation of the seasonal cycle concerns many sectors such as agriculture, forestry, tourism, and health because it is important to determine the correct season in which the most extreme rainfall events occur. However, different sectors are interested in different rainfall accumulation periods. For instance, while urban water resources engineers are mainly interested in information on sub-daily rainfall extremes to protect various engineering systems (e.g. sewer systems, urban drainage systems) against floods, farmers or tour operators are rather interested in monthly or seasonal rainfall extremes. One of the most commonly used tools in hydrologic engineering are *intensity-(or depth)-duration-frequency (IDF or DDF) curves*, and give an overall and consistent picture of rainfall extremes across different accumulation periods. Note that future projections of DDF-curves are recently investigated in a EURO-CORDEX ensemble by Berg et al. (2019).

Our main aim in this paper is to provide a diagnostic framework for RCMs that is partly based on the metrics of Westra et al. (2014) and Cortés-Hernández et al. (2016), but goes beyond in terms of different aspects of extreme value analysis. More specifically, our work provides inference for the statistical performance metrics, including uncertainty estimation and significance testing, and is implemented in freely available R-codes. The new evaluation tools are tested on a small ensemble at EUR-11 and convection-permitting resolutions (H-Res) of climate simulations over Belgium, produced within the context of the CORDEX.be project (Termonia et al. 2018b). Observed precipitation extremes over time scales of 1h-24h were derived by aggregating 10-min pluviograph data. We evaluate the added value of H-Res with respect to EUR-11 simulations.

The paper is organized as follows. Section 2 contains the definition of extremes often used in climate studies. This includes annual maxima, peaks-over-threshold and high quantiles. In Section 3, the extremes definitions are integrated in the evaluation metrics, and statistical models for these metrics are proposed. In Section 4, we describe the data (RCMs and observations) used in this study. The results of the application of the evaluation metrics are discussed in Section 5. Finally, in Section 6, some conclusions are drawn.

2. Extremes definition

a. Extremes in hourly precipitation series

Let $P_t(d)$ be the total amount of precipitation (mm) that has fallen in the time interval $[t-d,t]$ (time unit is per hour). If clock-hourly precipitation series, $P_1(1), P_2(1), \ldots$, are available, d-hourly precipitation is obtained as

$$P_t(d) = \sum_{i=t-d+1}^{t} P_i(1), \qquad t = 1,2,\ldots$$

The analysis mainly uses three types of extreme rainfall definitions:

- Annual maximum (AM):

$$M_n(d) = \max\{P_1(d), P_2(d), \ldots, P_n(d)\}, \qquad \text{with } n \text{ the number of hours per year.}$$

- Peaks-over-threshold (POT):

$$\{P_i(d)\,|\,P_i(d) > u, \text{ for a sufficiently high threshold } u\}.$$

  Note that POT-values regularly appear in clusters, and the series is then declustered by selecting the highest value in each cluster. Willems (2000) recommended to use an interevent time equal to the rainfall duration d, with a minimum value of 12h. Practice shows that the optimal threshold is such that we have an average of 3 to 5 exceedances per year (Coles 2001).

- High τ-quantiles: if *non-zero* d-hourly precipitation is characterized by its distribution function F, the τ-quantile is:

$$Q_\tau := \inf\{x : F(x) > \tau\}.$$

  Of particular importance is the *conditional* τ-quantile, which is conditioned on a certain predictor variable Y: $Q_\tau(y) = \inf\{x : F(x) > \tau \mid Y = y\}$. An example is the modeling of the influence of the temperature/humidity on extreme hourly rainfall, see Sec. 3a. Note that the present quantile-based definition of extremes, relative to wet periods, might be sensitive to changes in the fraction of wet hours (Ban et al. 2015). This does not, however, pose a problem for the present study, which is an evaluation of RCM performance, and not a study on climate change.

There exists a powerful mathematical theory of AM- and POT-extremes, which provides extreme value models and inferential techniques. The statistical modeling of extreme values may be considered as a well-established area of investigation (Leadbetter et al. 1983; Coles 2001; Beirlant et al. 2004). An essential difference with unconditional quantiles is that these extreme value models are able to estimate the probability of events that are more extreme than previously-observed events.

b. Clock-hourly aggregated maxima versus sub-hourly aggregated maxima

The above definition (AM) is concerned with clock-hourly aggregated maxima, but the disadvantage is that they underestimate the “sliding” d-hourly maxima. If sub-hourly series are available (which is the case for the pluviograph stations and some H-Res simulations), it is more convenient to consider sub-hourly aggregated maxima, which are computed as follows. For a series $P_1(\delta), P_2(\delta), \ldots$ with sub-hourly time resolution $\delta = t_i - t_{i-1}$ (e.g. 3-min, 5-min, 10-min, 15min), the d-hourly accumulated precipitation series, $P_1(d)$, $P_2(d), \ldots$, sampled at frequency δ, is calculated as

$$P_t(d) = \sum_{i=t-d/\delta+1}^{t} P_i(\delta).$$

The sub-hourly aggregated maximum is then

$$M^{(\delta)}(d) = \max\{P_1(d), P_2(d), \ldots, P_{n/\delta}(d)\}, \qquad \text{with } n \text{ the number of hours per year.}$$

This is particularly useful for the temporal scaling metric in Sec. 3b, where the scaling (IDF) relationships require a precise approximation of the precipitation maxima.

However, because an important amount of work is devoted to a meaningful comparison at different spatial resolutions (and EUR-11 simulations are at hourly resolution), we mainly limit the study of the other metrics to clock-hourly aggregated extremes.

3. Methods

We have selected a range of metrics that were proposed in the review of Westra et al. (2014), and Cortés-Hernández et al. (2016) to examine the ability of RCMs to reproduce the most important features of extreme precipitation, and tabulated them in Table 1 (first column).

The seasonal cycle is a particularly useful evaluation tool because convectively-driven rainfall extremes occur in Belgium mainly in the warm season, and are typically of short duration (e.g.

Accepted for publication in Journal of Applied Meteorology and Climatology. DOI 10.1175/JAMC-D-21-0004.1.

hourly). On the other hand, longer-duration extreme precipitation events (associated with stratiform weather types) are more likely to occur in autumn and winter. The diurnal cycle is commonly used to prove the triggering physical mechanisms of convective events.

In this section, we propose an overall framework for the statistical modeling of various aspects based on extreme value analysis, as tabulated in Table 1, e.g. extreme rainfall IDF relationships (Van de Vyver 2015a, 2018), spatial extremes (Cooley et al. 2012; Davison et al. 2012), or extremes of dependent time series (Leadbetter et al. 1983; Smith and Weissman 1994).

We compare RCM output at the nearest grid points of observation stations (see Sec. 4 for data description). Evaluation of models against station observations often raise issues of spatial-scale mismatch since the spatially averaged values of the corresponding grid cell values underestimate the point-scale precipitation. The coarser the resolution, the larger the underestimation. However, the daily summer extremes of ALARO at 10 km and 4 km resolution are shown to be in good agreement with the observations (De Troch et al. 2013). In the context of sub-daily summer extremes, Olsson et al. (2015) demonstrated that RCMs at 6 km are nearly unbiased, while a negative bias still exists at 12 km. In any case, the decrease in magnitude is unlikely to affect the evaluation metrics (Cortés-Hernández et al. 2016).

a. Temperature and humidity dependency

Extreme precipitation intensity is known to exponentially increase with daily mean (dew point) temperature at a rate of roughly 7%/°C, the so-called *Clausius-Clapeyron (CC)* rate (Trenberth et al. 2003; Westra et al. 2014). Recent studies using hourly precipitation observations from various locations in western Europe, showed that for temperatures above $\approx 10° - 12°$C, 1h-precipitation extremes increase approximately twice as fast as the CC-rate, a phenomenon that is usually called *super-CC scaling* (Lenderink and van Meijgaard 2008, 2010; Westra et al. 2014).

The aim here is to test to what extent the RCMs reproduce the super-CC scaling features. A statistical methodology was recently proposed to study the scaling of sub-daily rainfall extremes against the (dew point) temperature (Van de Vyver et al. 2019), and is briefly outlined here. Denote by $P(d)$, non-zero d-hourly precipitation and, unless otherwise stated, T the daily-mean dew point temperature. For brevity, we write P instead of $P(d)$.

The CC model – The scaling of high τ-quantiles of $\log(P)$ with the predictor variable T, is linearly modeled by

$$Q_\tau^{CC}(T) = \alpha + \beta T, \tag{1}$$

where the slope β is used to estimate the CC rate.

The CC$^+$ model – The "change-point model" uses piecewise linear quantile regression:

$$Q_\tau^{CC^+}(T) = \begin{cases} \alpha_1 + \beta_1 T & \text{for } T \leq T_c, \\ & \\ \alpha_2 + \beta_2 T & \text{for } T > T_c. \end{cases} \quad \text{with change point } T_c. \tag{2}$$

Similarly, β_1 models the CC scaling rate and, if super-CC scaling behavior is present, β_2 models the super-CC scaling rate. We require that the regression lines are continuous at the change point T_c, which results in the relation: $\alpha_2 = \alpha_1 + (\beta_1 - \beta_2)T_c$. This in turn gives rise to a four-parameter model $(\alpha_1, \beta_1, \beta_2, T_c)$.

It should be noted that the scaling rates and the change point are simultaneously estimated. The inference of the regression models is further explained in Appendix A. In addition, we used a goodness-of-fit (GoF) criterion to examine the influence of the predictor (temperature or dew point temperature) to extreme precipitation.

b. Temporal scaling

A fundamental property of precipitation extremes is temporal scaling invariance, which implies that the statistical properties of extremes across different rainfall durations can be related by a scaling factor. The intention is to investigate to what extent the RCMs are capable of reproducing the correct scaling relationships.

Let $I_t(d) = P_t(d)/d$ (mm h^{-1}) be the average intensity. For a one-year time window $[0, D]$, the maximum of the continuous process $\{I_t(d)\}$ is $M_D(d) = \max\{I_t(d)\,|\,0 \leq t \leq D\}$, which we briefly denote by $M(d)$.

Scaling invariant models include the power law:

$$M(\lambda d) \overset{distr}{=} \lambda^{-\eta} M(d), \qquad \text{with } 0 < \eta < 1. \tag{3}$$

This property is called *simple scaling*, and can be confirmed with linearity in the log-log plot of the statistical moments against d. Although simple scaling is a reasonable working hypothesis, significant departures from the log-log linearity are reported that give room for a generalization to the *multiscaling* concept (Gupta and Waymire 1990; Burlando and Rosso 1996). See Appendix B for details.

Assume that the extreme intensity $M(d)$ follows the generalized extreme value (GEV) distribution (Leadbetter et al. 1983; Coles 2001; Beirlant et al. 2004):

$$\Pr\{M(d) \leq z\} = \exp\left[-\left(1 + \xi(d)\,\frac{z - \mu(d)}{\sigma(d)}\right)_+^{-1/\xi}\right], \tag{4}$$

with $(v)_+ = \max\{0, v\}$, $\mu(d) \in \mathbb{R}$, $\sigma(d) \in \mathbb{R}^+$ and $\xi \in \mathbb{R}$. The location parameter, $\mu(d)$, describes the center of the distribution, the scale parameter, $\sigma(d)$, describes the size of the deviation around the location parameter, and the shape parameter, ξ, presents the tail of the distribution. In particular, the shape parameter ξ is the key parameter in extreme value theory, as it is very sensitive to high return level estimates.

In the case of simple scaling, one easily proves that

$$\mu(d) = d^{-\eta}\mu, \qquad \sigma(d) = d^{-\eta}\sigma, \qquad \text{with} \qquad 0 < \eta < 1, \tag{5}$$

which yields a four-parameter model: μ, σ, ξ are the GEV-parameters of the annual maximum 1h rainfall, and η characterizes the temporal scaling.

It was demonstrated in Van de Vyver (2018) that multiscaling implies that:

$$\mu(d) = d^{-\eta_1}\,\mu, \qquad \sigma(d) = d^{-\eta_2}\,\sigma, \qquad \text{with} \qquad 0 < \eta_1 \le \eta_2 < 1. \tag{6}$$

If $\eta_1 = \eta_2$, the simple scaling model Eq. (5) emerges. Note that Eqs. (5)–(6) determine the form of intensity-duration-frequency (IDF) relationships. The associated IDF-curves are graphical representations (on the log-log scale) of the intensity return levels against rainfall duration d, for a wide range of return periods. The slope of the IDF-curves are determined by the temporal scaling exponents, η or $\eta_{1,2}$.

Bayesian inference and information criterion-based model selection (i.e. AIC and BIC) for temporal scaling GEV-models was developed in Van de Vyver (2015a, 2018), and we refer to these works for the technical details. In particular, it was shown in Van de Vyver (2018) that there is a very strong evidence that the extreme intensities exhibit the multiscaling property, at least for the Belgian pluviograph dataset (see Sec. 4b).

Since classical extreme value theory is based on the assumption that the series under study is stationary (Leadbetter et al. 1983), it is important to examine a possible non-stationary behavior of the rainfall data. When pronounced non-stationarity in the data is present, a non-stationary GEV-model can be considered by introducing time-covariates in the parameters (Coles 2001). For example, the constant μ can be replaced by a time-dependent function $\mu(t) = \mu_0 + \mu_1 Y(t)$, with e.g. $Y(t)$ the global mean temperature or a climate oscillation index at year t. To our knowledge, Ouarda et al. (2018) presented the only non-stationary extension to date of the temporal scaling GEV-model, Eq. (5). Although, their IDF models will provide a better fit for non-stationary data, they are of limited use for the current validation framework because the temporal scaling parameter

η is the crucial metric here, and they considered it as constant since η does not show any trend in their data. To the best of our knowledge non-stationary IDF-relationships with a time-varying scaling parameter does not exist so far. Other non-stationary IDF-curves are also developed in Cheng and AghaKouchak (2014) and in Agilan and Umamahesh (2016), but they fitted a non-stationary GEV-distribution to the individual durations, without using a scaling relationship.

c. Spatial structure

The aim is to examine the spatial structure by a dependence measure for spatial extremes, the *madogram* (Cooley et al. 2006; Vannitsem and Naveau 2007; Naveau et al. 2009), which is the first moment version of the well-known variogram for general non-extreme events. The statistics of spatial extremes naturally extend the classical univariate extreme value models, and also include basic concepts of Gaussian geostatistics. A popular approach for modeling the maxima observed at different sites, is based on fitting *max-stable processes* (Davison and Gholamrezaee 2011; Davison et al. 2012; Cooley et al. 2012). Specifically, a max-stable process $Z(\mathbf{x})$ at location $\mathbf{x}$ has the GEV distribution (F) throughout the spatial domain, includes flexible correlation functions, and can therefore be thought of as the "extreme value analog" of Gaussian random fields.

We assume that the spatial process $Z(\mathbf{x})$ under study is stationary and isotropic, which means that the spatial dependence between $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$ depends only on the scalar distance $h = ||\mathbf{h}||$. The *spatial madogram* is

$$\nu_F(h) = \frac{1}{2}\mathbb{E}|F(Z(\mathbf{x})) - F(Z(\mathbf{x}+\mathbf{h}))|, \qquad \text{with } ||\mathbf{h}|| = h. \tag{7}$$

If $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$ are independent, we have $\nu_F(h) = 1/6$, and $\nu_F(h) < 1/6$ otherwise.

Let $z_t(x_i)$ denotes the annual maximum of year t, observed at location x_i, and assume that the K-year time series $z_1(x_i),\ldots,z_K(x_i)$ at each location follows the GEV-distribution, $F_i :=$

$\text{GEV}[\mu_i, \sigma_i, \xi_i]$. In the case that the extremes are non-stationary over time, a non-stationary GEV-distribution has to be introduced, i.e. $z_t(x_i) \sim \text{GEV}[\mu_i(t), \sigma_i(t), \xi_i(t)]$. A binning empirical madogram is

$$\hat{\nu}_F(h) = \frac{1}{K} \sum_{t=1}^{K} \frac{1}{2|\mathcal{N}_h|} \sum_{(x_i, x_j) \in \mathcal{N}_h} |F_i(z_t(x_i)) - F_j(z_t(x_j))|, \tag{8}$$

where $\mathcal{N}_h$ is the set of all pairs for which $\|x_i - x_j\| \in [h - \delta, h + \delta]$. The asymptotic probability distribution of the madogram estimator, as given in Cooley (2005), allows for the computation of confidence intervals. It should be noted that for RCMs, the distances between grid points are considered which, depending on the spatial resolution, may differ slightly from the distances between the stations.

The use of spatial statistical measures for the validation of RCMs is very limited. To our knowledge, the only applications can be found in (Cortés-Hernández et al. 2016; Hobæk Haff et al. 2015; Rasmussen et al. 2012), but there are some important differences with the present methodology. Firstly, Rasmussen et al. (2012) and Hobæk Haff et al. (2015) modeled spatial dependence over the full range of the precipitation distribution, without explicitly accounting for extremes. By contrast, the madogram focuses solely on the extremal spatial dependence, without regard to the main body of the precipitation distribution. Secondly, the binned madogram-estimates provide a local average that gives a clear view on the extremal dependence as a function of the distance, which is not the case with the pairwise extremal coefficient approach in Cortés-Hernández et al. (2016).

d. Temporal clustering

Since extreme precipitation events have the tendency to occur in temporal clusters, we examine how well RCMs reproduce the period of clustering. We introduce a measure of short-term temporal dependence for extremes in a series.

Consider a time series of discrete time stationary process $X_1, \ldots, X_n$, with distribution F. The distribution of the maximum is

$$G(x) = \Pr\{\max(X_1, \ldots, X_n) \leq x\}.$$

For an independent process, we have simply: $G(x) = F(x)^n$. Under fairly wide conditions (that is, the so-called mixing conditions), the following approximation can shown to be hold for dependent series:

$$G(x) = F(x)^{n\theta}, \quad \text{where } 0 < \theta \leq 1,$$

where θ is called the *extremal index*, which measures the degree of clustering of extremes. Observe that for independent X_i's, then trivially, the series has $\theta = 1$. For a more precise and technical description of the theory, we refer to (Leadbetter et al. 1983). An interesting interpretation of the extremal index θ is that it can be approximated by the inverse of the mean cluster size. It is therefore natural to estimate θ^{-1} as the cluster size in the observations. Clusters of high threshold exceedances are identified with a separation time r. For a series of length m, the *runs estimator* (Smith and Weissman 1994) is defined as

$$\hat{\theta} = N_e / N_c, \tag{9}$$

where N_e is the number of exceedances of the threshold u, and N_c is the number of clusters, i.e.

$$N_e = \sum_{i=1}^{m} W_i, \qquad N_c = \sum_{i=1}^{m} W_i\,(1 - W_{i+1})\,(1 - W_{i+r}), \qquad \text{where } W_i = \begin{cases} 1, & \text{if } X_i > u, \\ 0, & \text{if } X_i \leq u. \end{cases}$$

The main advantages of the runs estimator are the simplicity and the easy interpretability, it has a low bias and it is a nonparametric method. There are numerous other (more complicated) estimators for the extremal index, and research for new estimators is still ongoing; an overview can be found in Chavez-Demoulin and Davison (2012).

4. Model and observation data

a. Regional climate models

One of the goals of the CORDEX.be project was to contribute to EURO-CORDEX with three RCMs (Termonia et al. 2018b). Four Belgian climate modeling groups provided climate simulations for the EUR-11 domain corresponding to a 12.5 km horizontal resolution, in line with the EURO-CORDEX prescriptions. Additional to the 540 simulation years for the EUR-11 domain, 780 simulation years at convection-permitting resolution (H-Res) were produced, where the horizontal resolution ranges from 2.8 to 5 km (see Table 4 in Termonia et al. (2018b)). The EUR-11 and H-Res runs are performed on a limited geographical domain using a one-way nesting approach, i.e. by imposing meteorological conditions at the boundaries from model simulations at lower resolution. To validate the EUR-11 simulations, we consider the evaluation runs which use ERA-Interim as lateral boundary conditions over the EURO-CORDEX region. In turn, these simulations are used as boundary conditions for a second nesting over Belgium at convection-permitting scale. Precipitation is typically stored with an hourly frequency for the EUR-11 simulations, whereas the H-Res simulations are archived at sub-hourly frequency. There are four models used: ALARO-0, COSMO-CLM (two versions) and MAR. The MAR simulations are only run at H-Res resolution. The evaluation runs cover the period starting from 1979 up to 2009, 2010, 2014 or 2017 (according to the model). In Table 2 we provide their main features.

Prein et al. (2015) define a “convection-permitting model” as featuring a horizontal resolution of less than 4 km, and without a deep-convection parameterization scheme. Therefore, strictly speaking, only the H-Res simulations of COSMO-CLM qualify.

1) ALARO-0

The ALARO model (version ALARO-0) is a hydrostatic RCM which is based on ALADIN, a numerical weather prediction system developed by the international ALADIN consortium for operational weather forecasting and research purposes (Bubnová et al. 1995; Termonia et al. 2018a). The ALADIN model is the limited area model (LAM) version of the "Action de Recherche Petite Echelle Grande Echelle Integrated Forecast System" (ARPEGE-IFS). The ALARO model includes the precipitation and cloud scheme "Modular Multiscale Microphysics and Transport" (3MT), which is a parameterization of deep convection optimized for resolutions in the so-called grey zone (4km-10km). The main strength is scale awareness, i.e. the parameterization itself works out which processes are unresolved at the resolution used (Gerard et al. 2009). This opposes more conventional parameterization practices which either enable or disable parametrization schemes. This allows 3MT to generate consistent results across spatial scales, as shown by De Troch et al. (2013).

2) CCLM

The COSMO-CLM (CCLM) model is based on the COSMO model (Consortium for Small-scale Modeling), which is a non-hydrostatic limited area model of the Deutsche Wetterdienst (DWD) for operational purposes. The limited-area modeling (CLM) community implemented the COSMO model for climate simulations (Rockel et al. 2008). The EUR-11 simulation does not explicitly resolve convection and uses the Tiedtke convection scheme, while the H-RES simulation dynamically resolves deep convection. More model settings and technical recommendations for simulations at convection-permitting scale are given in Brisson et al. (2016a), and the added value was confirmed in Brisson et al. (2016b).

In this work we include two versions of CCLM:

- CCLM-UCL. A two-moment scheme with hail parameterization was implemented by Van Weverberg et al. (2014).

- CCLM-KUL. The computationally efficient urban land-surface scheme TERRA_URB was implemented by Wouters et al. (2016).

3) MAR

The MAR (Modèle Atmosphérique Régional) RCM is a hydrostatic primitive equation model. The atmospheric part of MAR is fully explained in Gallée and Schayes (1994), and the surface vegetation-atmosphere interface (called SISVAT for Soil Ice Snow Vegetation Atmosphere Transfer) is described in De Ridder and Gallée (1998). Although being originally developed for the polar regions, MAR was recently adapted and applied to the temperate climate of Belgium (Wyard et al. 2017). In this paper, MAR version 3.9 is used, which explicitly resolves a large part of precipitation (98%) at 5km-resolution, without a convective scheme. In future developments of the MAR model, more account must be taken of convective precipitation.

b. Observations

The observational dataset comprises 18 Belgian pluviograph series with 10-min precipitation, and cover the period of 1967–2004. This data was obtained by Hellmann–Fuess pluviographs, that were part of the hydro-meteorological network of the Royal Meteorological Institute (RMI) of Belgium (Fig. 1 and Table S1). The observation period does not entirely overlap with that of the RCM evaluation runs for different reasons. First, there is no clear trend in the extremes (see Sec. 4.c). Second, we are not looking at the one-to-one time correspondence between model and observations. Last but not least, it is unlikely that the evaluation metrics, related with the underlying physics of convective events, differ for the non-overlapping periods. As extremes are, by definition, rare, as much data as possible is used for a reliable inference.

For each station, we extracted the d-hourly precipitation annual maxima, for rainfall durations $d = 1, 2, 6, 12, 24, 48, 72$h. For each duration, we consider an annual maximum value as "missing" if (a) it is below the 40th-percentile of the annual maximum series, and (b) the number of missing values of that year is larger than one third. Finally, we consider a particular year as missing if at least three of the seven durations are missing.

The same data has been included in previous studies concerning spatial extremal dependence (Vannitsem and Naveau 2007), trends in historical extremes (Ntegeka and Willems 2008), spatial extreme value models (Van de Vyver 2012), sliding 24-h maxima (Van de Vyver 2015b), IDF-relationships (Willems 2000; Van de Vyver 2015a, 2018) and climate model validation (Tabari et al. 2016).

Satellite-based precipitation products, such as MSWEP (Beck et al. 2019) and GPM (Hou et al. 2014), provide complete spatial coverage and can be potentially interesting for the analysis of spatial extreme rainfall. However, none of these products have a sufficiently fine temporal resolution and / or a sufficiently long common evaluation period.

c. Trend testing

Most statistical methods assume that the series under investigation are stationary, but in reality this is usually not the case. To investigate a possible non-stationarity in the annual maximum series, we performed the Mann-Kendall monotonic trend test (Chandler et al. 2011), at the 5% significance level (Table S2). The d-hourly maximum time series are found to be sufficiently stationary. Indeed, among all RCM models and observations, at most 20% of the station- or grid-locations are found to feature (significant) increasing trends.

5. Results and discussion

We first illustrate that systematic and significant differences exist between observed and RCM-simulated rainfall extremes. We compute the bias as the difference between the modeled and observed mean annual maxima at each location (18 in total). Fig. 2 shows a boxplot of the bias, for different durations and spatial resolutions. None of the RCMs reproduced the observed extremes at all durations. In particular, the 1h- and 2h-extremes are poorly represented by the EUR-11 simulations, and are underestimated at almost every station. This is partly due to the scale mismatch between EUR-11 resolution and the point-observations. The H-Res simulations show better performance for these short-duration extremes. On average, ALARO also performs better for longer-duration extremes, but the spatial variability in the estimations is higher. On the other hand, CCLM tends to overestimate the longer-duration extremes. The Kolmogorov-Smirnov test (Fig. 3) confirms that the difference between the modeled and observed annual maximum distribution is statistically significant in many cases.

Closely related to this preliminary result, Fosser et al. (2015) and Kendon et al. (2017) reported that convection-permitting models did not improve the simulations of daily mean precipitation compared with large-scale climate models, although they provided more accurate simulations of heavy hourly precipitation. In the context of extremes, this was also observed in Tabari et al. (2016).

A brief overview of the overall conclusions for the different evaluation metrics results is given in Table 6, which we will discuss in more detail here.

a. Seasonality

We considered the AM- and POT-series and we recorded the seasons when the extreme events occurred, see Fig. 4 (EUR-11) and Fig. 5 (H-Res) for the AM-series. Similarly, Figs. S1-S2 show the results for the POT-series, which consist of cluster maxima of excesses above the 0.99-quantile (the quantile computation is considered for non-zero rainfall, viz. ≥ 0.1mm). We observe that the seasonal cycle was very well reproduced by all models and all seasons. Notable exceptions can be found for MAR due to strongly underestimated rainfall extremes in summer and large overestimations in winter. The results of the hourly summer POT-extremes of the CCLM-models are a bit improved when using a higher spatial resolution (H-Res).

b. Diurnal cycle

Fig. 6 and Fig. S3 show the discrete probability distributions of the time of occurrence (UTC) of the AM- and POT hourly rainfall, respectively, and are further referred to as the diurnal cycle. In addition to these graphs, we have calculated the Kullback-Leibler divergence (see Appendix C) to measure how close the observed and simulated diurnal cycles are to each other. The results are shown in Table 3. The diurnal cycle of the EUR-11 simulations is poorly represented by all models, except ALARO. Using a higher spatial resolution strongly improved the results for the CCLM-models. The conclusion is about the same for AM- and POT-extremes, except that the peaks in the observed daily cycle are less pronounced for POT than for AM.

c. Temperature and humidity dependency

In order to reduce the uncertainty in the estimation, the inference of the extreme precipitation/(dew point) temperature-relationship is based on a joint evaluation of all the locations in the study region (15 in total). Fig. 7 shows the binned high-quantiles of hourly precipitation against daily mean dew point temperature, for observations and RCMs at both spatial resolutions. The application of the quantile regression models is demonstrated in Tables S3-S4, for the following predictors: air temperature and dew point temperature, respectively. The estimations, together with the 95%-

confidence interval, are plotted in Figs. S4-S5 for the 0.95- and the 0.99-quantiles (predictor: dew point temperature). All RCMs except MAR are able to reproduce the super-CC scaling behaviour. The 0.95-quantiles of the EUR-11 simulations (Fig. 7) have a CC-scaling rate (β_1) that is significantly smaller than that of the observations (around 4% against the observed 6-7%). The H-Res simulations have disagreeing values for β_1 equal to around 4%, 6% and 7.5%. For ALARO, again the difference in β_1 between both spatial resolutions is less pronounced. For the 0.99-quantile (see Fig. S5), the β_1-values for observations and RCMs agree well generally, and the other parameters (β_2 and T_c) are satisfactorily reproduced by the high resolution models. Note that Figs. S4-S5 show estimates for which there are relatively large differences between EUR-11 and H-Res, but that the confidence intervals are so large that it cannot be concluded that these differences are statistically significant.

Next, the predictive skill of the (dew point) temperature was compared by means of the goodness-of-fit (GoF) criterion, see Fig. 8. Most importantly, the influence of the predictors is of the same order of magnitude as in the observations for all the models, except for MAR. Increasing the model resolution leads to increased predictive skill for CCLM but slightly decreased skill for ALARO.

d. Temporal scaling

1) Scaling invariance of simulated extremes

First, we checked whether the temporal scaling GEV-models, Eq. (5)-(6) , are suitable for the simulated extreme intensities, where we relied on the log-log linearity of the GEV-parameters (location and scale) with duration d, and in which the slopes correspond to the temporal scaling exponents (i.e. η or $\eta_{1,2}$). As said in Sec. 2b, we considered the clock-hourly aggregated maxima of the EUR-11 simulations, whereas for the observations and the H-Res simulations, the sub-hourly aggregated maxima are used. In Fig. 9, we have plotted the maximum likelihood estimators $(d_i, \hat{\mu}_i)$

and $(d_i, \hat{\sigma}_i)$ for station-location Uccle, at various rainfall durations $d = 1\text{h}, \ldots, 72\text{h}$. The scaling property is clearly visible for the observations and the H-Res simulations, except for MAR. As expected, the clock-hourly aggregated maxima of the EUR-11 simulations conform to the log-log linearity for the range $d = 6\text{h}, \ldots, 72\text{h}$, but not for the first six hours ($d = 1\text{h}, \ldots, 6\text{h}$).

Having verified the log-log linearity of the GEV-parameters, we investigated further which scaling model is the most likely for extreme intensities. Two model selection procedures were used: (i) based on AIC, and (ii) based on Bayesian hypothesis testing. The latter involves the zero hypothesis, $\mathcal{H}_0 : \eta_1 = \eta_2$ (simple scaling), and the alternative hypothesis $\mathcal{H}_1 : \eta_1 \neq \eta_2$ (multiscaling). Next, we evaluated the relative evidence of $\mathcal{H}_0$ over $\mathcal{H}_1$ with the posterior odd $R = p(\mathcal{H}_0 \,|\, y)/p(\mathcal{H}_1 \,|\, y)$, with y being the data. Technical details are further provided in Appendix D.a. In Table 4, we tabulate the number of stations for which the multiscaling hypothesis was selected on the basis of AIC or $R < 1$. To have an overall view of the hypothesis test, we consider the spatial product of R of the different locations. The MAR-model has been excluded from the table since the plots of the associated GEV-parameters against d strongly deviate from log-log linearity (Fig. 9).

Demonstrating the statistical significance of the multiscaling property is complicated here for several reasons: (i) the subjective choice of the number of series that should pass a particular test, (ii) different tests may give different results, and (iii) different time series lengths. However, instead of this determination, the aim here is to validate the models and therefore we limit ourselves to comparing the percentages of the RCMs with those of the observations. In general, we can observe that the RCM-simulated extremes (except MAR) broadly share the multiscaling property of the observed extremes.

2) Scaling model parameters for observed/simulated extremes: a comparison

We investigate to what extent the temporal scaling GEV-parameters of the simulations correspond to the observed ones. Recall that the model consists of two types of parameters: (i) the GEV-parameters of the 1h-extremes, (μ, σ, ξ), where μ and σ describe the main body of the GEV-distribution, and ξ describes the tail of the distribution, (ii) the temporal scaling parameters, η or $\eta_{1,2}$.

For each location, we plot the posterior mean and the 95%-credible intervals of the shape parameter, ξ, and the temporal scaling parameters, $\eta_{1,2}$ in Fig. S6 (EUR-11), Fig. S7 (H-Res) and Fig. S8 for the corresponding IDF-curves. We have chosen to display ξ because this parameter is critical for the estimation of high return levels. It can be already seen that for ALARO, the $\eta_{1,2}$-values are systematically underestimated, a features which is also retrieved below.

The significance of the parameter-preservation by the RCMs was again tested with AIC and Bayesian Hypothesis testing (Table 5). For example, regarding the shape parameter ξ, the hypotheses are formulated as $\mathcal{H}_0 : \xi^{(obs)} \neq \xi^{(mod)}$, and $\mathcal{H}_1 : \xi^{(obs)} = \xi^{(mod)}$, where $\xi^{(obs)}$ and $\xi^{(mod)}$ are the ξ-values corresponding to observed and RCM-simulated extremes. See Appendix D.b for the details of the computation of the posterior odd $R = p(\mathcal{H}_0 \,|\, y)/p(\mathcal{H}_1 \,|\, y)$. As somewhat expected, we encounter significant differences between observations and models. Therefore, in addition to the strict condition $R < 1$, we also set a weakened condition, $R < 3$. The spatial product of R is particularly large, but bear in mind that R-based tests are strongly penalizing the outliers such that R should rather be considered as a probability-based measure for making a meaningful comparative analysis.

From Table 5, it can be seen that the shape parameter ξ is generally one of the best reproduced parameters. Although Fig. S6-S7 seems to reveal differences in the ξ-values between models and with observations, it was found here that the differences are largely not statistically significant, highlighting the importance of a statistical framework for model validation. The scale parameter σ, on the other hand, was reproduced the least well, by all RCMs and for each spatial scale. Since σ describes the size of the deviation around the location parameter μ, we may conclude that the variability of the 1h-extremes is therefore not very well captured by the RCMs. The EUR-11 simulations represented the temporal scaling parameters $\eta_{1,2}$ better than the GEV-parameters concerning the moderate 1h-extremes, i.e. (μ, σ). For the H-Res simulations, the same is true for CCLM, while this is not immediately clear to ALARO.

The effect of moving to a finer spatial scale is remarkably different between ALARO and CCLM. The benefit for ALARO is mainly the relative good presentation of the μ-parameter, which means that the 1h-extremes are well simulated to some extent. This agrees with the preliminary analysis of the bias in the annual maxima (Fig. 2). The temporal scaling parameters $\eta_{1,2}$ are not as well represented as CCLM (for both spatial scales). As mentioned earlier, the $\eta_{1,2}$-values are indeed slightly underestimated by ALARO, and consequently the resulting IDF-curves are generally less steep. Opposed to ALARO, CCLM does not simulate the 1h-extremes very well, but it has the advantage that the temporal scaling parameters $\eta_{1,2}$ are better reproduced.

e. Spatial structure

Fig. 10 shows the spatial madogram, Eq. (7), for 1-, 6-, 12- and 24-h extremes of observations and ALARO at both spatial resolutions (EUR-11 and H-Res). The madogram estimations of the observed 1h-extremes agree well with a similar study on the same dataset (Vannitsem and Naveau 2007). We first note that the empirical estimates may be higher than the theoretical upper-bound, i.e. $\hat{\nu}_F(h) > 1/6$. Employing a parametric model of the madogram would be advantageous in this case, but the known max-stable processes provided a poor fit to the data. The main finding is that the confidence intervals are relatively large, and this makes it difficult to draw meaningful conclusions concerning systematic differences between RCMs and observations.

The decorrelation length has to increase for longer rainfall durations, because the short-duration extremes are more likely to be caused by small-scale convective events, while longer-duration events are associated with large-scale precipitation patterns (Westra et al. 2014). For distances larger than 50 km, it can be seen in Fig. 10 that the red horizontal line (i.e. the independent case, $\nu_F = 1/6$) falls in the 95%-confidence interval of the observed madogram for d = 1h and d = 6h. Similarly, the decorrelation lengths for d = 12h and d = 24h are higher in comparison with d = 1h.

It is clear that strong spatial correlations in the climate model simulated extremes are present within distances smaller than 50 km, and that the confidence intervals of the observed and simulated madograms do not overlap that much within this distance. This is consistent with the known overestimation of the spatial extent of rainfall events by RCMs (Maraun et al. 2010; Hobæk Haff et al. 2015). Using a finer spatial resolution only slightly reduced the correlations, but the overlap of the confidence intervals for the different resolutions is fairly small. Similar results were obtained for the other RCMs (not shown).

f. Temporal clustering

In Fig. 11, we plot the mean cluster size θ^{-1}, estimated with the runs estimator Eq. (9), as a function of a high threshold. Series of d-hourly precipitation were considered, for d = 1, 3, 6, 12h, and clusters of extremes were identified. The main conclusion is that hourly EUR-11 extremes gather in clusters of larger temporal extent than the observed extremes. The bias in the cluster size tends to be smaller for higher precipitation durations d, except for ALARO at d > 6h. There is little improvement for hourly H-Res extremes, but the benefit for ALARO is threshold-dependent. For d > 6h, ALARO did not produce improved cluster sizes when using a finer spatial resolution.

The temporal clustering in MAR is too high and features biases up to 20%, in particular for short durations, but the difference with the observations decreases for increasing d-values.

6. Conclusions

In this paper, we applied statistically-based metrics to evaluate whether RCMs (non-convection and convection-permitting) capture the spatio-temporal characteristics of heavy sub-daily rainfall events. The metrics provide a general picture of the RCM's performance, in contrast to the current evaluation strategy, which usually only compares observed and modeled high quantiles of the rainfall distribution. The metrics can also be useful in the context of climate change impact modeling to determine whether extreme precipitation simulations can be used directly as input for a specific impact model.

Since these metrics are usually calculated empirically from the data, they may suffer from large uncertainties. We attempted to overcome this problem by assuming various statistical extreme value models for a range of aspects of the performance metrics. Our strategy includes a better inference of the metrics such as uncertainty quantification and model selection, which could indicate whether RCMs differ significantly or not.

Our main conclusions are:

- The EUR-11 simulations poorly reproduced the hourly extremes, while the longer-duration extremes ($d \geq 6$h) were much better presented. Conversely, the hourly extremes were well simulated at finer spatial resolution, H-Res, but only one model succeeded to reproduce satisfactorily the longer-duration extremes. Consequently, the H-Res simulations are particularly useful for impact studies related with short-duration extreme rainfall events, such as local urban flooding risk assessment. On the other hand, there is no spatial resolution that consistently improved the simulation of extreme rainfall events across all the durations, so that it is not clear which spatial resolution is the most advantageous for an IDF analysis.

- The mean cluster length of extreme values in the hourly EUR-11 simulations is overestimated. Using a finer spatial resolution improves this estimate significantly.

- The spatial clustering is overestimated by every RCM, at all durations. Using a finer resolution only slightly improved the spatial dependence estimation. We conclude that the RCMs may be not suited for flood risk estimation over a catchment area, which often requires information of the joint probability between extreme rainfall at multiple sites (Westra et al. 2014).

- Except for the spatial extremes aspects, there is an acceptable performance of ALARO and the CCLM-models in terms of the physically meaningful metrics. In contrast, MAR is not able to correctly represent extreme rainfall in sub-daily resolution because it lacks a parametrized convective scheme.

The evaluation framework for simulated rainfall extremes has been demonstrated on a limited number of RCMs, but the methodology can be easily applied to a large ensemble of RCMs to arrive at a more comprehensive conclusion (Ban et al. 2021).

This study can be extended in various ways. First, the application of extreme value theory could be further explored for metrics such as the seasonal/diurnal cycle. An example of statistical modeling of the annual cycle of extreme 1-day precipitation events was proposed in Maraun et al. (2009) and Schindler et al. (2012), where GEV-distribution parameters of monthly maxima were modeled by oscillating functions with a 1-year period. Future research should indicate if the method can be extended to sub-daily precipitation extremes. Second, evaluation metrics could be involved to examine the similarity of the synoptic situation in observed and simulated extreme events (Westra et al. 2014). An example of a GEV-model to investigate the influence of the synoptic scale atmospheric circulation on extreme precipitation, was given in Maraun et al. (2012).

Finally, future research should aim to search for the most appropriate correction of future IDF relationships.

Acknowledgments. This work is supported by URCLIM and has received funding from EU's H2020 Research and Innovation Program under Grant Agreement 690462. The CORDEX.be project was financially supported by the Belgian Science Policy (BELSPO) under Contract No. BR/143/A2. The computational resources and services for the ALARO-0 regional climate simulations were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government department EWI.

We thank Dr. Sam Vanden Broucke for providing the model output of CCLM-KUL within the CORDEX.be project.

Data availability statement. The hourly and sub-hourly RCM simulations used in this study are openly available from the World Data Center for Climate (WDCC):

- ALARO-0 (https://doi.org/10.26050/WDCC/CORDEX.be_RMIB-UGent_ALARO-0)
- CCLM-UCL (https://doi.org/10.26050/WDCC/CORDEX.be_UCLouvain_CCLM6-0-6)
- CCLM-KUL (https://doi.org/10.26050/WDCC/CORDEX.be_KULeuven_CCLM5-0-6)
- MAR (https://doi.org/10.26050/WDCC/CORDEX.be_ULiege_MAR39)

The observed sub-daily precipitation extremes are openly available from Zenodo (http://doi.org/10.5281/zenodo.4741178).

The code availability is listed in Table 1. The *"Temperature and humidity dependency"* metric code, associated to the quantile regression models in Van de Vyver et al. (2019), is written in R and is available via Zenodo (http://doi.org/10.5281/zenodo.4644567). The *"Temporal scaling"* metric code, associated to the temporal scaling GEV-models and IDF statistics in Van de Vyver (2015a, 2018), is written in R and is available via Zenodo (http://doi.org/10.5281/zenodo.4644184). The code for the other metrics, *"Spatial structure"* and *"Temporal clustering"*, are not written by the authors and are available via existing CRAN packages.

APPENDIX A

Quantile regression

The methods in the following text are employed in Van de Vyver et al. (2019).

The quantile regression models are fitted to the set of n data pairs (T_i, P_i) by minimization the functions (Koenker 2005):

$$\begin{aligned} &\textit{CC model:} \quad \hat{S}^{CC} = \min_{(\alpha,\beta)} \sum_{i=1}^{n} \rho_\tau\big(\log(P_i) - Q_\tau^{CC}(T_i)\big), \\ &\textit{CC}^{+}\textit{ model:} \quad \hat{S}^{CC^+} = \min_{(\alpha_1,\beta_1,\beta_2,T_c)} \sum_{i=1}^{n} \rho_\tau\big(\log(P_i) - Q_\tau^{CC^+}(T_i)\big), \end{aligned} \tag{A1}$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}_{\{u<0\}})$ is called the loss function. Estimation of the CC model can be done using R-package "quantreg" (Koenker 2018; Wasko and Sharma 2014). A practical estimation procedure consists in minimizing over a range of T_c-values, and then selecting the T_c-value at which the minimum is achieved.

Choosing between the linear and the piecewise linear quantile regression model can be done with the Bayesian information criterion (BIC):

$$\mathrm{BIC} = -2 \log \hat{\mathcal{L}} + \ln(n)\, p, \tag{A2}$$

where $\hat{\mathcal{L}}$ is the associated maximized likelihood function of the quantile regression model, p is the number of parameters, and n, the data size. The model with the lowest BIC value is selected.

BIC-based model selection can be seen as a trade-off between goodness-of-fit and the simplicity of the model.

The prediction strength of T to the precipitation data can be measured by the relative success of a quantile regression model against the unconditional quantiles, $Q_\tau^{(0)}$. The latter is computed by minimizing the function:

$$\hat{\mathcal{S}}_0 = \min_{Q_\tau^{(0)}} \sum_{i=1}^{n} \rho_\tau \left(\log(P_i) - Q_\tau^{(0)}\right). \tag{A3}$$

Koenker and Machado (1999) defined the goodness-of-fit criterion for a particular τ-quantile as:

$$\mathrm{GoF} = 1 - \hat{\mathcal{S}} / \hat{\mathcal{S}}_0, \tag{A4}$$

where $\hat{\mathcal{S}}$ stands either for $\hat{\mathcal{S}}^{CC}$ or $\hat{\mathcal{S}}^{CC^+}$. GoF ranges between 0 and 1 and, the closer GoF is to 1, the higher the influence of T to the precipitation extremes.

APPENDIX B

Simple scaling versus multiscaling

Eq. (3) implies that the raw q-order moments obey the power law

$$E[M^q(\lambda d)] = \lambda^{-\alpha_q} E[M^q(d)], \qquad \text{with } \alpha_q = q\eta, \tag{B1}$$

which is called *wide sense simple scaling* (Gupta and Waymire 1990). It is often considered as a suitable working assumption in IDF-analysis, despite the fact that deviations from simple scaling are to be expected. Accordingly, Gupta and Waymire (1990) defined multiscaling as

$$E[M^q(\lambda d)] = \lambda^{-\alpha_q} E[M^q(d)], \qquad \text{with } \alpha_q = q\,\varphi_q\,\eta, \tag{B2}$$

where φ_q is called the dissipation function, and describes the departure from simple scaling.

APPENDIX C

Kullback-Leibler divergence

The Kullback-Leibler divergence measures how close two probability distributions are from each other. For two discrete probability density distributions f and g, the Kullback-Leibler divergence from g to f is defined as:

$$\mathrm{KL}(f;g) = \sum_t f(t)\,\log\left(\frac{f(t)}{g(t)}\right). \tag{C1}$$

APPENDIX D

Bayesian hypothesis testing

Since the previous works (Van de Vyver 2015a, 2018) on temporal scaling GEV-models were implemented in a Bayesian framework, some useful hypothesis tests must be translated into the Bayesian setting (Shikano 2019). In general, we set up two distinct hypotheses: a zero hypothesis $\mathcal{H}_0$, and an alternative hypothesis $\mathcal{H}_1$. Given the data y, the selection between the hypotheses relies on the ratio of posterior densities

$$R = \frac{p(\mathcal{H}_0\,|\,y)}{p(\mathcal{H}_1\,|\,y)}. \tag{D1}$$

The case $R \leq 1$ means that $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$. From the Bayesian rule, R can be computed as

$$\frac{p(\mathcal{H}_0\,|\,y)}{p(\mathcal{H}_1\,|\,y)} = \frac{p(y\,|\,\mathcal{H}_0)}{p(y\,|\,\mathcal{H}_1)}\,\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}, \tag{D2}$$

where it is commonly assumed that the priors $p(\mathcal{H}_0) = p(\mathcal{H}_1)$. The ratio $p(y\,|\,\mathcal{H}_0)/p(y\,|\,\mathcal{H}_1)$ in Eq. (D2) is called the *Bayes Factor* or BF, and can be obtained as follows

$$p(y\,|\,\mathcal{H}_0) = \int_{\Psi_{\mathcal{H}_0}} p(y\,|\,\psi_{\mathcal{H}_0})\,p(\psi_{\mathcal{H}_0}\,|\,\mathcal{H}_0)\,d\psi_{\mathcal{H}_0}, \tag{D3}$$

where $\psi_{\mathcal{H}_0}$ are the parameters of the model associated with $\mathcal{H}_0$. The integral, Eq. (D3), is approximated during the Markov chain Monte Carlo (MCMC) simulation in the Bayesian inference of the model. A similar approach holds for $\mathcal{H}_1$.

Two applications are given below.

a. Choosing between temporal scaling GEV-models

We consider here hypothesis testing as model selection, i.e. choosing between the simple scaling and the multiscaling GEV-models, Eqs. (5)–(6). We thus define:

- $\mathcal{H}_0$: $\eta_1 = \eta_2$ (simple scaling), with associated model parameters $\psi_{\mathcal{H}_0} = (\mu, \sigma, \xi, \eta)$.

- $\mathcal{H}_1$: $\eta_1 \neq \eta_2$ (multiscaling), with associated model parameters $\psi_{\mathcal{H}_1} = (\mu, \sigma, \xi, \eta_1, \eta_2)$.

The data used is the matrix of annual maximum intensities $\mathbf{i} = (i_j(d_k)) = (i_{jk})$, where $j = 1, \ldots, N$ refers to the year, and d_k, $k = 1, \ldots, M$ to the rainfall duration. Under the assumption that the maximum intensities are independent of each other, the conditional probability $p(y \,|\, \psi_{\mathcal{H}_0})$ in Eq. (D3) is equal to the "independence" likelihood (Van de Vyver 2015a, 2018):

$$L_{ind}(\mathbf{i} \,|\, \psi_{\mathcal{H}_0}) = \prod_{j=1}^{N} \prod_{k=1}^{M} g(i_{jk} \,|\, \psi_{\mathcal{H}_0}), \tag{D4}$$

where g is the density function of the simple scaling GEV$[\mu(d), \sigma(d), \xi]$. A similar approach holds for $p(y \,|\, \theta_{\mathcal{H}_1})$.

b. Testing for equality

We perform hypothesis testing to investigate the plausibility that a certain statistical parameter is different for observations and RCMs. Denote by $\psi^{(obs)} = (\mu^{(obs)}, \sigma^{(obs)}, \xi^{(obs)}, \eta^{(obs)})$ and $\psi^{(mod)} = (\mu^{(mod)}, \sigma^{(mod)}, \xi^{(mod)}, \eta^{(mod)})$, the set of parameters of the simple scaling GEV-model for the observed and RCM-simulated extremes, respectively. The data used is made of the observed and simulated extremes, $\mathbf{i}^{(obs)}$ and $\mathbf{i}^{(mod)}$.

To check the equality of a certain parameter, $\psi_j^{(obs)} = \psi_j^{(mod)}$, the hypothesis testing scenario is

- $\mathcal{H}_0$: $\psi_j^{(obs)} \neq \psi_j^{(mod)}$, with associated model parameters $\psi_{\mathcal{H}_0} = (\psi^{(obs)}, \psi^{(mod)})$.

- $\mathcal{H}_1$: $\psi_j^{(obs)} = \psi_j^{(mod)} =: \psi_j$, with associated model parameters $\psi_{\mathcal{H}_1} = (\psi_{-j}^{(obs)}, \psi_j, \psi_{-j}^{(mod)})$, where $\psi_{-j}^{(\cdot)}$ is the parameter vector $\psi^{(\cdot)}$ with the element ψ_j removed.

The conditional probabilities in Eq. (D3) are equal to the independence likelihoods:

- $p(y\,|\,\psi_{\mathcal{H}_0}) : L_{ind}(\mathbf{i}^{(obs)}, \mathbf{i}^{(mod)}\,|\,\psi_{\mathcal{H}_0}) = L_{ind}(\mathbf{i}^{(obs)}\,|\,\psi^{(obs)}) \times L_{ind}(\mathbf{i}^{(mod)}\,|\,\psi^{(mod)})$, where the form of L_{ind} in the righ-hand side is given in Eq. (D4).

- $p(y\,|\,\psi_{\mathcal{H}_1}) : L_{ind}(\mathbf{i}^{(obs)}, \mathbf{i}^{(mod)}\,|\,\psi_{\mathcal{H}_1}) = L_{ind}(\mathbf{i}^{(obs)}\,|\,\psi_{-j}^{(obs)}, \psi_j) \times L_{ind}(\mathbf{i}^{(mod)}\,|\,\psi_{-j}^{(mod)}, \psi_j)$.

References

Agilan, V., and N. Umamahesh, 2016: Modelling nonlinear trend for developing non-stationary rainfall intensity-duration-frequency curve. *Int. J. Climatol.*, **37 (3)**, 1265–1281, doi:10.1002/joc.4774.

Argüeso, D., J. P. Evans, L. Fita, and K. J. Bormann, 2014: Temperature response to future urbanization and climate change. *Clim. Dyn.*, **42 (7)**, 2183–2199, doi:10.1007/s00382-013-1789-6.

Ban, N., J. Schmidli, and C. Schär, 2015: Heavy precipitation in a changing climate: Does short-term summer precipitation increase faster? *Geophys. Res. Lett.*, **42 (4)**, 1165–1172, doi:10.1002/2014GL062588.

Ban, N., and Coauthors, 2021: The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: evaluation of precipitation. *Clim. Dyn.*, **57** (**1**), 275–302, doi:10.1007/s00382-021-05708-w.

Beck, H. E., E. F. Wood, M. Pan, C. K. Fisher, D. Miralles, A. I. van Dijk, T. R. McVicar, and R. F. Adler, 2019: MSWEP V2 global 3-hourly 0.1° precipitation : methodology and quantitative assessment. *Bull. Amer. Meteor. Soc.*, **100** (**3**), 473–502, doi:10.1175/bams-d-17-0138.1.

Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels, 2004: *Statistics of Extremes*. John Wiley & Sons, Ltd.

Berg, P., O. B. Christensen, K. Klehmet, G. Lenderink, J. Olsson, C. Teichmann, and W. Yang, 2019: Summertime precipitation extremes in a EURO-CORDEX 0.11° ensemble at an hourly resolution. *Nat. Hazards Earth Syst. Sci.*, **19** (**4**), 957–971, doi:10.5194/nhess-19-957-2019.

Brisson, E., M. Demuzere, and N. P. van Lipzig, 2016a: Modelling strategies for performing convection-permitting climate simulations. *Meteorol. Z.*, **25** (**2**), 149–163, doi:10.1127/metz/2015/0598.

Brisson, E., K. Van Weverberg, M. Demuzere, A. Devis, S. Saeed, M. Stengel, and N. P. M. van Lipzig, 2016b: How well can a convection-permitting climate model reproduce decadal statistics of precipitation, temperature and cloud characteristics? *Clim. Dyn.*, **47** (**9**), 3043–3061, doi:10.1007/s00382-016-3012-z.

Bubnová, R., G. Hello, P. Bénard, and J.-F. Geleyn, 1995: Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. *Mon. Weather Rev.*, **123** (**2**), 515–535, doi:10.1175/1520-0493(1995)123<0515:IOTFEE>2.0.CO;2.

Burlando, P., and R. Rosso, 1996: Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. *J. Hydrol.*, **187 (1-2)**, 45–64, doi:10.1016/S0022-1694(96) 03086-7.

Chandler, R. E., , and E. M. Scott, 2011: *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. John Wiley & Sons.

Chavez-Demoulin, V., and A. C. Davison, 2012: Modelling time series extremes. *REVSTAT*, **10 (1)**, 109–133.

Cheng, L., and A. AghaKouchak, 2014: Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate. *Sci. Rep.*, **4 (1)**, 7093, doi:10.1038/srep07093.

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London Ltd.

Cooley, D., 2005: Statistical analysis of extremes motivated by weather and climate studies: Applied and theoretical advances. Ph.D. thesis, University of Colorado, Boulder, USA, URL https://www.stat.colostate.edu/~cooleyd/Papers/cooley2.pdf.

Cooley, D., J. Cisewski, R. J. Erhardt, S. Jeon, E. Mannshardt, B. O. Omolo, and Y. Sun, 2012: A survey of spatial extremes: measuring spatial dependence and modeling spatial effects. *REVSTAT*, **10 (1)**, 135–165.

Cooley, D., P. Naveau, and P. Poncet, 2006: Variograms for spatial max-stable random fields. *Dependence in Probability and Statistics*, P. Bertail, P. Soulier, and P. Doukhan, Eds., Springer, 373–390.

Coppola, E., and Coauthors, 2020: A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean. *Clim. Dyn.*, **55**, 3–34, doi:10.1007/s00382-018-4521-8.

Cortés-Hernández, V. E., F. Zheng, J. Evans, M. Lambert, A. Sharma, and S. Westra, 2016: Evaluating regional climate models for simulating sub-daily rainfall extremes. *Clim. Dyn.*, **47 (5)**, 1613–1628, doi:10.1007/s00382-015-2923-4.

Davison, A. C., and M. M. Gholamrezaee, 2011: Geostatistics of extremes. *Proc. R. Soc. A.*, **468**, 581–608, doi:10.1098/rspa.2011.0412.

Davison, A. C., S. A. Padoan, and M. Ribatet, 2012: Statistical modeling of spatial extremes. *Statist. Sci.*, **27 (2)**, 161–186, doi:10.1214/11-STS376.

De Ridder, K., and H. Gallée, 1998: Land surface-induced regional climate change in Southern Israel. *J. Appl. Meteorol.*, **37 (11)**, 1470–1485, doi:10.1175/1520-0450(1998)037<1470:LSIRCC>2.0.CO;2.

De Troch, R., R. Hamdi, H. Van de Vyver, J.-F. Geleyn, and P. Termonia, 2013: Multiscale performance of the ALARO-0 model for simulating extreme summer precipitation climatology in Belgium. *J. Climate*, **26 (22)**, 8895–8915, doi:10.1175/JCLI-D-12-00844.1.

De Troch, R., and Coauthors, 2014: Overview of a few regional climate models and climate scenarios for Belgium. Tech. rep., Royal Meteorological Institute of Belgium.

EEA, 2019: Economic losses from climate-related extremes in Europe. *European Environment Agency*, Prod–ID: IND–182.

Fosser, G., S. Khodayar, and P. Berg, 2015: Benefit of convection permitting climate model simulations in the representation of convective precipitation. *Clim. Dyn.*, **44 (1-2)**, 45–60, doi: 10.1007/s00382-014-2242-1.

Fosser, G., S. Khodayar, and P. Berg, 2017: Climate change in the next 30 years: What can a convection-permitting model tell us that we did not already know? *Clim. Dyn.*, **48 (5)**, 1987–2003, doi:10.1007/s00382-016-3186-4.

Gallée, H., and G. Schayes, 1994: Development of a three-dimensional meso-γ primitive equation model: Katabatic winds simulation in the area of Terra Nova Bay, Antarctica. *Mon. Weather Rev.*, **122 (4)**, 671–685, doi:10.1175/1520-0493(1994)122<0671:DOATDM>2.0.CO;2.

Gerard, L., J.-M. Piriou, R. Bro ˇ zková, J.-F. Geleyn, and D. Banciu, 2009: Cloud and precipitation parameterization in a meso-gamma-scale operational weather prediction model. *Mon. Weather Rev.*, **137 (11)**, 3960–3977, doi:10.1175/2009MWR2750.1.

Giorgi, F., C. Jones, and G. R. Asrar, 2009: Addressing climate information needs at the regional level: the CORDEX framework. *WMO Bull.*, **58 (3)**, 175–183.

Gupta, V. K., and E. Waymire, 1990: Multiscaling properties of spatial rainfall and river flow distributions. *J. Geophys. Res.*, **95 (D3)**, 1999–2009, doi:10.1029/JD095iD03p01999.

Helsen, S., and Coauthors, 2020: Consistent scale-dependency of future increases in hourly extreme precipitation in two convection-permitting climate models. *Clim. Dyn.*, **54 (3–4)**, 1267–1280, doi:10.1007/s00382-019-05056-w.

Hobæk Haff, I., A. Frigessi, and D. Maraun, 2015: How well do regional climate models simulate the spatial dependence of precipitation? An application of pair-copula constructions. *J. Geophys. Res. Atmos.*, **120 (7)**, 2624–2646, doi:10.1002/2014JD022748.

Hou, A. Y., and Coauthors, 2014: The global precipitation measurement mission. *Bull. Amer. Meteor. Soc.*, **95 (5)**, 701–722, doi:10.1175/BAMS-D-13-00164.1.

Jacob, D., and Coauthors, 2014: EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg. Environ. Change*, **14 (2)**, 563–578, doi:110.1007/s10113-013-0499-2.

Jacob, D., and Coauthors, 2020: Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. *Reg. Environ. Change*, **20 (2)**, Article: 51, doi:10.1007/s10113-020-01606-9.

Kendon, E. J., N. M. Roberts, H. J. Fowler, M. J. Roberts, S. C. Chan, and C. A. Senior, 2014: Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nat. Climate Change*, **4 (7)**, 570–576, doi:10.1038/nclimate2258.

Kendon, E. J., and Coauthors, 2017: Do convection-permitting regional climate models improve projections of future precipitation change? *Bull. Amer. Meteor. Soc.*, **98 (1)**, 79–93, doi:10.1175/BAMS-D-15-0004.1.

Koenker, R., 2005: *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R., 2018: quantreg: Quantile Regression. http://CRAN.R-project.org/package=quantreg.

Koenker, R., and J. A. F. Machado, 1999: Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.*, **94 (448)**, 1296–1310, doi:10.1080/01621459.1999.10473882.

Leadbetter, M., G. Lindgren, and H. Rootzén, 1983: *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

Lenderink, G., and E. van Meijgaard, 2008: Increase in hourly extreme precipitation beyond expectations from temperature changes. *Nat. Geosci.*, **1 (8)**, 511–514, doi:10.1038/ngeo262.

Lenderink, G., and E. van Meijgaard, 2010: Linking increases in hourly precipitation extremes to atmospheric temperature and moisture changes. *Environ. Res. Lett.*, **5 (2)**, 025 208, doi: 10.1088/1748-9326/5/2/025208.

Lugrin, T., 2016: Bayesian uncertainty management in temporal dependence of extremes. *Extremes*, **19 (3)**, 491–515, URL http://dx.doi.org/10.1007/s10687-016-0258-0, R package version 0.3.2.

Maraun, D., T. J. Osborn, and H. W. Rust, 2012: The influence of synoptic airflow on UK daily precipitation extremes. Part II: regional climate model and E-OBS data validation. *Clim. Dyn.*, **23 (1)**, 287–301, doi:10.1007/s00382-011-1176-0.

Maraun, D., H. W. Rust, and T. J. Osborn, 2009: The annual cycle of heavy precipitation across the United Kingdom: a model based on extreme value statistics. *Int. J. Climatol.*, **29 (12)**, 1731–1744, doi:10.1002/joc.1811.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, **48 (3)**, RG3003, doi:10.1029/2009RG000314.

Martel, J.-L., A. Mailhot, and F. Brissette, 2020: Global and regional projected changes in 100-yr subdaily, daily, and multiday precipitation extremes estimated from three large ensembles of climate simulations. *J. Climate*, **33 (3)**, 1089–1103, doi:10.1175/JCLI-D-18-0764.1.

Met Office, 2021: UK Climate Projections (UKCP). https://www.metoffice.gov.uk/research/approach/collaboration/ukcp/index/.

Naveau, P., A. Guillou, D. Cooley, and J. Diebolt, 2009: Modeling pairwise dependence of maxima in space. *Biometrika*, **96 (1)**, 1–17, doi:10.1093/biomet/asp001.

Ntegeka, V., and P. Willems, 2008: Trends and multidecadal oscillations in rainfall extremes, based on a more than 100-year time series of 10 min rainfall intensities at Uccle, Belgium. *Water Resour. Res.*, **44 (7)**, W07 402, doi:10.1029/2007WR006471.

Olsson, J., P. Berg, and A. Kawamura, 2015: Impact of RCM spatial resolution on the reproduction of local, subdaily precipitation. *J. Hydrometeorol.*, **16 (2)**, 534 – 547, doi: 10.1175/JHM-D-14-0007.1.

Ouarda, T. B. M. J., L. A. Yousef, and C. Charron, 2018: Non-stationary intensity-duration-frequency curves integrating information concerning teleconnections and climate change. *Int. J. of Climatol.*, **39 (4)**, 2306–2323, doi:10.1002/joc.5953.

Pan, L.-L., S.-H. Chen, D. Cayan, M.-Y. Lin, Q. Hart, M.-H. Zhang, Y. Liu, and J. Wang, 2011: Influences of climate change on California and Nevada regions revealed by a high-resolution dynamical downscaling study. *Clim. Dyn.*, **37 (9)**, 2005–2020, doi:10.1007/s00382-010-0961-5.

Prein, A. F., R. M. Rasmussen, K. Ikeda, C. Liu, M. P. Clark, and G. J. Holland, 2017: The future intensification of hourly precipitation extremes. *Nature Clim. Change*, **7 (1)**, 48–52, doi:10.1038/nclimate3168.

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenge. *Rev. Geophys.*, **53 (2)**, 323–361, doi: 10.1002/2014RG000475.

Prein, A. F., and Coauthors, 2016: Precipitation in the EURO-CORDEX 0.11° and 0.44° simulations: high resolution, high benefits? *Clim. Dyn.*, **46 (1)**, 383–412, doi:10.1007/s00382-015-2589-y.

Rasmussen, S., J. Christensen, M. Drews, D. Gochis, and J. Refsgaard, 2012: Spatial scale characteristics of precipitation simulated by regional climate models and the implications for hydrological modelling. *J. Hydrometeorol.*, **13 (6)**, 1817–1835, doi:10.1175/JHM-D-12-07.1.

Reszler, C., M. B. Switanek, and H. Truhetz, 2018: Convection-permitting regional climate simulations for representing floods in small- and medium-sized catchments in the Eastern Alps. *Nat. Hazards Earth Syst. Sci.*, **18 (10)**, 2653–2674, doi:10.5194/nhess-18-2653-2018.

Ribatet, M., R. Singleton, and R Core Team, 2013: *SpatialExtremes: Modelling Spatial Extremes*. URL https://CRAN.R-project.org/package=SpatialExtremes, R package version 2.0-0.

Rockel, B., A. Will, and A. Hense, 2008: The regional climate model COSMO-CLM (CCLM). *Meteorol. Z.*, **17 (4)**, 347–348, doi:10.1127/0941-2948/2008/0309.

Schär, C., and Coauthors, 2020: Kilometer-scale climate models: Prospects and challenges. *Bull. Amer. Meteor. Soc.*, **101 (5)**, E567–E587, doi:10.1175/BAMS-D-18-0167.1.

Schindler, A., D. Maraun, and J. Luterbacher, 2012: Validation of the present day annual cycle in heavy precipitation over the British Islands simulated by 14 RCMs. *J. Geophys. Res.*, **117**, D18 107, doi:10.1029/2012JD017828.

Shikano, S., 2019: Hypothesis testing in the Bayesian framework. *Swiss Political Science Review*, **25 (3)**, 288–299, doi:10.1111/spsr.12375.

Smith, R. L., and I. Weissman, 1994: Estimating the extremal index. *J. R. Stat. Soc. B*, **56 (3)**, 515–528, doi:10.1111/j.2517-6161.1994.tb01997.x.

Tabari, H., and Coauthors, 2016: Local impact analysis of climate change on precipitation extremes: are high-resolution climate models needed for realistic simulations? *Hydrol. Earth Syst. Sci.*, **20 (9)**, 3843–3857, doi:10.5194/hess-20-3843-2016.

Termonia, P., and Coauthors, 2018a: The ALADIN system and its canonical model configurations AROME CY41T1 and ALARO CY40T1. *Geosci. Model Dev.*, **11 (1)**, 257–281, doi:10.5194/gmd-11-257-2018.

Termonia, P., and Coauthors, 2018b: The CORDEX.be initiative as a foundation for climate services in Belgium. *Clim. Serv.*, **11**, 49–61, doi:10.1016/j.cliser.2018.05.001.

Torma, C., F. Giorgi, and E. Coppola, 2015: Added value of regional climate modeling over areas characterized by complex terrain–Precipitation over the Alps. *J. Geophys. Res. Atmos.*, **120 (9)**, 3957–3972, doi:10.1002/2014JD022781.

Trenberth, K. E., A. Dai, R. M. Rasmussen, and D. B. Parsons, 2003: The changing character of precipitation. *Bull. Amer. Meteor. Soc.*, **84 (9)**, 1205–1218, doi:10.1175/BAMS-84-9-1205.

Van de Vyver, H., 2012: Spatial regression models for extreme precipitation in Belgium. *Water Resour. Res.*, **48 (9)**, W09 549, doi:10.1029/2011WR011707.

Van de Vyver, H., 2015a: Bayesian estimation of rainfall intensity-duration-frequency relationships. *J. Hydrol.*, **529 (3)**, 1451–1463, doi:10.1016/j.jhydrol.2015.08.036.

Van de Vyver, H., 2015b: On the estimation of continuous 24-h precipitation maxima. *Stoch. Environ. Res. Risk Assess.*, **29 (3)**, 653–663, doi:10.1007/s00477-014-0912-5.

Van de Vyver, H., 2018: A multiscaling-based intensity-duration-frequency model for extreme precipitation. *Hydrol. Process.*, **32 (11)**, 1635–1647, doi:10.1002/hyp.11516.

Van de Vyver, H., 2021: R code for Bayesian Estimation of Rainfall Intensity-Duration-Frequency (IDF) Relationships (Version V1). URL http://doi.org/10.5281/zenodo.4644184, Zenodo.

Van de Vyver, H., B. Van Schaeybroeck, R. De Troch, R. Hamdi, and P. Termonia, 2019: Modeling the scaling of short-duration precipitation extremes with temperature. *Earth Space Sci.*, **6 (10)**, 2031–2041, doi:10.1029/2019EA000665.

Van de Vyver, H., B. Van Schaeybroeck, R. De Troch, R. Hamdi, and P. Termonia, 2021: R code for Modeling the Scaling of Short-Duration Precipitation Extremes with Temperature (Version V1). URL http://doi.org/10.5281/zenodo.4644567, Zenodo.

Van Weverberg, K., E. Goudenhoofdt, U. Blahak, E. Brisson, M. Demuzere, P. Marbaix, and J.-P. van Ypersele, 2014: Comparison of one-moment and two-moment bulk microphysics for high-resolution climate simulations of intense precipitation. *Atmos. Res.*, **147–148**, 145–161, doi:10.1016/j.atmosres.2014.05.012.

Vannière, B., and Coauthors, 2019: Multi-model evaluation of the sensitivity of the global energy budget and hydrological cycle to resolution. *Clim. Dyn.*, **52 (11)**, 6817–6846, doi:10.1007/s00382-018-4547-y.

Vannitsem, S., and P. Naveau, 2007: Spatial dependences among precipitation maxima over Belgium. *Nonlin. Processes Geophys.*, **14 (5)**, 621–630, doi:10.5194/npg-14-621-2007.

Wasko, C., and A. Sharma, 2014: Quantile regression for investigating scaling of extreme precipitation with temperature. *Water Resour. Res.*, **50 (4)**, 3608–3614, doi:10.1002/2013WR015194.

Westra, S., and Coauthors, 2014: Future changes to the intensity and frequency of short-duration extreme rainfall. *Rev. Geophys.*, **52 (3)**, 522–555, doi:10.1002/2014RG000464.

Willems, P., 2000: Compound intensity/duration/frequency-relationships of extreme precipitation for two seasons and two storm types. *J. Hydrol.*, **233 (1-4)**, 189–205, doi:10.1016/S0022-1694(00)00233-X.

Wouters, H., M. Demuzere, U. Blahak, K. Fortuniak, B. Maiheu, J. Camps, D. Tielemans, and N. P. M. van Lipzig, 2016: The efficient urban canopy dependency parametrization (SURY) v1.0 for atmospheric modelling: description and application with the COSMO-CLM model for a Belgian summer. *Geosci. Model Dev.*, **9 (9)**, 3027–3054, doi:10.5194/gmd-9-3027-2016.

Wyard, C., C. Scholzen, X. Fettweis, J. Van Campenhout, and L. Fran¸ois, 2017: Decrease in climatic conditions favouring floods in the south-east of Belgium over 1959-2010 using the regional climate model MAR. *Int. J. of Climatol.*, **37 (5)**, 2782–2796, doi:10.1002/joc.4879.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

LIST OF TABLES

Table 1. Summary of the evaluation metrics. (AM: annual maxima, POT: peaks-over-threshold, CQ: conditional quantiles). 47

Table 2. The RCM models contributed to the CORDEX.be project (Termonia et al. 2018b), and their characteristics. 48

Table 3. Kullback-Leibler divergence, Eq. (C1), between the modeled and the observed diurnal cycle. 49

Table 4. Number of stations (out of 18) where the multiscaling GEV-model is better than the simple scaling GEV-model. Model selection is based on (i) AIC and (ii) evaluation of the posterior odd $R = p(\mathcal{H}_0 \,|\, y)/p(\mathcal{H}_1 \,|\, y)$, with $\mathcal{H}_0$: simple scaling hypothesis, and $\mathcal{H}_1$: multiscaling hypothesis. 50

Table 5. Number of stations (out of 18) where the parameters of the multiscaling GEV-model of the observed and the RCM-simulated extremes are found to agree. Testing for equality of model parameters $\psi = (\mu, \sigma, \xi, \eta_1, \eta_2)$ is based on (i) AIC and (ii) evaluation of the posterior odd $R = p(\mathcal{H}_0 \,|\, y)/p(\mathcal{H}_1 \,|\, y)$, with $\mathcal{H}_0$: $\psi_j^{(obs)} \neq \psi_j^{(mod)}$, and $\mathcal{H}_0$: $\psi_j^{(obs)} = \psi_j^{(mod)}$. 51

Table 6. General overview of the RCM-evaluation results. EUR-11 versus H-Res simulations. 52

TABLE 1. Summary of the evaluation metrics. (AM: annual maxima, POT: peaks-over-threshold, CQ: conditional quantiles).

Metric	Statistical methodology	Extremes definition	Code availability
Seasonality	Extreme rainfall occurrences per season (Cortés-Hernández et al. 2016).	AM/POT	-
Diurnal cycle	Extreme rainfall occurrences per hour (Cortés-Hernández et al. 2016).	AM/POT	-
Temperature and humidity dependency	Model the increase of extremes against (dew point) temperature (cfr. CC-scaling & super-CC scaling) with piecewise linear quantile regression (Van de Vyver et al. 2019).	CQ	R-code in Van de Vyver et al. (2021)
Temporal scaling	Model scaling relationships between different levels of aggregation (IDF-relationships) with multiscaling extreme value distributions (Van de Vyver 2018).	AM	R-code in Van de Vyver (2021)
Spatial structure	Model spatial dependence with the madogram (Cooley et al. 2006; Vannitsem and Naveau 2007; Naveau et al. 2009).	AM	R-package: `SpatialExtremes` (Ribatet et al. 2013)
Temporal clustering	Model cluster properties with the extremal index (Leadbetter et al. 1983; Smith and Weissman 1994).	POT	R-package: `tsxtreme` (Lugrin 2016)

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

Table 2. The RCM models contributed to the CORDEX.be project (Termonia et al. 2018b), and their characteristics.

Model Version	Name	Resolution	No. vertical levels	Precip. freq.	Available years	Important scheme
1) ALARO-0	ALARO-EUR11	12.5 km	46	1h	1979-2010	3MT
	ALARO-HRes	4 km		3min	1979-2010	
2) COSMO-CLM (UCL) V. 5.0-CLM6	CCLM-UCL-EUR11	12.5 km	40	15min	1979-2009	Two-moment microphysical scheme
	CCLM-UCL-HRes	2.8 km		15min	1979-2009	
3) COSMO-CLM (KUL) V. 6.0-CLM6	CCLM-KUL-EUR11	12.5 km	40	1h	1979-2014	TERRA_URB
	CCLM-KUL-HRes	2.8 km		15min	1979-2009	
4) MAR V. 3.9	MAR-HRes	5 km	30	15min	1978-2017	Snow variables

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

TABLE 3. Kullback-Leibler divergence, Eq. (C1), between the modeled and the observed diurnal cycle.

	ALARO		CCLM-UCL		CCLM-KUL		MAR
	EUR-11	H-Res	EUR-11	H-Res	EUR-11	H-Res	H-Res
AM	0.069	0.094	0.37	0.12	0.44	0.16	0.24
POT	0.026	0.026	0.14	0.031	0.18	0.050	0.11

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

TABLE 4. Number of stations (out of 18) where the multiscaling GEV-model is better than the simple scaling GEV-model. Model selection is based on (i) AIC and (ii) evaluation of the posterior odd $R = p(\mathcal{H}_0 \,|\, y)/p(\mathcal{H}_1 \,|\, y)$, with $\mathcal{H}_0$: simple scaling hypothesis, and $\mathcal{H}_1$: multiscaling hypothesis.

	OBS	ALARO		CCLM-UCL		CCLM-KUL	
		EUR-11	H-Res	EUR-11	H-Res	EUR-11	H-Res
Selection criterion: AIC	16	7	17	14	16	13	14
Selection criterion: $R < 1$	17	12	17	16	18	15	17

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

TABLE 5. Number of stations (out of 18) where the parameters of the multiscaling GEV-model of the observed and the RCM-simulated extremes are found to agree. Testing for equality of model parameters $\psi = (\mu, \sigma, \xi, \eta_1, \eta_2)$ is based on (i) AIC and (ii) evaluation of the posterior odd $R = p(\mathcal{H}_0 \,|\, y)/p(\mathcal{H}_1 \,|\, y)$, with $\mathcal{H}_0$: $\psi_j^{(obs)} \neq \psi_j^{(mod)}$, and $\mathcal{H}_0$: $\psi_j^{(obs)} = \psi_j^{(mod)}$.

		μ	σ	ξ	$\eta_{1,2}$	μ	σ	ξ	$\eta_{1,2}$
		EUR-11				H-Res			
ALARO	Selection criterion: AIC	0	1	10	4	10	0	9	6
	Selection criterion: R <1	2	1	14	3	10	1	7	3
	Selection criterion: R <3	4	1	16	8	12	2	11	6
	Spatial product of R ($\log_{10}$-scale)	34.2	38.7	-1.3	15.4	9.7	30.1	6.3	15.2
CCLM-UCL	Selection criterion: AIC	13	0	13	16	4	0	12	10
	Selection criterion: R <1	15	2	14	15	3	0	12	8
	Selection criterion: R <3	16	3	15	16	5	3	17	11
	Spatial product of R ($\log_{10}$-scale)	-4.3	18.1	-2.4	-5.1	27.5	26.3	1.4	9.6
CCLM-KUL	Selection criterion: AIC	11	1	9	14	3	3	8	11
	Selection criterion: R <1	12	3	13	13	3	3	8	7
	Selection criterion: R <3	14	5	15	16	5	5	12	11
	Spatial product of R ($\log_{10}$-scale)	0.6	20.3	-2.1	-3.0	21.7	20.7	4.7	5.0

Table 6. General overview of the RCM-evaluation results. EUR-11 versus H-Res simulations.

Metric	EUR-11 simulations
Seasonality	Perform all similar and reasonably well.
Diurnal cycle	(i) ALARO reproduces the diurnal cycle, (ii) CCLM-models does not represent the diurnal cycles well.
Temperature and humidity dependency	(i) Reproduce super-CC scaling behavior, (ii) CC-scaling rate was underestimated for 0.95-quantiles, but correctly estimated for 0.99-quantiles, (iii) Relative strength of the predictors for 1h-extremes was reasonably represented.
Temporal scaling	(i) Reproduce multiscaling properties for extreme intensities with duration $d = 6, \ldots, 72$h. (ii) Capture very well the GEV-parameters: shape parameter ξ, and temporal scaling parameters $\eta_{1,2}$, except ALARO that underestimates $\eta_{1,2}$.
Spatial structure	Spatial dependency is generally too high.
Temporal clustering	Overestimate the cluster size of 1h-extremes.
	H-Res simulations
Seasonality	No significant improvement compared to EUR-11.
Diurnal cycle	(i) ALARO: no improvement compared to EUR-11, (ii) CCLM-models produce better diurnal cycles.
Temperature and humidity dependency	(i) Reproduce super-CC scaling behavior, (ii) CC-scaling rates are mostly higher than for EUR-11.
Temporal scaling	(i) Reproduce multiscaling properties for extreme intensities with duration $d = 1, \ldots, 72$h. (ii) The temporal scaling parameters were not improved. The scaling GEV-distribution for ALARO (evaluated at $d = 1$h) agrees well with the observed 1h-extremes.
Spatial structure	Spatial dependency is slightly decreased compared to EUR-11, but it is still too high.
Temporal clustering	Cluster size estimations are improved compared to EUR-11.

LIST OF FIGURES

Fig. 1. Elevation map (m) of Belgium, together with the location of the 10-min pluviograph stations. Geospatial information of the locations (latitude/longitude/elevation) is listed in Table S1. . . . 54

Fig. 2. Boxplot of the bias in the mean annual maxima, for different durations and model resolutions. Every box includes the 18 pluviograph stations/gridpoints. Positive (negative) bias means over- (under)estimation. 55

Fig. 3. Number of locations with p-value> 0.05, for the Kolmogorov-Smirnov test statistic, under the hypothesis that the observed and modeled annual maxima follow the same distribution. Horizontal dashed line: total number of locations (i.e. 18). 56

Fig. 4. Seasonal cycle analysis with the average seasonal occurrence of (i) observed annual maximum d-hourly rainfall of the 18 pluviograph stations, and (ii) the 18 gridpoints of the EUR-11 simulations. 57

Fig. 5. Same as Fig. 4, but for the H-Res simulations. 58

Fig. 6. Diurnal cycle analysis with the average occurrence at each time of day (UTC) of (i) observed annual maximum hourly rainfall of the 18 pluviograph stations, and (ii) the 18 gridpoints of the RCM-simulations (EUR-11 and H-Res). 59

Fig. 7. High τ-quantiles of hourly precipitation as a function of the daily mean dew point temperature. Solid lines: estimated with binning. Dotted lines: estimated with piecewise linear quantile regression, Eq. (2). Light grey solid lines: piecewise regression lines of the observed quantiles, which serve as a reference (cfr. top left). 60

Fig. 8. The strength of the relationship between the predictor (air temperature or dew point temperature) and 0.99-quantiles of hourly precipitation. Bars: the goodness-of-fit criterion GoF, Eq. (A4). The vertical lines indicate the 95% confidence intervals of GoF. 61

Fig. 9. GEV-parameters (location/scale) plotted against rainfall duration d, for station Uccle. First row: location. Second row: scale. Points: maximum likelihood estimators $(d_i, \hat{\mu}_i)$ and $(d_i, \hat{\sigma}_i)$. Solid lines: generalized least squares fits to the estimated GEV-parameters. 62

Fig. 10. Spatial extremal dependence: the madogram $\nu_F(h)$, Eq. (7), as a function of distance h for (i) the observed annual maxima at the 18 pluviograph stations, and for (ii) the 18 gridpoints from ALARO. Dots: binned empirical madogram, Eq. (8). Dashed lines: 95%-confidence bounds of the empirical madogram. Red line: independent case, $\nu_F = 1/6$. 63

Fig. 11. Mean cluster size (unit: hour) against the threshold u, expressed in probability level. The runs estimator, Eq. (9), was used. Upper (lower) rows are for EUR-11 (H-Res) simulations. . . 64

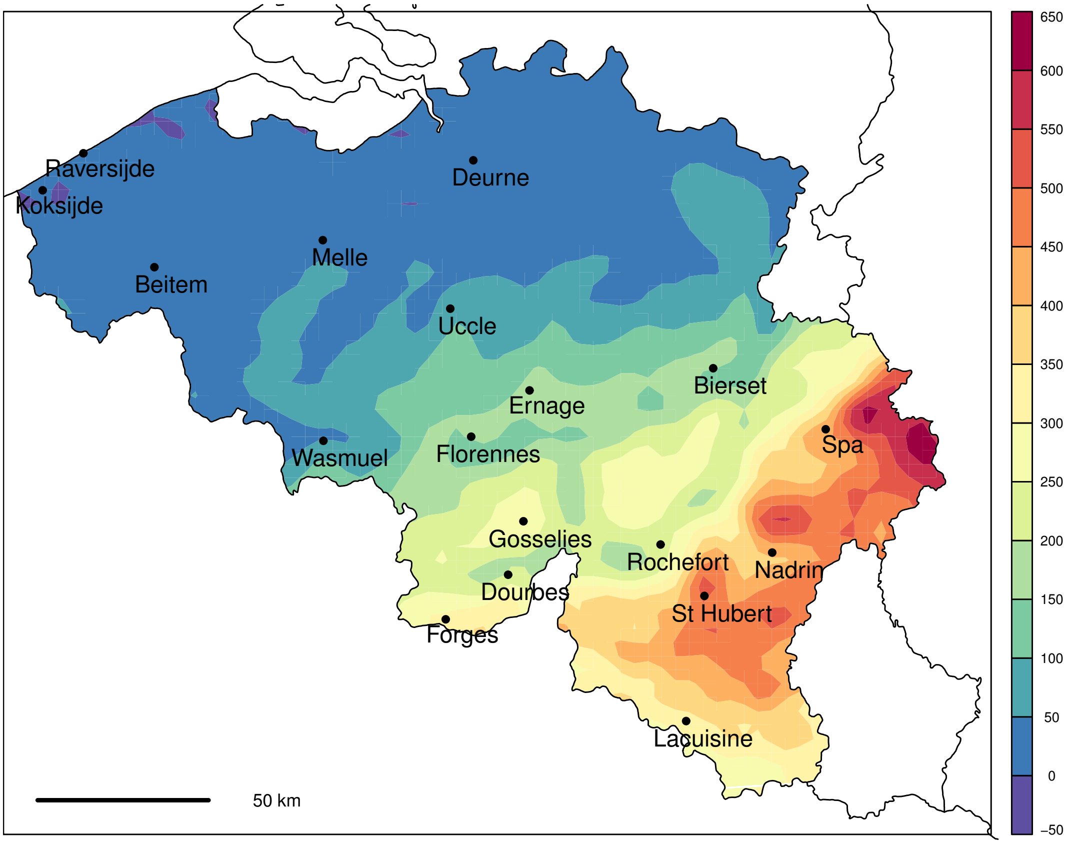

FIG. 1. Elevation map (m) of Belgium, together with the location of the 10-min pluviograph stations. Geospatial information of the locations (latitude/longitude/elevation) is listed in Table S1.

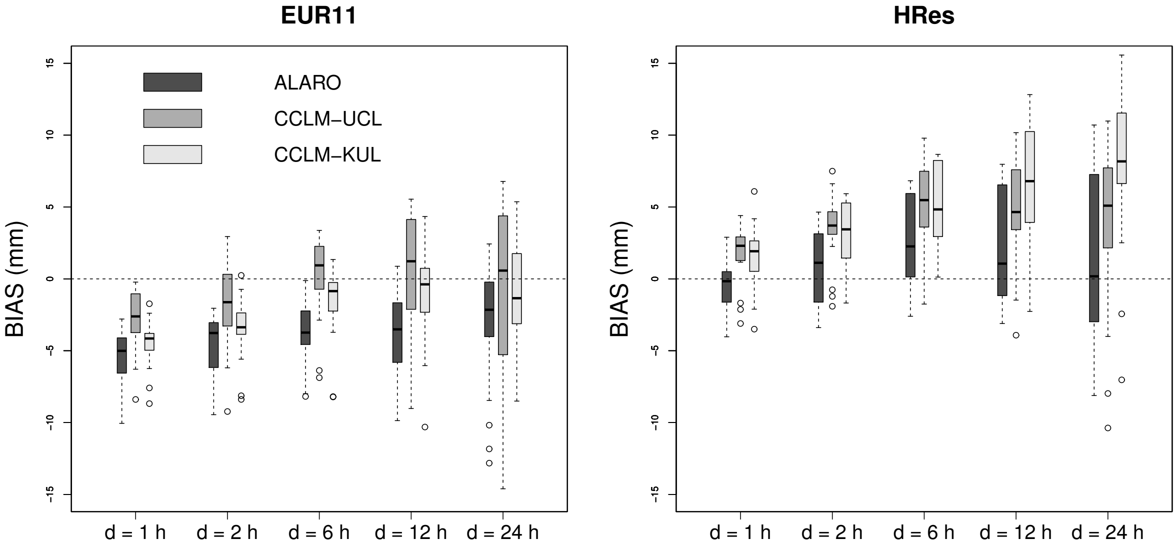

FIG. 2. Boxplot of the bias in the mean annual maxima, for different durations and model resolutions. Every box includes the 18 pluviograph stations/gridpoints. Positive (negative) bias means over- (under)estimation.

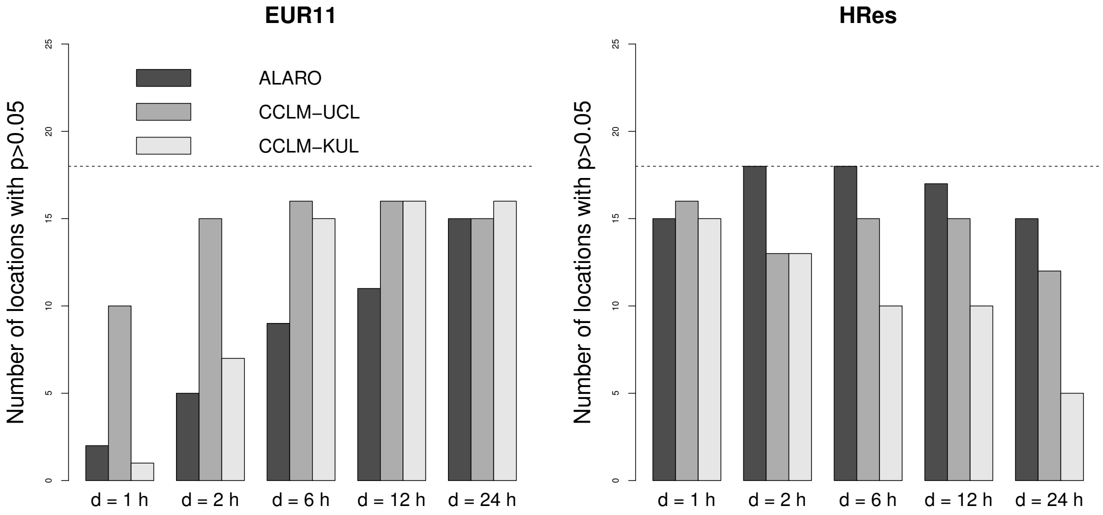

FIG. 3. Number of locations with p-value>0.05, for the Kolmogorov-Smirnov test statistic, under the hypothesis that the observed and modeled annual maxima follow the same distribution. Horizontal dashed line: total number of locations (i.e. 18).

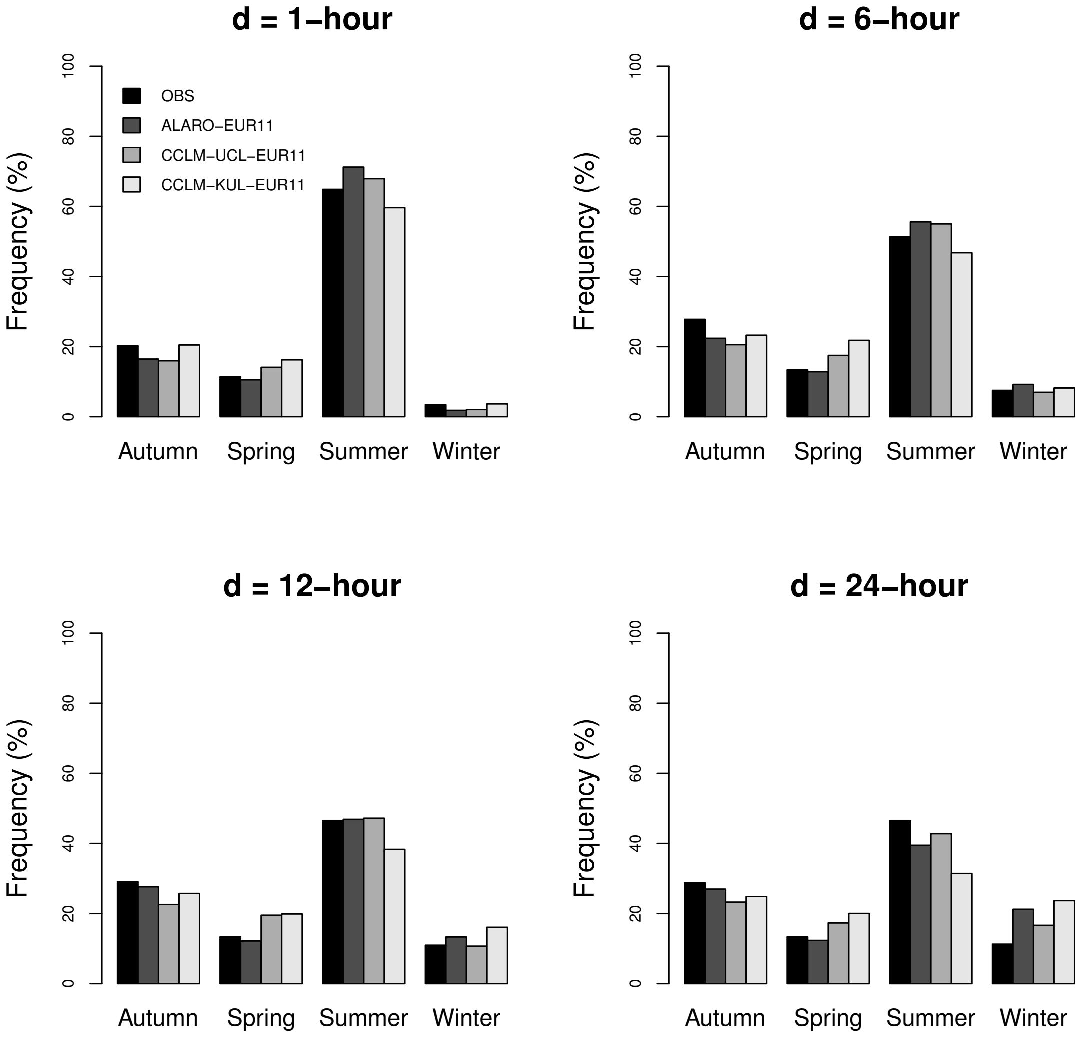

FIG. 4. Seasonal cycle analysis with the average seasonal occurrence of (i) observed annual maximum d-hourly rainfall of the 18 pluviograph stations, and (ii) the 18 gridpoints of the EUR-11 simulations.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.

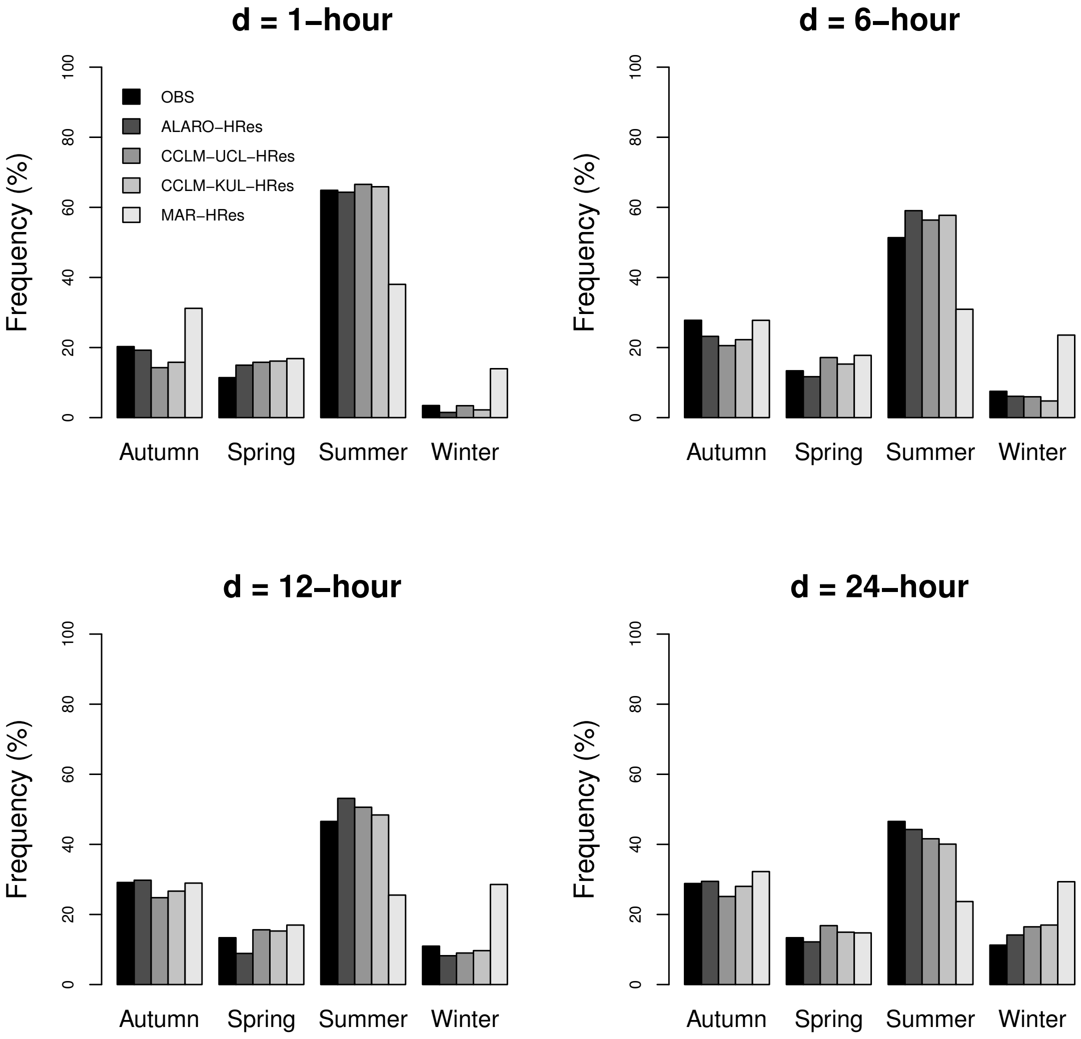

FIG. 5. Same as Fig. 4, but for the H-Res simulations.

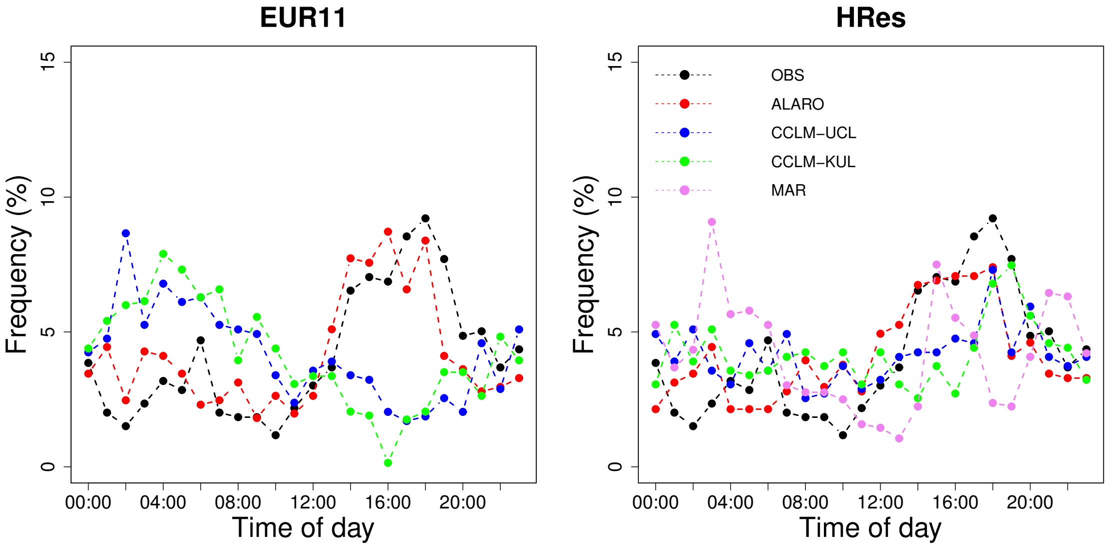

FIG. 6. Diurnal cycle analysis with the average occurrence at each time of day (UTC) of (i) observed annual maximum hourly rainfall of the 18 pluviograph stations, and (ii) the 18 gridpoints of the RCM-simulations (EUR-11 and H-Res).

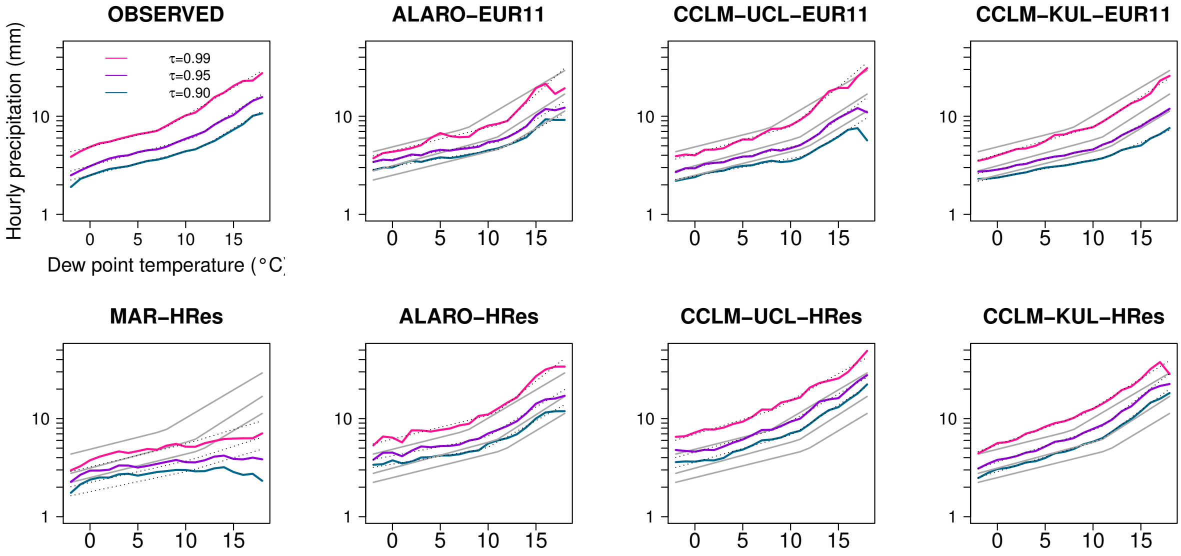

FIG. 7. High τ-quantiles of hourly precipitation as a function of the daily mean dew point temperature. Solid lines: estimated with binning. Dotted lines: estimated with piecewise linear quantile regression, Eq. (2). Light grey solid lines: piecewise regression lines of the observed quantiles, which serve as a reference (cfr. top left).

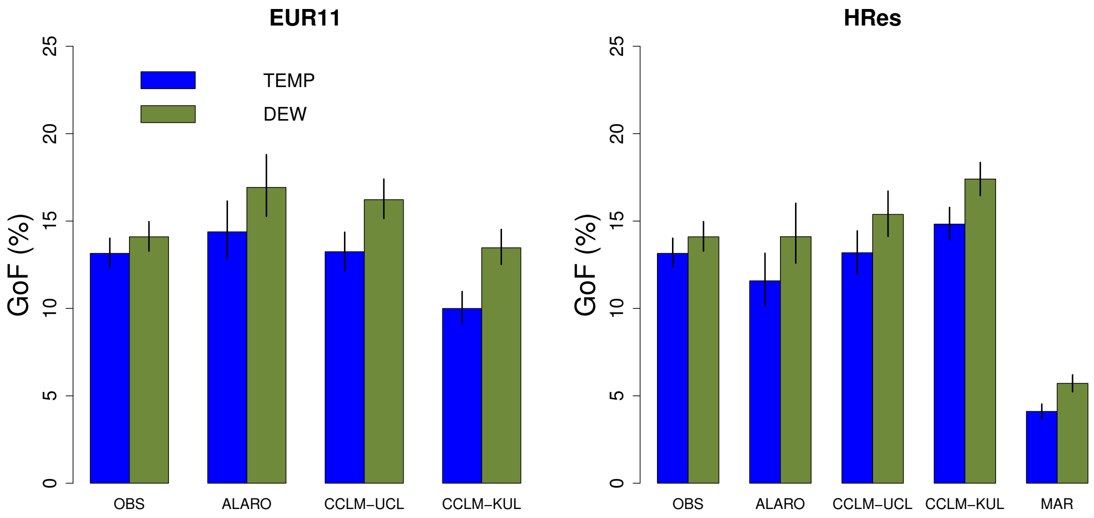

FIG. 8. The strength of the relationship between the predictor (air temperature or dew point temperature) and 0.99-quantiles of hourly precipitation. Bars: the goodness-of-fit criterion GoF, Eq. (A4). The vertical lines indicate the 95% confidence intervals of GoF.

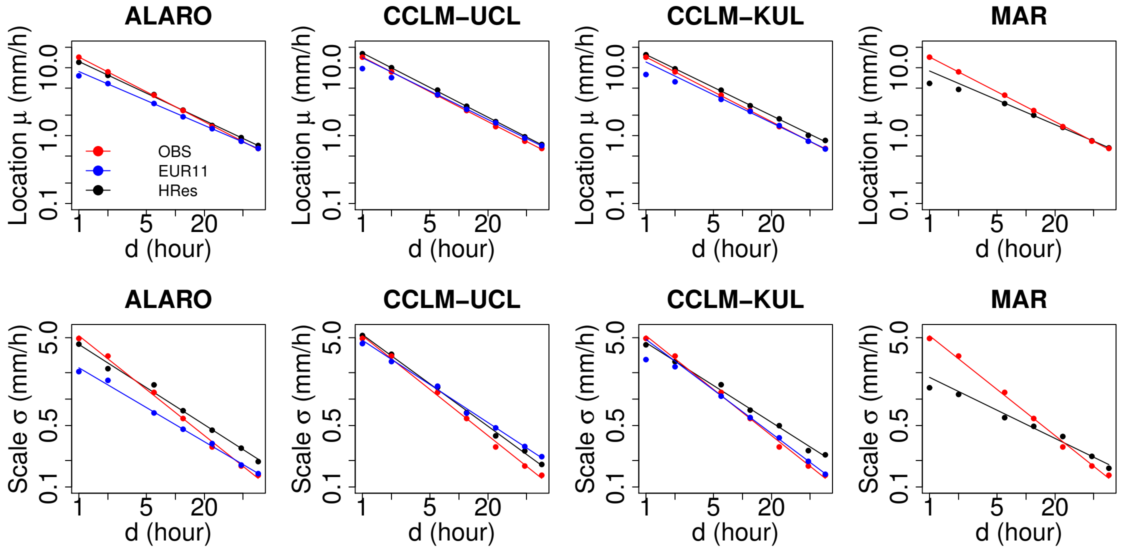

FIG. 9. GEV-parameters (location/scale) plotted against rainfall duration d, for station Uccle. First row: location. Second row: scale. Points: maximum likelihood estimators $(d_i, \hat{\mu}_i)$ and $(d_i, \hat{\sigma}_i)$. Solid lines: generalized least squares fits to the estimated GEV-parameters.

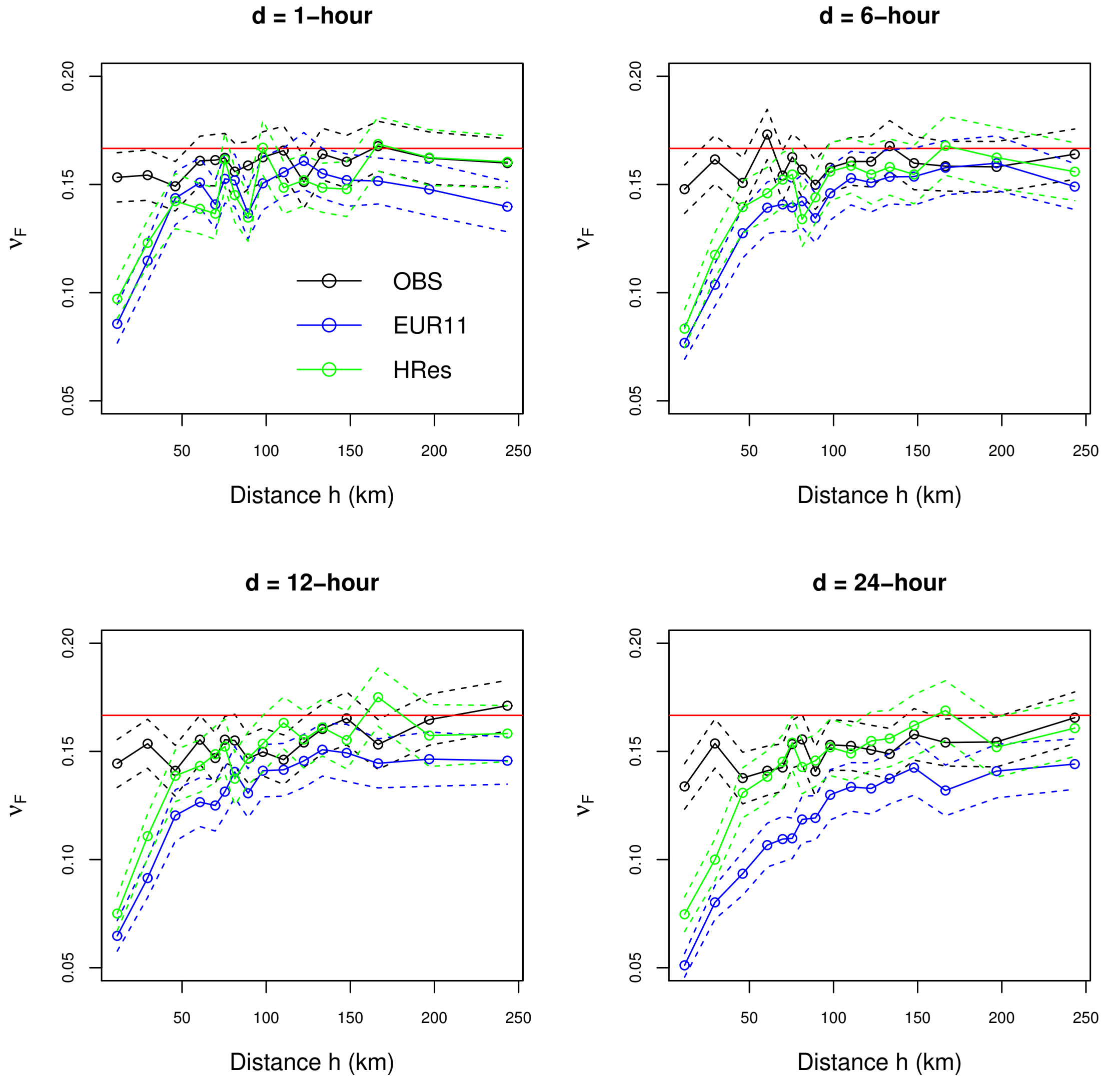

FIG. 10. Spatial extremal dependence: the madogram $\nu_F(h)$, Eq. (7), as a function of distance h for (i) the observed annual maxima at the 18 pluviograph stations, and for (ii) the 18 gridpoints from ALARO. Dots: binned empirical madogram, Eq. (8). Dashed lines: 95%-confidence bounds of the empirical madogram. Red line: independent case, $\nu_F = 1/6$.

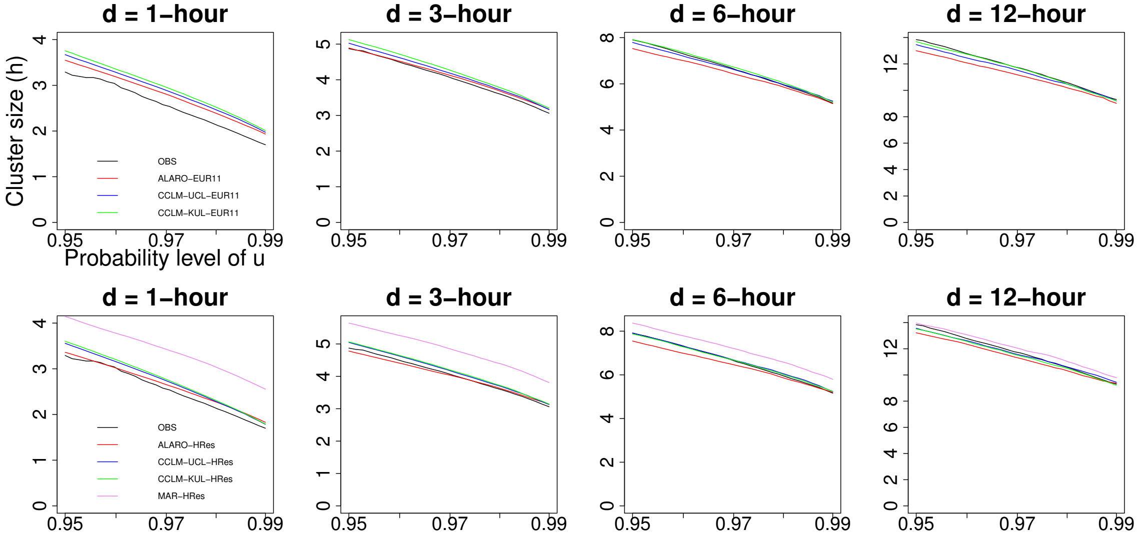

FIG. 11. Mean cluster size (unit: hour) against the threshold u, expressed in probability level. The runs estimator, Eq. (9), was used. Upper (lower) rows are for EUR-11 (H-Res) simulations.

Accepted for publication in *Journal of Applied Meteorology and Climatology*. DOI 10.1175/JAMC-D-21-0004.1.