Dear Editor,

Please find enclosed the reply to the interactive comments on the manuscript **"Simple Scaling of Extreme Precipitation in North America"** by Innocenti et al. published in HESS-D.

We provide below a detailed response to each of the referees' comments which are reported in blue for the first referee and in red for the second referee. A copy of the revised manuscript in "track changes" mode is also provided. In the "track changes" manuscript specific colors are used to link corrections and reviewers' comments:

- blue^{R1} is used to underline changes related to first referee's comments.
- while red^{R2} is used for changes related to the second referee's comments,
- and green is used for other changes.

Line numbering **(in bold)** refers to the revised manuscript with "track changes" attached to this reply.

Sincerely,

Silvia Innocenti, on behalf of the co-authors.

Authors' detailed response to 1st referee's comments

The article is overall well written, and the scientific subject is very topic and well addressed. The literature review and scientific context given in the introduction and Section 2 are very well written! The data description is also very clearly outlined. I had difficulties understanding (some technical details of) Section 4, and consequently (some of the results of) Sections 5 and 6. Therefore most of my comments aim to improve (my understanding) of this section. (I had also some perplexities about the chosen regions, which I address in the major comments too). I hope my comments lead to an improvement of the manuscript.

Major comments:

1) It is very difficult for me to comprehend Section 4 (mainly page 7 and Figures 1 and 2)

1a) The definition of scaling interval needs to be clearer: from page 7, lines 7-8, I understand that for each data-set (SD, ID and LD) you define several scaling intervals. These intervals have fixed durations equal to 6, 12,18 and 24 times the reference duration unit d* (which is 15', 1h and 6h, respectively, for each data-set SD, ID and LD). I assume the 6, 12, 18 and 24 durations are associated to a dilatation factor of = 6, 12, 18, 24 (the same as in equations 2,3,4). The several scaling intervals you consider have all the same length (duration) but are distinct for their initial time. If this is correct, the text at line 11 (and thereafter) need to be revised: I suggest replacing "its first duration" with "its initial time", because using the word "duration" for both (initiation and duration) indices is very confusing.

We agree that the definition of scaling interval needed to be clarified, probably because the concept of reference duration d^*, initial duration d_1, scale ratio λ, and interval length were confusing [see also comment 1) of J. Blanchet, second referee].

For all theoretical developments presented in Sect. 1, 2, and 4, we considered the reference duration $d^* = 1$ [defined at **Line 27, Page 5**] to express the scale ratio λ [i.e. the ratio between two durations defined at **Line 20, Page 3**]. Choosing $d_* = 1$, the scale ratio $\lambda = d/d^*$ can be simplified to $\lambda = d$ [as stated at **Line 27, Page 5** in the original version of the paper, and also at **Lines 5 to 11, Page 4** in the revised manuscript]. Hence, we can express the Moment Scaling Analysis (MSA) regression coefficients and the GEV parameters as functions of d only [see Eq.(4), Eq. (7), and reply to comment 1b)].

In Sect. 4, for all the SD, ID, and LD datasets and all scaling intervals, $d^* = 1\ h$ has then been used as reference duration for defining the SS samples $\boldsymbol{x}_{d^*}$ [i.e. the samples of all AMS observation rescaled at $1h$ using Eq. (8)]. To clarify the definition of $\boldsymbol{x}_{d^*}$, **Lines 17 to 12, Page 8** have been rephrased [please refer to our reply to comment 1e)].
In Sect. 6, $d^* = 1\ h$ has also been used as the reference duration for estimating the SS-GEV parameters μ_*, σ_*, and ξ_* from the SS sample $\boldsymbol{x}_{d^*}$. This has been described at **Lines 27 to 30, Page 15**. However, without loss of generality, one could have been chosen $d_* \neq 1h$ with the only effect of rescaling the SS-GEV parameter values and without affecting the shapes of the estimated distributions. For this reason, while Fig. 8 (a) and (b) present the distribution of the SS-GEV parameters for $d^* = 1h$ for the SD and ID dataset, in Fig. 8 (d) we rescaled μ_*, σ_*, and ξ_* at $d_* = 24h$ for comparing SS- and Non-SS-GEV estimates, as described at **Line 13, Page 16**.

The definition of the scaling intervals, on the contrary, does not depend on λ nor d^*. Each scaling interval is defined by the following three characteristic:

1. *The initial duration* d_1, corresponding to the first/smallest duration included in the scaling interval [see definition at **Line 5, Page 4** and the examples in Fig. 1 d)]; note that d_1 is effectively a "duration" [as in AMS definition] and not an "initial time" [which sounds more like the time at which the interval begins]. Hence, the terminology "first duration" has been kept.

2. *The time-step* used to construct each dataset [i.e., time-increment between successive durations for which we constructed AMS]: 15min for the DS dataset, 1h for ID, and 6h for LD, respectively [see **Line 33, Page 6** and Table 2)].

3. *The interval length*, i.e. the number of consecutive durations [either 6, 12, 18 or 24] included in the scaling interval $[d_1, d_D]$ [see **Line 28, Page 7**]: in other words, the scaling intervals do not all have the same length, as you suggested. For this reason we used the terminology "interval length", instead of "interval duration", which could be confusing.

In order to clarify the terminology and notations, the following modifications have been made in Sect. 2 and in the definition of the scaling intervals at page 7:

- We eliminated the notation $D = \lambda d$, keeping λd only, as you also suggested in minor comment 2) [**Line 20, Page 3** and following paragraphs]. Accordingly we also modified Eq. (2)-(4).

- We modified the discussion of Eq. (4) to [**Lines 5 to 11, Page 4**]:

  "Moreover, without loss of generality, λ can always be expressed as the scale ratio $\lambda = d/d^*$ defined for a reference duration d^* chosen, for simplicity, as $d^* = 1$. Therefore, the SS model can be estimated and validated over a set of durations $d_1 < d_2 < .. < d_D$ by simply checking the linearity in a log-log plot of the X moments versus the observed durations d_j, $j = 1, 2, \ldots, D$ [see, for instance, Gupta and Waymire, 1990; Burlando and Rosso, 1996; Fig. 1 of Nhat et al, 2007; and Fig. 2 (a) of Panthou et al., 2014]. If H estimated for the first moment equals the exponents (slopes) for the other moments, the precipitation intensity X can be considered scale invariant under SS in the interval of durations d_1 to d_D."

- We rearranged the definition of the scaling intervals by explicitly mentioning that d_1 corresponds to the first duration of the 6, 12, 18, or 24 durations included within each scaling interval, by improving the definition of "interval length", and by adding other examples of scaling intervals [**Line 27, Page 7 to Line 5, Page 8**]:

  "In order to identify possible changes in the SS properties of AMS distributions, various scaling intervals were defined for the MSA. In particular, all possible subsets with 6, 12, 18 and 24 contiguous durations were considered within each dataset. Figure 1 and Figure 2 show the 136 scaling intervals thereby defined: 40 scaling intervals for SD and IS, and 56 scaling intervals for LD. For instance, the first matrix on the left of Fig. 1(a) presents the 6-duration scaling intervals 15min - 1.5h, 30min - 1.75h, …, 4.75h - 6h defined for the SD dataset [i.e. the 19 scaling intervals containing six contiguous durations defined with a 15min increment], while Fig. 1(d) shows an example of the first four 6-duration scaling intervals for the ID dataset [i.e., 1h-6h, 2h-7h, 3h-8h, and 4h-9h, containing six contiguous durations defined with an increment of 1h]. This procedure was defined in order to evaluate the sensitivity of the SS estimates to changes in the first duration d_1 of the scaling interval and in the interval length [i.e. the number of durations included in the scaling interval]."

1b) I am trying to understand how you estimate H: I understand that you use Eq. 4, with a fixed q and a fixed (say, 6h). You have to regress however on more than one estimated moment, so I assume you consider the aforementioned scaling intervals with the different initial times (e.g. for the ID data-set and 6h-duration you consider the 19 6h-long scaling intervals you show in Figure 1b, left panel). For each (of the 19) initial times, you select the annual max, and then an AMS, from which you compute the q- moment (so you have 19 q-moments for the LHS of equation 4). What do you use for the first term of the RHS of Equation 4? will you have 6×19 q-moments corresponding to the AMS of 1h accumulated precipitation within each 6h long scaling interval?

We apologize for this lack of clarity but the estimation of the scaling exponent H is not based on the methodology you described. In particular, each scaling interval does not represent a period of time but an interval of durations [see the reply to the previous comment] and the estimation of H for one scaling interval is independent from the

estimation over other scaling intervals.

In fact, for each moment order q and each scaling interval, the scaling exponent can be estimated through Eq. (4), by the use of a linear regression between $E[X_d^q]$ and the scale ratio λ in a log-log plot [see **Lines 5 to 17, Page 4**]. Moreover, since $d_* = 1$ implies $\lambda = (d/d_*) = d$, as stated at **Line 27, Page 5**, the MSA regression can be computed for each scaling interval by simply using $\ln(d)$ as covariate for $\ln(E[X_d^q])$. In other words, for each q we estimate the following linear regression model:

$$Y_j = \alpha + K_q Z_j$$

where $Y_j = \ln(E[X_{d_j}^q])$, $Z_j = \ln(d_j)$, and $j = 1, 2, \ldots, D$. In practice, the empirical $q-$moments $\left\langle X_{d_j}^q \right\rangle$ of X_{d_1},X_{d_2}, ... , X_{d_D} were used for defining Y_j. This is a standard procedure called Moment Scaling Analysis (MSA) which is used to estimate the scaling exponents K_q (the linear regression slopes) for different moment orders, and, at the same time, to validate the linearity of the scaling exponents with q, i.e. to test if $K_q \approx Hq$ [see, for instance, Gupta and Waymire, 1990; Burlando and Rosso, 1996; Fig.1 of Nhat et al, 2007; and Fig. 2 (a) of Panthou et al, 2014].

In our application, we estimated the scaling exponents K_q as the slope of the MSA regression for fifteen moment orders $q = 0.2, 0.4, \ldots, 2.8, 3$ as described at **Lines 6 to 11, Page 8**. Then we checked the linearity condition $K_q \approx Hq$ using the "slope test" [**Line 12, Page 8 to Line 16, Page 8**]: a second regression model, $K_q = \beta_0 + \beta_1 q$, was fitted between the fifteen estimated values K_q and the moment orders q. Then, a Student's t-test was used to test the null hypothesis $\boldsymbol{H}_0$: $\beta_1 = K_1$, i.e to test if β_1 is equal to the scaling exponent estimated for $q = 1$ [i.e. for the mean of the AMS]. If $\boldsymbol{H}_0$ was not rejected at the significance level α = 0.05, the SS assumption $K_q \approx Hq$ was considered appropriate for the specific scaling interval and the simple scaling exponent $H = K_1$ was retained.

To improve the description of the MSA procedure and the slope test, we modified **Lines 5 to 11, Page 4** [see reply to comment 1a)] and we added a few details from **Line 6, Page 8 to Line 16, Page 8**.

1c) Page 7, line 16 "and the corresponding durations" are different durations, or different initial times for a fixed duration ?

Please see our replies to previous comments: for each scaling interval including durations $[d_1, d_D]$, we constructed the AMS for durations $d_1, d_2, \ldots, d_D$ while λ only represents the scale ratio $\lambda = d/d^*$ needed to express X_d [i.e. the variable "extreme precipitation intensity observed over time interval of duration d"] as a function of X_{d^*}. Note also that, throughout the manuscript, the word "duration" always refers to the AMS temporal scale. We hope that the modifications made for addressing the previous comments have already clarified these points.

1d) I am not sure I understand what are you testing with the slope test. My hypothesis is that you have evaluated the regression coefficient $Kq = -Hqlog(\lambda)$ (the equality is from equation 4). Since the final goal is to estimate H, you want to regress the Kq versus q (for all the q=0.2,0.4, .. ,3), with a fixed λ, to finally find H. Please explain this better.

The slope test described at **Line 12, Page 8 to Line 16, Page 8** is used to validate/invalidate the SS hypothesis $K_q \approx Hq$; if the slope test did not reject the null hypothesis of linearity of the MSA regression slopes with q, $H = k_1$ was retained. Please also see our reply to comment $1b$). However, modifications made in response to previous comments should have clarified these points.

1e) GOF test: Page 7, line 23, "for each duration d": this time seems to refer to a real duration. Whereas four lines later "for all durations dj in the scaling interval" suggests dj are initial times. I do not understand what are you testing with the GOF test: what is the "pooled sample of the rescaled AMS x'_{d_j} for all durations dj in the scaling interval"?

I got completely lost in Eq. 8 and 9 ... (I came back to this page few times, in separate days, to make sure my lack of understanding was not due to a particular bad moment. This is why I am describing in details what I do not understand, I hope this helps to point out what needs to be rephrased).

Please, see replies to comments $1a$) et $1c$). Both expressions "for each duration d" and "for all durations d_j in the scaling interval" were used to effectively refer to "durations" since AMS were constructed for each duration $d_1, d_2, \ldots, d_D$ included in a given scaling interval $[d_1, d_D]$ (identified by its first duration d_1).

Since the scale invariance property states that $X_{d^*} \overset{d}{=} d^H X_d$ [Eq. (2) with $\lambda = d/d^* = d$], it is possible to rescale $\boldsymbol{x}_{d_1}, \ldots, \boldsymbol{x}_{d_j}, \ldots, \boldsymbol{x}_{d_D}$ (each of these vectors representing the AMS observed for the duration d_j included in the scaling interval $[d_1, d_D]$) to the reference $d^* = 1h$ by simply applying the rescaling factor d^H. Accordingly, if D durations d_j, with $j = 1, 2, ..., D$, are included in the scaling interval $[d_1, d_D]$, the SS hypothesis implies that the rescaled AMS $\boldsymbol{x'}_{d_1}, \ldots, \boldsymbol{x'}_{d_j}, \ldots, \boldsymbol{x'}_{d_D}$ can be pooled in a single sample $\boldsymbol{x'}_{d^*}$, where:

- for each d_j, $\boldsymbol{x'}_{d_j}$ is defined according to Eq. (9) as the samples of observations $x_{d_j,1}, x_{d_j,2}, \ldots, x_{d_j,i}, \ldots x_{d_j,n}$ sampled at duration d_j and rescaled to d^* through the rescaling factor ${d_j}^H$:

$$\boldsymbol{x'}_{d_j} = \left({d_j}^H x_{d_j,1}, {d_j}^H x_{d_j,2}, \ldots, {d_j}^H x_{d_j,i}, \ldots {d_j}^H x_{d_j,n}\right)$$

- and $\boldsymbol{x}_{d^*} = \left(\boldsymbol{x'}_{d_1}, \ldots, \boldsymbol{x'}_{d_j}, \ldots, \boldsymbol{x'}_{d_D}\right)$ [Eq. (8)] is the pooled sample of AMS for durations $d_1, d_2, \ldots, d_D$ rescaled to the reference duration d^* [i.e., $\boldsymbol{x}_{d^*}$ contains $n \times D$ observations rescaled at $d^* = 1h$].

For this reason, the AD and KS tests have been applied to compare the cdf of $\boldsymbol{x}_{d,ss} = d^{-H}\boldsymbol{x}_{d^*}$ [i.e. the pooled sample $\boldsymbol{x}_{d^*}$ scaled back to the duration d] to the cdf of the observed $\boldsymbol{x}_d$ at its original scale d, as a second test for the validity of the SS hypothesis.

In order to clarify this point the paragraph describing GOF tests has been rewritten as [**Line 17, Page 8 to Line 12, Page 9**]:

" Goodness-of-Fit (GOF) test: for each duration d, the goodness of fit of the X_d distribution under SS was tested using the Anderson-Darling (AD) and the Kolmogorov-Smirnov (KS) tests. These tests aim at validating the appropriateness of the scale invariance property for approximating the X_d cdf by the distribution of $X_{d,ss} = d^{-H}X_{d^*}$. To this end, the pooled sample

$$\boldsymbol{x}_{d^*} = \left(\boldsymbol{x'}_{d_1}, \ldots, \boldsymbol{x'}_{d_j}, \ldots, \boldsymbol{x'}_{d_D}\right) \qquad (8)$$

of the D rescaled AMS, $\boldsymbol{x'}_{d_j}$, was used to define X_{d^*} under the SS assumption, considering all the durations d_j, with $j = 1, \ldots, D$, in the scaling interval. Each rescaled sample $\boldsymbol{x'}_{d_j}$ of the annual maxima $x_{d_j,i}$, $i = 1, \ldots, n$, observed for the duration d_j was obtained by simply inverting Eq. (2) for d^* :

$$\boldsymbol{x'}_{d_j} = \left({d_j}^H x_{d_j,1}, {d_j}^H x_{d_j,2}, \ldots, {d_j}^H x_{d_j,i}, \ldots {d_j}^H x_{d_j,n}\right) \qquad (9)$$

where n represents the number of observations (years) available for each duration. In this way, $n \times D$ rescaled observations were included in $\boldsymbol{x}_{d^*}$.
As in previous applications (e.g., Panthou et al., 2014), the AD and KS tests were then applied at significance level $\alpha = 0.05$ to compare the empirical distributions (Cunnane plotting formula, Cunnane, 1973) of the SS and non-SS samples, $\boldsymbol{x}_{d,ss} = d^{-H}\boldsymbol{x}_{d^*}$ and $\boldsymbol{x}_d$. "

1f) Page 8, lines 10-11: here duration seems again to refer to a time duration. So, in my understanding, you consider 1h, 2h, 4h, 5h and 6h; from these you evaluate H; then you evaluate the 3h AMS with the SS. Then you evaluate the RMSE for the high quantiles of this synthetic 3h AMS versus the high quantiles of the empirical 3h AMS. I am

not sure this is correct, but this is my guess .

Yes, the cross-validations was constructed as described in your example.

I think it would be very useful if you could show a figure with one (or maybe two) concrete example of your regression, for a fixed q and , so that the reader can better understand what you regress, and what are the statistical tests you perform to be confident in your results.

We added a reference to two graphical representations of the MSA procedure and SS equality $K_q \approx Hq$ at **Line 17, Page 4**:

"... see, for instance, Gupta and Waymire (1990), Burlando and Rosso (1996); Fig. 1 of Nhat et al, 2007; and Fig. 2 (a) of Panthou et al., 2014 ".

However, considering that the MSA is the standard tool for estimating and validating the SS models, we decided not to increase the number of figures in the paper.

- Nhat, L. M., Y. Tachikawa, T. Sayama, and K. Takara (2007), *A simple scaling charateristics of rainfall in time and space to derive intensity duration frequency relationships*, Ann. J. Hydraul. Eng, 51, 73–78.
- Panthou, G., T. Vischel, T. Lebel, G. Quantin, and G. Molini (2014), *Characterising the spacetime structure of rainfall in the Sahel with a view to estimating IDAF curves*. Hydrol. Earth Syst. Sci., 18(12), 5093–5107, doi:10.5194/hess-18-5093-2014.

2) When you perform your analysis, in my understanding, you consider scaling intervals of a duration of 6h (as an example) spanning the whole diurnal cycle as intervals with equivalent statistical properties. However, physically, precipitation occurring in the afternoon (which in the summer is triggered by convection) can present very different temporal structure (and physical properties) than precipitation occurring early in the morning: can you pull together these data, or (given that your focus is on extremes, which in summer often is related to convective events), should you consider a stratification based on the diurnal cycle? (The inhomogeneity of your phenomena along the dyurnal cycle might be also the cause of test rejections for longer scaling intervals, as you state at page 8, line 28-29)

We agree that the temporal structure of the observed precipitation events can be highly affected by the time of the day at which the events occur. This could be the case, for instance, for regions and time of the year when convection is the main generating process of extreme rainfall events. Accordingly, this will also be more important for short duration extreme events (e.g. duration less than an hour) for which convection is the main driver. In this regard, the analysis of the statistical characteristics of the temporal processes of precipitation should at some point consider the diurnal and seasonal precipitations cycles. However, in order to analyze the impact of these cycles on extreme precipitation intensity, one should consider methods and datasets which are different than the ones considered in our study. For instance one would have to deal with the time of the year and the time of the day at which the precipitation events occurred, as you suggested. Our study focuses, instead, on Annual Maxima Series (AMS) that do not consider the time at which the precipitation occurred but only the temporal scale (duration) over which the precipitation has been observed. This correspond to the classical definition of AMS typically used for constructing IDF curves and for scaling analysis [e.g., *Burlando and Rosso, 1996; Koutsoyiannis et al., 1998; CSA, 2012; Panthou et al., 2014*]. Even if it would be interesting to analyze the occurring time and other characteristics of the events from which annual maxima have been extracted [Sect. 5 of our study, for instance, addresses some of these issues], our extreme precipitation analysis did not considered the event-based definition of extremes needed to analyze the diurnal and seasonal cycles of precipitation.

For these reasons, in our opinion it is not possible to connect the higher proportion of GOF test rejections for longer scaling intervals [i.e., scaling interval considering an higher number of durations] to the diurnal cycle of rainfall, as you suggested. Instead, we can affirm that, for longer scaling intervals, there exists an *" higher probability of observing large differences between $\boldsymbol{x}_d$ and $\boldsymbol{x}_{d,ss}$ quantiles when $\boldsymbol{x}_{d,ss}$ had larger sample size and included data from more distant durations. "* [see **Line 12, Page 10**;].

- Koutsoyiannis, D., D. Kozonis, and A. Manetas (1998), A mathematical framework for studying rainfall intensity-duration-frequency relationships, Journal of Hydrology, 206(1-2), 118–135, doi:10.1016/S0022-1694(98)00097-3.
- Burlando, P., and R. Rosso (1996), Scaling and multiscaling models of DDF for storm precipitations, Journal of Hydrology, 187, 45–64.
- CSA (2012), Technical guide: Development, interpretation and use of rainfall intensity- duration-frequency (IDF) information: Guideline for Canadian water resources practi- tioners, Tech. Rep. PLUS 4013 - 2nd ed.
- Panthou, G., T. Vischel, T. Lebel, G. Quantin, and G. Molini (2014), *Characterising the spacetime structure of rainfall in the Sahel with a view to estimating IDAF curves*. Hydrol. Earth Syst. Sci., 18(12), 5093–5107, doi:10.5194/hess-18-5093-2014.

3) Figure 1: your lowest confidence (largest proportion of rejected stations) is associated at the "durations" in the beginning of the scaling interval: why? This seems a sampling problem (is this what you "expect" in your statement at page 8, lines 26-27)? All the possible causes you list from page 8, line 29 to page 9 line 6 are plausible, but should hold also for 15m durations within the scaling interval (not just in the begin- ning). Possibly, my lack of understanding is linked to my lack of understanding on how the calculation is performed (page 7).

We agree that largest proportions of rejected stations occurring for the shortest durations of the scaling intervals can be partially explained by a sampling effect. In particular, as you mentioned, our interpretation of this result underlined three possible causes [**Line 14, Page 10 to Line 24, Page 10**] :

1. GOF tests could be more often rejected due to a relatively more important presence of very large values in short-duration samples, i.e. a difference in statistical features of X_d between short and long durations that makes shorter durations more prone to rejection of GOFs.
2. When considering durations close to the temporal resolution of the recorded series [which is,generally 15 min ou 1 h], the measure of precipitation may be underestimated because intense rainfall events may be more likely split between two consecutive time steps. Obviously, this underestimation affect shortest durations more than longer durations.
3. Largest GOF test rejections for shortest series seem also to be connected to the coarser measurement resolution of 15PD series, which, similarly to the temporal resolution effect at point 2, induces larger measurement errors for shortest duration AMS.

These explanations were probably unclear due to the confusing definition of the scaling intervals reported in the previous version of the paper. The corrections made to sections 4 [the improvement of the paragraphs describing the construction of the scaling intervals and the scaling exponent estimation] should have now clarified this point.

4) Figure 2 (and related text, at page 9, lines 15-21): your largest relative errors are always associated at the durations in the beginning and at the end of the scaling interval: similarly to Figure 1, is it possible that this is a sampling problem? (Or maybe, for longer time scaling, the inhomogeneity of the weather phenomena along the diurnal cycle plays a role ..)

Thanks for the interesting comment. To our opinion, this result is not really surprising considering that Fig. 2 shows the cross-validation estimations of the normalized RMSE averaged over all valid SS stations. In a cross-validation setting, we expect that SS estimations are more impacted by the exclusion of a duration at the border of a scaling interval than by the exclusion of an inner duration.

Consider, for instance, the cross-validation estimations on the 6-duration scaling interval 1h-6h for the ID dataset. One would expect a greater change in SS estimation when the duration 6h is excluded than when an inner duration, e.g. 3h, is excluded. This is related to two factors:

i) an OLS linear regression is used for estimating H: excluding the last point of the regression (i.e. the last duration of the scaling interval) may have a greater impact on the slope estimation than excluding an inner point.

ii) the SS sample $\boldsymbol{x}_{d,ss}$ pools and rescales observations coming from AMS observed over various durations: approximating the X_{3h} distribution with rescaled observations from both durations $\leq$ 3h and $\geq$ 3h uses, in average, more information than approximating X_{6h} with rescaled observations coming from durations $\leq$ 6h only .

Moreover, as stated at **Line 5, Page 11**, the quality of the SS approximation seems also to deteriorate with decreasing d_1 and with increasing scaling interval length due to an effective decrease of the model performances.

In order to complete the discussion of Fig. 2, we added a reference to the greater sensitivity of SS estimation to the exclusion of durations at the border of the scaling interval [**Lines 1 to 5, Page 11**]:

"Larger errors were observed for durations at the border of the scaling intervals. Not surprisingly, this result underlines that, in a cross-validation setting, both the MSA estimation of H and the $X_{d,ss}$ approximation are less sensitive to the exclusion of an inner duration of the scaling interval than to the exclusion of d_1 or d_D. Conversely, the extrapolation under SS of the X_d distribution is generally less accurate if d is outside the range of durations used to estimate H. ".

5) Figure 3: I find it very difficult to understand the results shown in Figure 3:

5a) It is not well set what is the scope of this section is (in fact, at my first reading, I had the feeling it was a aimless technical analysis ... which instead is not). Later, I came to this hypothesis: in my understanding, H should be a scale-invariant parameter. Therefore large changes in H (aka large ΔH) are "bad", whereas if ΔH is near zero the SS model is a good approximation. Is this correct? Can you please state this clearly in the beginning of this section. Then, the reader will be able to search for the wanted results while analyzing Figure 3.

In order to clarify the topic of the section and the aim of our analysis we added the following introductory paragraph at **Lines 11 to 13, Page 11**:

" In order to evaluate the sensitivity of SS to the considered scaling interval, the variability of H with d_1 has been analyzed. Then, the spatial distribution of the scaling exponents for each scaling interval was studied to assess the uncertainty in H estimation and the dependence of SS exponents on local geoclimatic characteristics."

Then, we added the following explanations to the definition of Δ_H:

- *" Investigating the variability of the scaling exponent with the scaling interval is particularly important since, if SS is assumed to be valid between some range of durations, one should expect that H remains almost unchanged over the various scaling intervals included in this range. For this reason, the variation $\Delta_{H_{(j)}}$ of the scaling exponents computed for overlapping scaling intervals having the same d_1 but different lengths was analyzed. "* [**Lines 14 to 17, Page 11**]

- *"If SS is appropriate over a range of durations, $\Delta_{H_{(j)}}$ is expected to be small for scaling intervals defined within this range. "* [**Line 25, Page 11**]

5b) Technical question: (the distribution of) H is computed for each duration (e.g. 1h, 2h, 3h, ... in Figure 3 i-b). From Equations 2,3,4 I understood that H is scale invariant; then there should be one H which enables to describe all time scales. Why this is not the case? (Similar for the following sections, e.g Figure 8). Again, it is possible that my lack of understanding is linked to my lack of understanding on how the calculation of H is performed (page 7).

The distribution of H over stations is presented in Fig 3(i) for each d_1, i.e. for each scaling interval having d_1 as first duration [please, see replies to comments 1a) and 1b)]. In fact, only one value of H is estimated for all durations included in a scaling interval [as you suggested, H enables to describe all time scales with only one rescaled distribution], and this is not in contradiction with what is showed in Fig. 3. The misunderstanding is due to the fact that d_1 is used to identify a scaling interval and not any general duration d: once the dataset and the scaling interval length are fixed, the first duration d_1 is used to identify the scaling interval that includes the durations $d_1, d_2, \ldots, d_D$.

To simplify result interpretation, we added the label "d_1" to all x-axis of Fig. 3, 8 [corresponding to Fig. 9 in the previous version], and 11 [Fig. 12 in the previous version].

5c) I suggest as new title for Section 4.2: "Variability of the estimated scaling exponents" (to mirror the results shown in Figure 3, which shows ΔH).

The title of the subsection 4.3 has been modified to *"Estimated scaling exponents and their variability"*.

6) Section 5: from the maps you show in Figures 4 and 5, the only two clearly distinct homogeneous regions (at a first eye analysis) seem to be SW_pacif and NW_pacif. I can see (the signal is more mild) also the regions C and E. It seems to me that the South-East of the United Stated could also be split in two regions (e.g. from Fig 4 ID and LD, and figure 5 ID). The Boreal region (as you conlude yourself at page 12, line 16) is very etherogeneous, and I do not see any reason for clustering these stations together (sole common factor is the network sparseness ....). Have you attempted a cluster analysis to define your own regions, rather than considering the Bukovsky regions?

We agree that for some scaling intervals (e.g., intervals showed in Fig. 5) many regions may appear quite heterogeneous. However, the regional analysis presented in Sect. 5 seems to confirm that, with the exception of A1 and B, the distribution of H within the considered regions is concentrated about its mean value for most of the scaling intervals . For instance, it seems clear from Fig. 8 that the variability of H in regions E and F is fairly low, while these regions contain the greater proportions of available stations. On the contrary, the heterogeneity observed in regions A1 and B could be probably connected to the low station density in these two large areas [as mentioned at **Line 12, Page 14**]. In this respect, we agree that a finer definition/separation of northern regions would improve region homogeneity.

However, our primary interest was to define a simple partition of the study area into regions with distinct climatic characteristics. In this sense, other climatological classifications could have also been considered. In other words, our analysis did not seek for a rigorous identification of "SS regions" but intended to perform a preliminary regional analysis based on the climatoilogical features that could influence SS regimes.

We agree that it would be interesting to apply methods, such as clustering, to further analyzed H spatial distribution. More refined methods would be necessary in oder to asses the homogeneity of the *Bukosky* (or other) regions in terms of the scaling exponent values and to precisely describe which climate and meteorological pocesses drive the local distribution of the H. However, the intent of our analysis was, at this point, mainly descriptive and aimed at validating (or not) the hypothesis that basic climatological characteristics may influence the spatial distribution of H.

7) Figure 7 and S4 are needed -in my view- solely for supporting the description of Fig.8d (page 12, lines 26-31). (The results related to the other panels are less interesting, in my view). I suggest to move also Figure 7 in the supporting material, along with the text at page 11, lines 11-22.

As suggested, the figure and the details of calculations of N_{eve} have been moved in the supplementary material: see **page 13**.

Minor comments

1) There is a typo at line 7 of the abstract: should be 15', and not 15h.

Corrected.

2) Page 4, equation 2: I suggest to eliminate "D" and explicitly write "λd" instead (the less symbols you introduce, the more readable is the article).

Correction made. Please see the reply to major comment 1a).

3) Page 5, line 1: eliminate “and the frequency ... F(x)”.

Done.

4) Page 5, line 9: I suggest writing “Approaches aimed at increasing the sample size may be used ... ”

Done.

5) Page 5, Equation 7: notation is too complex, and should be simplified.

Thank you for the suggestion but we think that no superfluous symbol is used in Eq. (7), which directly follow from Eq. (2) and the scale invariance property of the GEV expressed at **Line 25, Page 5**:

since $X \overset{d}{=} GEV(\mu,\sigma,\xi)$ implies $\lambda X \overset{d}{=} GEV(\lambda\mu,\lambda\sigma,\xi)$,

if one consider $X_{d^*} \overset{d}{=} GEV(\mu_{d^*},\sigma_{d^*},\xi)$, and Eq.(2) applies, i.e. $X_d = \lambda^{-H} X_{d^*} = d^{-H} X_{d^*}$,

then $X_d \overset{d}{=} GEV(d^{-H}\mu_{d^*}, d^{-H}\sigma_{d^*},\xi)$.

See also Blanchet et al (2016), Eq. (7), page 84 [J. Blanchet, D. Ceresetti, G. Molinie, J.-D. Creutin, A regional GEV scale-invariant framework for Intensity–Duration–Frequency analysis. Journal of Hydrology, Volume 540, September 2016, Pages 82–95.].

6) Page 6, lines 6-9: the cause-effect is not entirely clear to me: why two different periods where chosen (JJAS for the north and MJJASO for the south), rather the same period (either JJAS or MJJASO) for all stations?

For each station in the study region [i.e. no matter the latitude], a year was considered valid if it had minimally 85% valid values (otherwise it was considered as a missing year). Then, only valid years were considered for constructing AMS. However, many Canadian stations -especially those equipped with tipping bucket rain gage- do not record precipitation during winter period, namely from November to April for stations located south of the 52nd Parallel N and from October to May for stations located north of the 52nd Parallel N (CSA, 2012). This means that, without restricting the definition of "years" to the annual period during which stations are in operation, we could have not use the records from many northern stations. At the same time, the ’block maxima’ definition of extreme precipitation series need long annual periods (i.e. large blocks) to be consistent with the GEV approximation of AMS distribution. Therefore, a trade-off between "year length" and "number of valid stations" existed. We therefore chose a definition of ”year” (block) based on the latitude: for stations located north of the 52nd Parallel N we defined the year as the period from June to September (i.e. 122 days a year were considered), while for stations located south of the 52nd Parallel N we used the period from May to October (i.e. 184 days a year were considered). Note that the same criteria ware used for Canadian and US stations.

- CSA: *Development, interpretation and use of rainfall intensity-duration-frequency (IDF) information: Guideline for Canadian water resources practitioners*, Tech. Rep. Canadian Standard Association, Tech. Rep. PLUS 4013, Mississauga, Ontario, 2nd ed., http://shop.csa.ca/en/canada/infrastructure-and-public-works/plus-4013-2nd-ed-pub-2012/invt/27030802012,2012.

7) Page 6, line 13-14 (and thereafter): rather than saying “coarser resolution” I suggest writing “more discretized recording procedure” (or something similar). The effect of recording discretizations are a well known problem in statistics, which can be bypassed simply by adding a uniformly distributed random noise (ranging between 0 and 2.54) to your data.

Thank you for the suggestion but for consistence with the expression "temporal resolution" we kept the terminology "instrument resolution" for indicating the minimum amount of precipitation detectable from rainfall gauges [Note that these are two "resolutions" have similar impacts on our results]. For clarity, however, we changed *"resolution"* to *"instrument resolution"* when needed [e.g., **Lines 24 to 28, Page 6**, **Line 9, Page 18**, and **Table 1**].

Moreover we agree that several statistical methods exist for dealing with highly discretized data. Although we did not use such methods, we evaluate how this discretization may affect our results [some complementary analyses are presented in the supplementary material, Sect. S2 and Fig. S2-S3] and we considered it when applying our statistical analysis [e.g., when applying GOF test, see **Line 14, Page 9**]. However, we choose not to modify the raw data since it could possibly have, in our opinion, an unpredictable impact on our results [a non-quantifiable impact on AMS scaling].

8) Page 6, line 30: this condition is not clear to me, please explain it more explicitely.

From the definition of daily maxima [note that the DM dataset has been renamed Daily Maxima Precipiation Data (DMPD) according to minor comment 4 of the second reviewer; see **Line 9, Page 6**], if the intensity value $x_{d_1} \geq 0$ has been recorded at duration d_1 and the intensity value $x_{d2} \geq 0$ has been recorded at duration d_2, with $d_1 \leq d_2$, the following conditions must be met:

i) The rainfall intensity x_{d_1} observed for the shorter duration d_1 must be larger or equal to the intensity x_{d_2} observed over the longer duration d_2 (equality occurs if rainfall intensity would be constant during a period of time d_2): $x_{d_1} \geq x_{d2}$, i.e. $0 \leq \frac{x_{d_2}}{x_{d1}} \leq 1$.

ii) The rainfall depth $d_1 x_{d_1}$ recorded during the time period d_1 must be smaller or equal to the rainfall depth $d_2 x_{d_2}$ recorded during time period d_2 (equality occur if rainfall is recorded only during the time period d_1): $d_1 x_{d_1} \leq d_2 x_{d_2}$, i.e. $\frac{d_1}{d_2} \leq \frac{x_{d_2}}{x_{d_1}}$

Combining these two condition we have: $0 \leq \frac{d_1}{d_2} \leq \frac{x_{d_2}}{x_{d_1}} \leq 1$. Therefore, if maximum intensities recorded over the various durations within a given day do not satisfy this condition among pairs of durations, the values were considered "suspicious" and the day were discarded (i.e. all the daily maxima observed over the various durations for that day were discarded).

To clarify this issue, the paragraph has been rewritten as [**Lines 13 to 16, Page 7**]:

"For instance, each pair of DMPD intensity (x_{d_1}, x_{d_2}) observed at durations $d_1 < d_2$ must respect the conditions $x_{d_2}/x_{d_1} \leq 1$ and $d_1 x_{d_1} \leq d_2 x_{d_2}$ derived from the definitions of daily maxima rainfall intensity and depth; otherwise all DMPD values recorded that day were discarded and assimilated to missing data. "

9) Figures 1 and 2: I suggest using a notation as 2h30', rather than 2.5h. Similarly for the days, 2d12h (or 50h) rather than 2.5d.

Done. Note that a notation like "2h30min" instead of " 2h30' " has been used to respect HESS standard for units and figures.

10) Page 7, line 10-12: "This procedure ... evaluate the variability of the SS estimates... " could it be the sensitivity instead? This text is not clear.

The text has been changed integrating your suggestions [see also our reply to major comment 1a)]:

"This procedure was defined in order to evaluate the sensitivity of the SS estimates to changes in the first duration d_1 of the scaling interval and in the interval length [i.e. the number of durations considered]."

11) Page 7, line 20: student t-test (the t is associated to the test, not to the student).

Corrected.

12) Page 8, lines 14-20: r is usually used in statistics for correlation: it is possible to use a different notation? Please specify that the normalized RMSE is zero for a good fit, and it gets larger and larger for a worse fit.

The symbol ϵ has been substituted to r in the whole text [see, for instance, Eq. (10), (11), and Fig. 2]. Moreover, we integrated your second suggestion adding the following note at **Line 4, Page 10** :

"Note that the normalized RMSE is a measure of error, meaning that values of $\bar{\epsilon}_{x_{d,s}}$ closer to 0 correspond a better fit than larger

values. "

13) Page 9, line 23: the acronym for inter-quartile range is, traditionally, IQR.

Modified.

14) Figures 7,8 would be more easy to read if you put a title on each panel with the name of the region.

Thank for the suggestion but the list of the *Bukosky* regions considered in each panel was too long to be added to the legend. For this reason we named the regions with names (A1), (A2), (B), ..., (F).

For clarity, we added "Region" in each panel and the following sentence to the legend of Fig 7 [corresponding to Fig. 8 of the original version of the paper]:

"See Fig. 6 for region definition."

15) Page 11 lines 6-9: join this paragraph to the previous (they both pertain to the physical explanation of the different panels of Figure 8).

Done.

16) Page 12, lines 2-3: this apply to the SWpac region as well.

We agree that the particular topography characterizing the pacific coast may also impact the results of the curves in Fig. 7 (d). However the different synoptic regime characterizing the south-west areas of the continent seems to be the most important factor differentiating the curves observed for the northern and southern parts of the west coast [according to N_{eve} and T_{wet} results]. To underline this point, the following comment has been added at **Line 33, Page 14** :

"These results suggests that both the distinctive topography of the west coast and the characteristic large-scale circulation of the south-west areas of the continent are crucial factors determining the transition between the two scaling regimes in region D. "

17) Page 13, lines 10-12: I agree that scaling regimes are weaker for short d1 than for longer d1, however you need to rephrase the sentence at line 12. In fact, de- spite “smaller”, the scaling regimes for short duration exceed 0.5 (most of them 0.6). Therefore effect of the scaling factor ($\lambda\hat{}-H$) is not negligible on the AMS distribution moments.

The sentence has been simply rephrased to [**Line 14, Page 15**] :

"In general, the weakest scaling regimes were observed for short d_1 and along the west coast of the continent and seem to be connected to scaling intervals and climatic areas characterized by homogeneous weather processes. "

Authors' detailed response to 2nd referee's comments

The article is well written and mainly clear. There is a substantial amount of work and many interesting results. However

1) Although Sections 1 and 2 are very clear, I had at first reading some difficulties understanding the rest of the paper, mainly because I got confused with the concepts of "d1" and "interval length". For example, if I understood correctly, an interval length of 6 durations with d1=1h for SD corresponds to durations 1h, 1h15, . . ., 2h30, whereas an interval length of 6 durations with d1=1h for ID corresponds to durations 1h, 2h, . . ., 6h. This may be confusing, so the authors may want to clarify these concepts, maybe giving examples or a table with the different intervals.

We apologize for this lack of clarity, which was also pointed out by the first referee [see our reply to comment 1a) of the first reviewer]. To improve the description of the set of scaling intervals and their characteristic d_1 (initial duration), length (number of durations considered), and time-step (the time-increment separating contiguous durations within each dataset, equal to 15min in SD, 1h in ID, and 6h in LD) we modified the paragraph describing these characteristics to [**Lines 27 to 5, Page 7**]:

"In order to identify possible changes in the SS properties of AMS distributions, various scaling intervals were defined for the MSA. In particular, all possible subsets with 6, 12, 18 and 24 contiguous durations were considered within each dataset. Figure 1 and Figure 2 show the 136 scaling intervals thereby defined: 40 scaling intervals for SD and IS, and 56 scaling intervals for LD. For instance, the first matrix on the left of Fig. 1(a) presents the 6-duration scaling intervals 15 min - 1.5 h, 30min - 1.75h, . . . , 4.75 h - 6 h defined for the SD dataset [i.e. the 19 scaling intervals containing six contiguous durations defined with a 15min increment], while Fig. 1(d) shows an example of the first four 6-duration scaling intervals for the ID dataset [i.e. 1 h - 6 h, 2 h - 7 h, 3 h - 8 h, and 4 h - 9 h, containing six contiguous durations defined with an increment of 1h]. This procedure was defined in order to evaluate the sensitivity of the SS estimates to changes in the first duration d_1 of the scaling interval and in the interval length [i.e. the number of durations included in the scaling interval]. "

2) The authors use databases with different measurement frequencies. So I expect, e.g., the 1h-annual maxima at a given location to be larger when they stem from accumulating 15min rainfall than hourly rainfall. Thus I'm concerned about all the comparisons mixing these different measurement frequencies: do we expect H for example to be the same for different measurement frequencies? Likewise for the GEV parameters. In a pretty related study, Blanchet et al 2016 addresses this issue.

We agree that the temporal resolution of observed series, as well as the measurement resolution of rain gauges, may have an impact on the estimations of the AMS and of the scaling exponents.

Note however, that for many stations both DMPD and HCPD, or both 15PD and HPD series are available over a common period [note also that the names of the 4 datasets have been changed according to your minor comment 4)]. For these stations, annual maxima measured at the two different temporal resolutions were compared and combined [for each year, the annual maximum value of the two AMS was retained] In this way, the impact of the temporal resolution should have been partially reduced. [note that, to improve the description of this preliminary step in the construction of our AMS **Lines 17 to 20, Page 7** have been rephrased. Please, see our reply to minor comment 5)].

At the same time, the double resolution effects [temporal and measurement resolution effects which could add and have a greater impact on shortest durations] may still affect the estimation of AMS Simple Scaling. In fact, some of these issues have been raised and briefly discussed while analysing GOF test results [see **Lines 17 to 24, Page 10**]. Some complementary analyses have also been presented in the supplementary material, Sect. S2 and Fig. S2-S3. [Note that, for completeness, we added the reference to *Blanchet et al (2016)* at **Line 23, Page 10**].

However, despite the obvious interest of studying the effects of these factors, it would have been difficult, with the available datasets, to separate the influence of the temporal and measurement resolutions on extreme precipitation inference from the effect of other factors, such as, sampling errors associated to the series length.
Moreover, pooling the four datasets allow the construction of a rainfall dataset with an exceptional extent and density for North America. Hence, we decided to use these four datasets to construct the three AMS dataset SD, ID, and LD which constitute a remarkable data source for the study of Simple Scaling of extreme precipitation at a regional scale.

For completeness, however, the following comment has been added to the Conclusion [**Line 11, Page 18**]:

"... and (these results) show the importance of a deeper analysis to evaluate the impact of dataset characteristics (e.g., their temporal and measurement resolutions, or the series length) on the scale invariant properties of extreme precipitation."

Detailed Comments

1 p.3 l.5 "a deeper analysis . . . needed": Blanchet et al 2016 make such a regional analysis in South of France. The study region is much smaller but rainfall variability seems quite comparable.

The reference has been added at **Line 4, Page 3**.

2 p.4 l.11 "$H_{intensity}$ and H_{depth}": not defined

The sentence has been rephrased in order to add the explicit definition of H_{depth} [**Line 20, Page 4**]:

"(note that for the rainfall depth the scaling exponent $H_{depth} = 1 - H$ applies)".

3 section 2.2: Blanchet et al. 2016 use a GEV-ML estimation in a single step.

The following paragraph has been moved from Sect. 6 to **Line 5, Page 6** [Sect. 2] in order to clarify that a one-step procedure can also be used for the estimation of the SS-GEV parameters:

"In a few other cases, a Generalized Additive Model ML (GAM-ML) framework (Coles, 2001; Katz, 2013) has also been used to obtain the joint estimate of H, μ_*, σ_*, and ξ_* through the introduction of the duration as model covariate (e.g., Blanchet et al, 2016). "

Note that this procedure has been tested in our preliminary analyses, as mentioned at **Lines 30 to 33, Page 15**. Please, see also our reply to minor comment 12.

4 section 3: it may be clearer for the reader to call the 4 databases 15PD, H1PD, H2PD and DPD.

As suggested, we homogenized the dataset acronyms changing DM to DMPD and H to HCPD [**Line 9, Page 6** and following paragraphs]. However, we kept the acronyms "HPD" and "15PD" for US datasets since these are the official acronyms used by the NOAA [http://www.ncdc.noaa.gov/data-access/land-based-station-data].

5 p.6 l.20: so if I understand correctly SD comprises stations from 15PD only, ID from both 15PD and HPD, and LD from both 15PD, 1HPD and DPD (DPD only for the du- ration intervals >=1day). I'm correct? The authors may want to clarify it. In which case, the authors are analysing annual maxima with different measurement frequencies, without taking this at all into account. I wonder how the results/parameters you're comparing later are really comparable.

We apologize for this lack of clarity but the stations included within each of the SD, ID, and LD dataset were selected according to a procedure slightly different from the one you mentioned. In particular, while it is correct that the SD dataset uses 15PD data only [**Line 6, Page 7**], both ID and LD datasets use all relevant series to construct AMS for the sampled durations [Please, see also Tables 1 and 2, and reply to major comment 2)]. To improve the description of dataset construction **Lines 17 to 20, Page 7** have been modified to:

(iv) For each selected station, annual maxima were extracted for each valid year and duration. For stations having both DMPD and HCPD series, or 15PD and HPD series, for each year, the annual maxima extracted from these two series were compared and the maximum value was retained as the annual maximum for that year.

Moreover, we agree that the inhomogeneity of the series temporal resolution should be taken into account when interpreting our results, as well as other factors such as the different series length, measurement resolution, etc [Please, refer to the reply to major comment 2)]. However, it is *a priori* difficult to asses and separate the impacts of each of these inhomogeneities on the SS estimation. In particular, in our opinion, it is difficult to rigorously relate one or some of these inhomogeneities to any specific feature observed for the H distribution or GEV parameters. Hence, we assumed that they will not globally nor significantly affect our results nor the main conclusions of our study. For completeness, however, a brief comment about this issue has been added in the Conclusion [see **Line 11, Page 18** and major comment 2)].

6 p.6 l. 23-26: Papalexiou and Koutsoyannis 2013 and Blanchet et al. 2016 consider also the rank of the observed maxima to decide whether they should consider it or not in the analysis.

As in Papalexiou and Koutsoyiannis (2013), the following criterion was used: for each series, observation that are at least one order of magnitude larger than the series second largest value were excluded. This procedure was repeated until the ratio between the two largest values of the series was less than an order of amplitude. This detail, as well as your suggested reference, has been added at **Line 7, Page 7**:

"Note that, in order to exclude outliers possibly associated with recording or measurement errors, extremely large observations were discarded and assimilated to missing data. In particular, as in some previous studies (e.g., Papalexiou and Koutsoyiannis, 2013; Papalexiou et al., 2013), an iterative procedure was applied prior to step (ii)-1) to discard observations larger than 10 times the second largest value of the series. "

7 p.8 l. 8: I don't understand what are the "SS" and "non-SS" samples.

The definitions of SS and non-SS samples have been added at **Line 27, Page 8**:

"As in previous applications (e.g., Panthou et al., 2014), the AD and KS tests were then applied at significance level $\alpha = 0.05$ *to compare the empirical distributions (Cunnane plotting formula, Cunnane, 1973) of the SS sample,* $\boldsymbol{x}_{d,ss} = d^{-H}\boldsymbol{x}_{d^*}$*, and the non-SS sample,* $\boldsymbol{x}_d$*."*

Then, we rephrased the sentence at **Line 16, Page 9** as:

"According to this approach, data in $\boldsymbol{x}_d$ *and* $\boldsymbol{x}_{d,ss}$ *were pooled and randomly reassigned to two samples having same sizes of the SS and non-SS samples. "*

8 Figures 1 and 2: it took me time to understand these figures, partly because the x-axis are not labeled. Please add the labels (d1?).

Done.

9 Figure 3: isn't there also an effect of measurement frequency in the plots for ID and LD?

We would appreciate the reviewer be more specific and explain how she concluded that an affect of measurement frequency can be seen in Figure 3. In our opinion, further analyses would be needed to evaluate this effect, as previously mentioned. Please, see also our replies to major comment 2) and minor comment 5).

10 section 6: do I understand correctly that "non-SS" cases mean that the GEV parameters are estimated using the data from d* only? Please make it clearer.

We apologize for this lack of clarity but the non-SS GEV parameters were not estimated using the data from d* only: for each duration d, non-SS GEV paramaters were estimated on the non-SS sample $\boldsymbol{x}_d$, independently from other durations . To clarify this point, **Lines 27 to 30, Page 15** have been modified to:

"In our study, the PWM procedure was applied to estimate SS-GEV parameters μ_*, σ_*, and ξ_* [Eq. (7)] from $\boldsymbol{x}_{d^*}$ [Eq. (8)]. For each duration d, PWM were also used to estimate non-SS parameters μ_d, σ_d, and ξ_d from each of the non-SS samples $\boldsymbol{x}_d$. "

11 Figure 4: it might be clearer for comparison to use the same US map for the three rows (the first row is different so far). Also there might be here an effect of the measurement frequency for LD and ID, although the spatial patterns are pretty coherent.

When using the same map limits for all maps of Fig. 4 and 5, the nine maps becomes really small while a lot of blank space is present in the first row because there are no stations in Canada and Alaska. The figure is thus less clear when the same latitude and longitude limits are used for the nine maps. Moreover, since the focus is the comparison of the maps placed in the same row (i.e., in the same dataset), we prefer to keep the map limit for SD as they are now.

Concerning the eventual effect of the measurement frequency, please see replies to major comment 2) and minor comments 5) and 9). We agree that spatial patterns are fairly homogeneous suggesting that the series temporal resolution has a weak impact on the estimation of H.

12 Figure 5: same as Fig. 4.

Please, see the reply to the previous comment.

12 p.13 l.26: So if I understand correctly, here you use the H estimated previously and estimation is just for the GEV parameters. Please make it clearer. Have you also tried to estimate all parameters at once (mu*, sigma*, xi*, H) with ML estimators as in Blanchet et al 2016 for example? Theoretically, this should reduce the bias.

Yes, SS-GEV μ_*, σ_*, and ξ_* presented in our results are estimated applying PWM on the rescaled sample $\boldsymbol{x}_{d^*}$ [i.e. using a two-step procedure, as described at **Lines 1 to 5, Page 6**]. To address the reviewer's comment, we rephrased **Lines 27 to 30, Page 15** as reported in our reply to minor comment 10.

The one-step procedure of *Blanchet et al* (2016) has also been tested. To point this out, the relevant paragraph has been rephrased as [**Lines 30 to 33, Page 15**]:

"Preliminary comparisons of various estimation methods [PWM, classical ML estimators, and one-step GAM-ML; see Sect. 2.2], showed that PWM slightly outperformed the other methods".

13 Figure 9: isn't there also an effect of measurement frequency in the plots for ID and LD?

Please, see our reply to previous comments, in particular reply to major comment 2) and minor comments 5) and 11).

14 Figure 10: idem

Please, see our reply to the previous comment.

15 Figure 11: please add in the legend "with Hosking test at level 5%"

Done.

- J. Blanchet, D. Ceresetti, G. Molinie, J.-D. Creutin, A regional GEV scale-invariant framework for Intensity–Duration–Frequency analysis. Journal of Hydrology, Volume 540, September 2016, Pages 82–95.
- Papalexiou, S.M., Koutsoyiannis, D., 2013. Battle of extreme value distributions: a global survey on extreme daily rainfall. Water Resour. Res. 49 (1), 187–201.

Simple Scaling of extreme precipitation in North America

Silvia Innocenti[1], Alain Mailhot[1], and Anne Frigon[2]

[1]Centre Eau-Terre-Environnement, INRS, 490 de la Couronne, Québec, Canada, G1K 9A9
[2]Consortium Ouranos, 550 Sherbrooke Ouest, Montrèal, Canada, H3A 1B9

Correspondence to: S. Innocenti (silvia.innocenti@ete.inrs.ca)

Abstract. Extreme precipitation is highly variable in space and time. It is therefore important to characterize precipitation intensity distributions at several temporal and spatial scales. This is a key issue in infrastructure design and risk analysis, for which Intensity-Duration-Frequency (IDF) curves are the standard tools used for describing the relationships among extreme rainfall intensities, their frequencies, and their durations. Simple Scaling (SS) models, characterizing the relationships among extreme probability distributions at several durations, represent a powerful means for improving IDF estimates. This study tested SS models for approximately 2700 stations in North America. Annual Maxima Series (AMS) over various duration intervals from 15 min ~~h~~[R1] to 7 days were considered. The range of validity, magnitude, and spatial variability of the estimated scaling exponents were investigated. Results provide additional guidance for the influence of both local geographical characteristics, such as topography, and regional climatic features on precipitation scaling. Generalized Extreme Value (GEV) distributions based on SS models were also examined. Results demonstrate an improvement of GEV parameter estimates, especially for the shape parameter, when data from different durations were pooled under the SS hypothesis.

1 Introduction

Extreme precipitation is highly variable in space and time as various physical processes are involved in its generation. Characterizing this spatial and temporal variability is crucial for infrastructure design and to evaluate and predict the impacts of natural hazards on ecosystems and communities. Available precipitation records are however sparse and cover short time periods, making a complete and adequate statistical characterization of extreme precipitation difficult. The resolution of available data, whether observed at meteorological stations or simulated by weather and climate models, often mismatches the resolution needed for applications (e.g., Blöschl and Sivapalan, 1995; Maraun et al., 2010; Willems et al., 2012), thus adding to the difficulty of achieving complete and adequate statistical characterizations of extreme precipitation.

The need for multi-scale analysis of precipitation has been widely recognized in the past (Rodriguez-Iturbe et al., 1984; Blöschl and Sivapalan, 1995; Hartmann et al., 2013; Westra et al., 2014, among others) and much effort has been put into the development of relationships among extreme precipitation characteristics at different scales. The conventional approach for characterizing scale transitions in time involves the construction of Intensity-Duration-Frequency (IDF) or the equivalent Depth-Duration-Frequency (DDF) curves (Bernard, 1932; Burlando and Rosso, 1996; Sivapalan and Blöschl, 1998; Koutsoyiannis et al., 1998; Asquith and Famiglietti, 2000; Overeem et al., 2008; Veneziano and Yoon, 2013). These curves are a standard tool for hydraulic design and risk analysis as they describe the relationships between the frequency of occurrence of

extreme rainfall intensities (depth) X_d and various durations d (e.g., CSA, 2012). Analysis is usually conducted by separately estimating the statistical distributions of X_d at the different durations (see Koutsoyiannis et al., 1998; Papalexiou et al., 2013, for discussions about commonly used probability distributions). The parameters or the quantiles of these theoretical distributions are then empirically compared to describe the variations of extreme rainfall properties across temporal scales.

Despite its simplicity, this procedure presents several drawbacks. In particular, it does not guarantee the statistical consistency of precipitation distributions, independently estimated at the different durations, and it limits IDF extrapolation at non-observed scales or ungauged sites. Uncertainties of estimated quantiles are also presumably larger because precipitation distribution and IDF curve parameters are fitted separately.

Scaling models (Lovejoy and Mandelbrot, 1985; Gupta and Waymire, 1990; Veneziano et al., 2007) based on the concept of scale invariance (Dubrulle et al., 1997), have been proposed to link rainfall features at different temporal and spatial scales. Scale invariance states that the statistical characteristics (e.g., moments or quantiles) of precipitation intensity observed at two different scales d and λd can be related to each other by a power law of the form:

$$f(X_{\lambda d}) = \lambda^{-H} f(X_d) \tag{1}$$

where $f(.)$ is a function of X with invariant shape when rescaling the variable X by a multiplicative factor λ and for some values of the exponent $H \in \mathbb{R}$. In the simplest case, a constant multiplicative factor adequately describes the scale change. The corresponding mathematical models are known as *Simple Scaling* (SS) models (Gupta and Waymire, 1990). SS models are attractive because of the small number of parameters involved, as opposed to multiscaling (MS) models which involve more than one multiplicative factor in Eq. (1) (e.g., Lovejoy and Schertzer, 1985; Gupta and Waymire, 1990; Burlando and Rosso, 1996; Veneziano and Furcolo, 2002; Veneziano and Langousis, 2010; Langousis et al., 2013). A single *scaling exponent* H is used to characterize the extreme rainfall distribution at all scales over which the scale invariance property holds. As a consequence, a consistent and efficient estimation of extreme precipitation characteristics is possible, even at non-sampled temporal scales, and a parsimonious formulation of IDF curves based on analytical results is available (e.g., Menabde et al., 1999; Burlando and Rosso, 1996; De Michele et al., 2001; Ceresetti, 2011).

Theoretical and physical evidence of the scaling properties of precipitation intensity over a wide range of durations has been provided by several studies. MS has been demonstrated to be appropriate for modeling the temporal scaling features of the precipitation process (i.e., not only the extreme distribution) and for the extremes in event-based representations of rainfall (stochastic rainfall modeling) (e.g., Veneziano and Furcolo, 2002; Veneziano and Iacobellis, 2002; Langousis et al., 2013, and references therein). These multifractal features of precipitation last within a finite range of temporal scales (approximatively between 1 hour and 1 week) and concern the temporal dependence structure of the process. They have been connected to the large fluctuations of the atmospheric and climate system governing precipitation which are likely to produce a "cascade of random multiplicative effects" (Gupta and Waymire, 1990).

At the same time, many studies confirmed the validity of SS for approximating the precipitation distribution tails in IDF estimation (for examples of durations ranging from 5 min to 24 h see Menabde et al., 1999; Veneziano and Furcolo, 2002; Yu et al., 2004; Nhat et al., 2007; Bara et al., 2009; Ceresetti et al., 2010; Panthou et al., 2014). This type of scaling is substantially

different from the temporal scaling since it only refers to the power law shape of the marginal distribution of extreme rainfall. Application of the SS models to precipitation records showed that the scaling exponent estimates may depend on the considered range of durations (e.g., Borga et al., 2005; Nhat et al., 2007) and the climatological and geographical features of the study regions (e.g., Menabde et al., 1999; Bara et al., 2009; Borga et al., 2005; Ceresetti et al., 2010, Blanchet et al, 2016). However, the application of the SS framework has been mainly restricted to specific regions and small observational datasets. A deeper analysis of the effects of geoclimatic factors on the SS approximation validity and on estimated scaling exponent is thus needed.

The present study aims to deepen the knowledge of the scale-invariant properties of extreme rainfall intensity by analyzing SS model estimates across North America using a large number of station series. The specific objectives of this study are: a) asses the ability of SS models to reproduce extreme precipitation distribution; b) explore the variability of scaling exponent estimates over a broad set of temporal durations and identify possible effects of the dominant climate and pluviometric regimes on SS; c) evaluate the possible advantages of the introduction of the SS hypothesis in parametric models of extreme precipitation.
The article is structured as follows. In Sect. 2 the statistical basis of scaling models is presented, while data and their preliminary treatments are described in Sect. 3. Sect 4 presents the distribution-free estimation of SS models and their validation using available series. Section 5 focusses on to the spatial variability of SS exponents and discusses the scaling exponent variation from a regional perspective. Finally, the SS IDF estimation based on the Generalized Extreme Value (GEV) assumption is discussed in Sect. 6, followed by a discussion and conclusions [Sect. 7].

2 Simple Scaling models for precipitation intensity

When the equality in Eq. (1) holds for the cumulative distribution function (cdf) of the precipitation intensity X, considered at two different durations d and λd ~~$D = \lambda d$~~[R1], Simple Scaling can be expressed as (Gupta and Waymire, 1990; Menabde et al., 1999):

$$X_d \overset{d}{=} \lambda^H X_{\lambda d}, \tag{2}$$

~~$X_D \overset{d}{=} \lambda^{-H} X_d$ (2)~~[R1]

where $H \in \mathbb{R}$ and $\overset{d}{=}$ means that the same probability distribution applies for X_d and $X_{\lambda d}$ ~~X_D~~[R1], up to a dilatation or contraction of size λ^H ~~$\lambda^{-H} = (D/d)^{-H}$~~[R1]. An important consequence of the SS assumption is that, if X_d ~~X_D~~[R1] has finite moments $E[X_d^q]$ ~~$E[X_D^q]$~~[R1] of order q, then X_d^q ~~X_D^q~~[R1] and $\lambda^H X_{\lambda d}^q$ ~~$\lambda^{-H} X_d^q$~~[R1] have the same distribution. Their moments are thus linked by the following relationship (Gupta and Waymire, 1990; Menabde et al., 1999):

$$E[X_d^q] = \lambda^{Hq} E[X_{\lambda d}^q]. \tag{3}$$

~~$E[X_D^q] = \lambda^{-Hq} E[X_d^q]$. (3)~~[R1]

This last relationship is usually referred to as the *wide sense* simple scaling property (Gupta and Waymire, 1990) and signifies that simple scaling results in a simple translation of the log-moments between scales:

$$\ln\{E[X_d^q]\} = \ln\{E[X_{\lambda d}^q]\} + Hq\ln\lambda \quad (4)$$

~~$\ln\{E[X_D^q]\} = \ln\{E[X_d^q]\} - Hq\ln\lambda$ (4)~~ R1

Moreover, without loss of generality, λ can always be expressed as the scale ratio $\lambda = d/d^*$ defined for a reference duration d^* chosen, for simplicity, as $d^* = 1$. R1 Therefore, the SS model can be estimated and validated over a set of durations $d_1 < d_2 < .. < d_D$ by simply checking the linearity in a log-log plot R1 of the X moments versus the observed durations d_j, $j = 1, 2, \ldots, D$ ~~the scale ratio λ in a log-log plot~~ R1 [see, for instance, Gupta and Waymire (1990); Burlando and Rosso (1996); Fig. 1 of Nhat et al. (2007); and Fig. 2 (a) of Panthou et al. (2014)]~~(Gupta and Waymire, 1990; Gupta and Waymire, 1996)~~ R1.

If H estimated for the first moment equals the exponents (slopes) for the other moments, the precipitation intensity X can be considered scale invariant under SS in the interval ~~range~~ R1 of durations d_1 to d_D.

More sophisticated methods have also been proposed for detecting and estimating scale invariance [for instance, dimensional analysis, Lovejoy and Schertzer (1985); Tessier et al. (1993); Bendjoudi et al. (1997); Dubrulle et al. (1997); spectral analysis and wavelet estimation Olsson et al. (1999); Venugopal et al. (2006) Ceresetti (2011); and empirical probability distribution function (pdf) power law detection Hubert and Bendjoudi (1996); Sivakumar (2000); Ceresetti et al. (2010)]. However, estimation through the moment scaling analysis is by far the simplest and most intuitive tool to check the SS hypothesis for a large dataset. For this reason, the presented analyses are based on this method.

According to the literature, the values of the scaling exponents H generally range between 0.4 and 0.8 for precipitation intensity considered at daily and shorter time scales (e.g., Burlando and Rosso, 1996; Menabde et al., 1999; Veneziano and Furcolo, 2002; Bara et al., 2009) (note that for the rainfall depth the scaling exponent $H_{depth} = 1 - H$ applies)~~(note that $H_{intensity} = 1 - H_{depth}$)~~ R2. Values from 0.3 to 0.9 have also been reported for some specific cases (e.g., Yu et al., 2004; Panthou et al., 2014, for scaling intervals defined within 1 h and 24 h).

Higher H values have been generally observed for shorter-duration intervals, and regions dominated by convective precipitation (e.g., Borga et al., 2005; Nhat et al., 2007; Ceresetti et al., 2010; Panthou et al., 2014, and references therein). Nonetheless, some studies performing spatio-temporal scaling analysis reached a different conclusion. For instance, Eggert et al. (2015), analyzing extreme precipitation events from radar data for durations between 5 min and 6 h and spatial scales between 1 km and 50 km, indirectly showed that stratiform precipitation intensity generally displays higher temporal scaling exponents than convective intensity. For short-duration intervals (typically less than one hour), previous studies have also reported more spatially homogeneous H estimates than for long-duration intervals (e.g., Alila, 2000; Borga et al., 2005, and references therein). This suggests that processes involved in the generation of local precipitation are comparable across different regions.

More generally, higher H values are associated with larger variations in moment values as the scale is changed (i.e. a stronger scaling), while H close to zero means that the X_d distributions for different durations d more closely match each other.

2.1 Simple Scaling GEV models

Annual Maximum Series (AMS) are widely used to select rainfall extremes from available precipitation series. Various theoretical arguments and experimental evidences support their use for extreme precipitation inference (e.g., Coles et al., 1999; Katz et al., 2002; Koutsoyiannis, 2004a; Papalexiou et al., 2013).

Based on the asymptotic results of the Extreme Value Theory (Coles, 2001), the AMS distribution of a random variable X is well described by the Generalized Extreme Value (GEV) distribution family. If we represent the AMS by $(x_1, x_2, ..., x_n)$, the GEV cdf can be written as (Coles, 2001):

$$F(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{5}$$

where $\xi \neq 0$, $-\infty < x \leq \mu + \sigma/\xi$ if $\xi < 0$ (bounded tail), and $1/\mu + \sigma\xi \leq x < +\infty$ if $\xi > 0$ (heavy tail). $\mu \in \mathbb{R}$, $\sigma > 0$ and ξ respectively represent the location, scale, and shape parameters of the distribution. The shape parameter describes the characteristics of the distribution tails ~~and the frequency of the extremes generated by $F(x)$~~[R1]. Thus, high order quantile estimation is particularly affected by the value of ξ. If $\xi = 0$ (light-tailed shape, Gumbel distribution), Eq. (5) reduces to:

$$F(x) = \exp\left\{-\exp-\left\{\frac{x-\mu}{\sigma}\right\}\right\} \tag{6}$$

where $-\infty < x < +\infty$.

In applications, the GEV distribution is frequently constrained by the assumption that $\xi = 0$ (i.e., to the Gumbel distribution), due to the difficulty of estimating significant values of the shape parameter when the recorded series are short (e.g., Borga et al., 2005; Overeem et al., 2008; CSA, 2012). However, based on theoretical and empirical evidence, many authors have shown that this assumption is too restrictive for extreme precipitation, and may lead to important underestimations of the extreme quantiles (e.g., Koutsoyiannis, 2004a, b; Overeem et al., 2008; Papalexiou et al., 2013, *Papalexiou and Koutsoyiannis*, 2013).

Instead, approaches aimed at increasing the sample size ~~series length~~[R1] may be used to improve the estimation of the GEV distribution shape parameter (for instance, the Regional Frequency Analysis (RFA), Hosking and Wallis, 1997). Among these approaches, SS models constitute an appealing way to pool data from different samples (durations) and reduce uncertainties in GEV parameters.

For the GEV distribution it is straightforward to verify that, if $X \overset{d}{=} GEV(\mu, \sigma, \xi)$ then $\lambda X \overset{d}{=} GEV(\lambda\mu, \lambda\sigma, \xi)$ for any $\lambda \in \mathbb{R}$. This means that the GEV family described by Eq. (5) and (6) satisfies Eq. (2) and thus complies with scale invariance for any constant multiplicative transformation of X. Under this assumption the wide sense SS definition [Eq. (3)] gives:

$$\mu_d = d^{-H}\mu_*, \;\; \sigma_d = d^{-H}\sigma_*, \quad \text{and} \;\; \xi_d = \xi_* \tag{7}$$

where μ_*, σ_*, and ξ_* represent the GEV parameters for a reference duration d^* chosen, for simplicity, as $d^* = 1$, so that $\lambda = d$.

2.2 SS GEV estimation

Taking advantage of the scale invariant formulation of the GEV distribution, many authors have proposed simple scaling IDF and DDF models for extreme precipitation series (e.g., Yu et al., 2004; Borga et al., 2005; Bougadis and Adamowski, 2006;

Bara et al., 2009; Ceresetti, 2011). In these cases, the scaling exponent and the GEV parameters are generally estimated in two separate steps: first, the H value is empirically determined through a log-log linear regression, as described above; then, GEV parameters μ_*, σ_*, and ξ_* for the reference duration d^* are estimated on the pooled sample of all available durations. Classical estimation procedures, such as GEV Maximum-Likelihood (ML) (Coles, 2001) or Probability Weighted Moment (PWM) (Greenwood et al., 1979; Hosking et al., 1985), can be used. In a few other cases, a Generalized Additive Model ML (GAM-ML) framework (Coles, 2001; Katz, 2013) has also been used to obtain the joint estimate of H, μ_*, σ_*, and ξ_* through the introduction of the duration as model covariate (e.g. Blanchet et al., 2016). R2

3 Data and study region

Four station datasets were used for the construction of intensity AMS at different durations: the Daily Maxima Precipitation Data[R2] (DMPD ~~DM~~[R2]) and the Hourly Canadian Precipitation Data[R2] (HCPD ~~H~~[R2]) datasets provided by Environment and Climate Change Canada (ECCC) and the MDDELCC [in french Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques] for Canada, and the Hourly Precipitation Data (HPD) and 15-Min Precipitation Data (15PD) datasets made available by the National Oceanic and Atmospheric Administration (NOAA) agency [http://www.ncdc.noaa.gov/data-access/land-based-station-data] for United States. The total number of stations was approximately 3400, with roughly 2200 locations having both DMPD ~~DM~~[R2] and HCPD ~~H~~[R2] series, or both HPD and 15PD series. The majority of stations are located in the United States and in the southern and most densely populated areas of Canada. In northern regions the station network is sparse and the record length does not generally exceed 15 or 20 years. Moreover, for most of DMPD~~DM~~[R2] and HCPD ~~H~~[R2] stations, the annual recording period does not cover the winter season and available series generally include precipitation measured from May to October. For this reason, the year was defined as the period from June to September for the stations located north of the 52^{nd} Parallel [122 days a year were used], while the period from May to October was used for remaining stations [184 days a year].

Data were collected through a variety of instruments [e.g., standard, tipping-bucket, and Fischer-Porter rain gauges] and precipitation values were processed and checked using both automated and manual methods (CSA, 2012, *HPD and 15PD online documentation*). Most often, observations were recorded by tipping-bucket gauges with tip resolution from 0.1 mm to 2.54 mm (CSA, 2012; Devine and Mekis, 2008). 15 min series usually present the coarser instrument[R1] resolution, with a minimum non-zero value of 2.54 mm, observed for about 80.5% of 15PD stations. The effects of such a coarse instrument[R1] resolution on simple scaling estimates could be important leading to empirical X_d cdfs becoming step-wise functions with a low number of steps. Some preliminary analyses aiming at evaluating these effects on SS estimates are presented in the supplementary material [see Fig. S2 and S3]. However, the 15PD dataset is important considering the associated network density and its fine temporal resolution, and thus it has been retained for our study. The main characteristics of the available datasets are summarized in Table 1.

The scaling AMS datasets were constructed according to the following steps:

(i) Three duration sets were defined: a) 15 min to 6 h with a 15min step; b) 1 h to 24 h with a 1h step; c) 6 h to 168 h (7

days) with a 6h step. These duration sets are hereinafter referred to as Short-Duration (SD), Intermediate-Duration (ID), and Long-Duration (LD) datasets, respectively.

(ii) Meteorological stations that were included in each final dataset were selected according to the following criteria: 1) precipitation series must have at least 85% of valid observations each year, otherwise the year was considered as missing; 2) each station must have at least 15 valid years; 3) for each station, it was possible to compute AMS for all durations considered in the scaling dataset (e.g., HCPD ~~H~~[R2] and HPD stations were not included in the SD dataset because only hourly durations were available). Note that, in order to exclude outliers possibly associated with recording or measurement errors, extremely large observations were discarded and assimilated to missing data. In particular, as in some previous studies (e.g., Papalexiou and Koutsoyiannis, 2013; Papalexiou et al., 2013), an iterative procedure was applied prior to step (ii)-1) to discard observations larger than 10 times the second largest value of the series. [R2]

(iii) A moving window was applied to 15PD, HCPD ~~H~~[R2], and HPD series to estimate aggregated series at each duration. For DMPD~~DM~~[R2] series, a quality check was also implemented in order to guarantee that precipitation intensities recorded each day at different durations were consistent with each other . For ~~(for~~ finstance, each pair of DMPD~~DM~~[R2] intensity (x_{d_1}, x_{d_2}) observed at durations $d_1 < d_2$ must respect the conditions $x_{d_2}/x_{d_1} \leq 1$ and $d_1 x_{d_1} \leq d_2 x_{d_2}$ derived from the definitions of daily maxima rainfall intensity and depth; otherwise all DMPD values recorded that day were discarded and assimilated to missing data. ~~the condition $d_1/d_2 \leq x_{d_2}/x_{d_1} \leq 1$, otherwise the day was considered as missing)~~[R1].

(iv) For each selected station, annual maxima were extracted for each valid year and duration. For stations having both DMPD and HCPD series, or 15PD and HPD series, for each year, the annual maxima extracted from these two series were compared and the maximum value was retained as the annual maximum for that year. ~~Then, for stations having both DM and H series, or 15PD and HPD series, over a common time period, annual maximum values between these two series were retained.~~[R2]

Major characteristics of each scaling AMS dataset are reported in Table 2.

4 SS estimation through Moment Scaling Analysis (MSA)

Moment Scaling Analysis (MSA) for the SD, ID, and LD datasets was carried out to empirically validate the use of SS models for intensity AMS. Assessing the validity of the SS hypothesis for various duration intervals also aimed at determining the presence of different scaling regimes for precipitation intensity distributions. ~~The spatial distribution of the scaling exponents was then analyzed to assess the dependence of SS on local geoclimatic characteristics.~~

In order to identify possible changes in the SS properties of AMS distributions, various scaling intervals were defined for the MSA. In particular, all ~~All~~[R1] possible subsets with 6, 12, 18 and 24 contiguous durations were considered within ~~for~~[R1] each dataset. Figure 1 and Figure 2 show the 136 scaling intervals thereby defined: 40 scaling intervals for SD and IS, and 56 scaling intervals for LD. For instance, the first matrix on the left of Fig. 1(a) presents the 6-duration scaling intervals 15min - 1.5h, 30min - 1.75h, . . . , 4.75h - 6h defined for ~~from~~ the SD dataset [i.e. the 19 scaling intervals containing six contiguous durations defined with a 15min increment]~~(with 15min step)~~[R1] , while Fig. 1(d) shows an example of the first four 6-duration scaling

intervals for the ID dataset [i.e. 1h-6h, 2h-7h, 3h-8h, and 4h-9h, containing six contiguous durations defined with an increment of 1h][R1]. This procedure was defined in order to evaluate the sensitivity of the SS estimates to changes in the first duration d_1 of the scaling interval and in the interval length [i.e. the number of durations included in the scaling interval] ~~variability of the SS estimates when changing the position of the scaling interval [hereinafter identified by its first duration d_1; see examples in Figure 1(d)] and the number of durations considered~~[R1].

For each scaling interval (for simplicity, their index has been omitted), the validity of the SS hypothesis was verified according to the following steps:

1. *MSA regression:* for $q = 0.2, 0.4, \ldots, 2.8, 3$, the slopes K_q of the log-log linear relationships between the empirical $q-$moments $\langle X_d^q \rangle$ of $X_{d_1}, X_{d_2}, \ldots, X_{d_D}$ and the corresponding durations $d_1, d_2, \ldots, d_D$ in the scaling interval $[d_1, d_D]$[R1] were estimated by Ordinary Least Squares (OLS). Order $q \geq 3$ were not considered because of the possible biases affecting empirical high order moment estimates.

2. *Slope test:* Regressing the MSA slopes[R1] K_q on q, the hypothesis that the estimated K_q-exponents vary linearly with the order of moment q was verified. To this end, a Student's t-test ~~t-Student test~~[R1] was used to test the null hypothesis $\boldsymbol{H}_0$: $\beta_1 = K_1$, where β_1 is the slope coefficient of the simple regression model $K_q = \beta_0 + \beta_1 q$. If $\boldsymbol{H}_0$ was not rejected at the significance level $\alpha = 0.05$, the SS assumption ~~for the scaling interval~~[R1] was considered appropriate for the scaling interval[R1] and the simple scaling exponent $H = K_1$ was retained.

3. *Goodness-of-Fit (GOF) test:* for each duration d, the goodness of fit of the X_d distribution under SS was tested using the Anderson-Darling (AD) and the Kolmogorov-Smirnov (KS) tests. These tests aim at validating the appropriateness of the scale invariance property for approximating the X_d cdf by the distribution of $X_{d,ss} = d^{-H} X_{d^*}$. To this end, the pooled sample[R1]

$$\boldsymbol{x}_{d^*} = \left(\boldsymbol{x'}_{d_1}, \ldots, \boldsymbol{x'}_{d_j}, \ldots, \boldsymbol{x'}_{d_D}\right) \tag{8}$$

of the D rescaled AMS, $\boldsymbol{x'}_{d_j}$, was used to define X_{d^*} under the SS assumption, considering all the durations d_j, with $j = 1, \ldots, D$, in the scaling interval. Each rescaled sample $\boldsymbol{x'}_{d_j}$ of the annual maxima $x_{d_j,i}$, $i = 1, \ldots, n$, observed for the duration d_j was obtained by simply inverting Eq. (2) for d^* : [R1]

$$\boldsymbol{x'}_{d_j} = \left({d_j}^H x_{d_j,1}, {d_j}^H x_{d_j,2}, \ldots, {d_j}^H x_{d_j,i}, \ldots {d_j}^H x_{d_j,n}\right) \tag{9}$$

where n represents the number of observations (years) available for each duration. In this way, $n \times D$ rescaled observations were included in $\boldsymbol{x}_{d^*}$.

As in previous applications (e.g., Panthou et al., 2014), the AD and KS tests were then applied at significance level $\alpha = 0.05$ to compare the empirical distributions (Cunnane plotting formula, Cunnane, 1973) of the SS sample, $\boldsymbol{x}_{d,ss} = d^{-H} \boldsymbol{x}_{d^*}$, and the non-SS sample, $\boldsymbol{x}_d$.

~~for each duration d, the goodness of fit of the X_d distribution under SS was tested using the Anderson-Darling (AD) and the Kolmogorov-Smirnov (KS) tests. These tests aim at validating the appropriateness of the scale invariance property for approximating the X_d cdf by the distribution of $\lambda^{-H} X_{d^*}$. To this end, the pooled sample~~

~~$\boldsymbol{x}_{d^*} = (\boldsymbol{x'}_{d_1}, \ldots, \boldsymbol{x'}_{d_j}, \ldots, \boldsymbol{x'}_{d_D})$ (8)~~

~~of the rescaled AMS, $\boldsymbol{x'}_{d_j}$, for all durations d_j, $j = 1, \ldots, D$, in the scaling interval, was used to define X_{d^*} under the SS assumption. Using $d^* = 1$ h, the rescaled sample $\boldsymbol{x'}_{d_j}$ of the annual maxima $x_{d_j,i}$, $i = 1, \ldots, n$, observed for d_j was:~~

~~$\boldsymbol{x'}_{d_j} = \left(d_j{}^H x_{d_j,1}, d_j{}^H x_{d_j,2}, \ldots, d_j{}^H x_{d_j,i}, \ldots d_j{}^H x_{d_j,n}\right)$ (9)~~

~~where n represents the number of observations (years) available for each duration. Hence, $n \times D$ rescaled observations were included in $\boldsymbol{x}_{d^*}$.~~

~~The tests were then applied at significance level $\alpha = 0.05$ to compare $\boldsymbol{x}_d$ and $\boldsymbol{x}_{d^*}$ empirical distributions (Cunnane plotting formula, *Cunnane, 1973*) as in previous precipitation scaling applications (e.g., *Panthou et al., 2014*).~~ R1 In fact, despite the low power of KS and AD tests for small sample tests, they represent the only suitable solution to the problem of comparing empirical cdfs when the data do not follow a normal distribution. Because both AD and KS are affected by the presence of ties in the samples (e.g., repeated values due to rounding or instrument resolution), a permutation test approach (Good, 2013) was used to estimate test p-values. According to this approach, data in $\boldsymbol{x}_d$ and $\boldsymbol{x}_{d,ss}$ were pooled and randomly reassigned to two samples having same sizes as ~~pooled data of $\boldsymbol{x}_d$ and $\boldsymbol{x}_{d^*}$ were randomly reassigned~~ R2 the SS and non-SS samples. Then, the test statistic distribution under the null hypothesis of equality of the $X_{d,ss}$ ~~$d^{-H} X_{d^*}$~~ R1 and X_d distributions was approximated by computing its value over a large set of random samples. Finally, the test p-value was obtained as the proportion of random samples presenting a test statistic value larger than the value observed for the original sample.

The mean error resulting from approximating the X_d distribution by the SS model was then evaluated in a cross-validations setting. For this analysis, each duration was iteratively excluded from each scaling interval and the scaling model re-estimated at each station. Predictive ability indices, such as the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) between empirical and SS distribution quantiles, were estimated for highest quantiles. In particular, to focus on return periods of practical interest for IDF estimation, only quantiles larger than the median were considered (i.e., only return periods greater than 2 years).

The average over all stations of the normalized RMSE, $\overline{\epsilon}_{x_d}$ ~~$\overline{r}_{x_d}$~~ R1, for each scaling interval and duration was used:

$$\overline{\overline{\epsilon}}_{x_d} = \frac{1}{n_s} \sum_{s=1}^{n_S} \overline{\epsilon}_{x_{d,s}} \tag{10}$$

~~$\overline{\overline{r}}_{x_d} = \frac{1}{n_s} \sum_{s=1}^{n_S} \overline{r}_{x_{d,s}}$ (10)~~ R1

where n_s is the number of valid SS stations in the dataset ~~, and~~

$$\overline{\epsilon}_{x_{d,s}} = \frac{\epsilon_{x_{d,s}}}{\overline{x}_{d,s}} \tag{11}$$

$\overline{r}_{x_{d,s}} = \frac{r_{x_{d,s}}}{\overline{x}_{d,s}}$ (11) R1

and $\epsilon_{x_{d,s}}$ ~~$r_{x_{d,s}}$~~ R1 and $\overline{x}_{d,s}$ are, respectively, the RMSE and the mean value of all X_d quantiles of order $p > 0.5$ at station s. Note that the normalized RMSE is a measure of error, meaning that values of $\overline{\epsilon}_{x_{d,s}}$ closer to 0 correspond to a better fit than larger values. R1

4.1 Model estimation and validation

Figure 1 presents the results of points 1 to 3 of the procedure for the evaluation of the SS validity. It shows, for each scaling interval and duration, the proportion of valid SS stations [Fig. 1(a)-(c)]. As showed in the example in Fig. 1(e), for each scaling interval, valid SS stations were defined as stations having not rejected both the Slope test for the scaling interval and the GOF tests for each duration included in this scaling interval.

As expected, the proportion of valid SS stations decreased when the number of durations within the scaling interval increased and with decreasing d_1. This is particularly evident for short d in SD and ID datasets. More GOF test rejections were observed for longer scaling intervals [not shown], due to the higher probability of observing large differences between $\boldsymbol{x}_d$ and $\boldsymbol{x}_{d,ss}$ ~~$\boldsymbol{x}_{d^*}$~~ quantiles when $\boldsymbol{x}_{d,ss}$ had larger sample size ~~$\boldsymbol{x}_{d^*}$ was larger~~ and included data from more distant durations. However, several factors can impact GOF test results when shorter d_1 are considered ~~included in the scaling intervals~~. First, the SS hypothesis could be rejected due to the presence of very large values in short-duration samples, to which GOF tests are particularly sensitive. Second, when considering durations close to the temporal resolution of the recorded series [i.e., 15 min in SD and 1 h in ID and LD], stronger underestimations could affect the measure of precipitation because intense rainfall events are more likely to be split between two consecutive time steps. Finally, preliminary analyses [Fig. S2 and S3 in the supplementary material] showed that the largest GOF test rejections could also be connected to the coarse instrument resolution of 15PD series, which, similar to the temporal resolution effect, induces larger measurement errors in the shortest duration ~~precipitation~~ series. Note that comparable resolution issues were previously reported by some authors while estimating fractal and intermittency properties of rainfall processes (e.g., Veneziano and Iacobellis, 2002; Mascaro et al., 2013) and IDF (e.g., Blanchet et al., 2016) R2 .

Valid SS station proportions between 0.99 and 1 were always observed for GOF tests in ID and LD datasets, except for some durations shorter than 3 h (ID dataset) or 6 h (LD dataset). For all three datasets, no particular pattern was observed for slope test results [not shown], with at most 2% of the stations within each scaling interval displaying a non linear evolution of the scaling exponent with the moment order.

When considering both GOF and Slope test, with the exception of some durations $\leq$1 hour, the proportion of stations satisfying SS was higher than 0.9, and the majority of scaling intervals [65%, 90%, and 98% of the scaling intervals in SD, ID, and LD, respectively] included at least 95% of valid SS stations. For each scaling interval, only valid SS stations were considered in the rest of the analysis.

Figure 2 presents, for each scaling interval and duration, the station average, $\overline{\epsilon}_{x_d}$ ~~$\overline{r}_{x_d}$~~ R1, of the normalized RMSE. These graphics show that mean relative errors on intensity quantiles did not generally exceed 5% of the precipitation estimates for 6-

duration scaling intervals [Fig. 2, first col.]. Larger errors were observed for durations at the border of the scaling intervals. Not surprisingly, this result underlines that, in a cross-validation setting, both the MSA estimation of H and the $X_{d,ss}$ approximation are less sensitive to the exclusion of an inner duration of the scaling interval than to the exclusion of d_1 or d_D. Conversely, the extrapolation under SS of the X_d distribution is generally less accurate if d is outside the range of durations used to estimate H. [R1] Moreover, as ~~As~~ for the valid SS station proportion, the performances of the model deteriorated with decreasing d_1 and with increasing scaling interval length ~~, especially for durations at the border of the scaling intervals.~~ However, for more that 70% of 12-, 18-, and 24-duration scaling intervals, $\bar{\bar{\epsilon}}_{x_d}$ ~~$\bar{\bar{r}}_{x_d}$~~[R1] ≤ 0.1 for each duration included in the scaling interval. $\bar{\bar{\epsilon}}_{x_d}$ ~~$\bar{\bar{r}}_{x_d}$~~[R1] ≥ 0.25 were observed for 15 min in 12-duration or longer scaling intervals, pointing out the weaknesses of the model in approximating short duration extremes when the scaling interval included durations ≥ 3 h.

4.2 Estimated scaling exponents and their variability

In order to evaluate the sensitivity of SS to the considered scaling interval, the variability of H with d_1 has been analyzed. Then, the spatial distribution of the scaling exponents for each scaling interval was studied to assess the uncertainty in H estimation and the dependence of SS exponents on local geoclimatic characteristics.[R1]

Investigating the variability of the scaling exponent with the scaling interval is particularly important since, if SS is assumed to be valid between some range of durations, one should expect that H remains almost unchanged over the various scaling intervals included in this range. For this reason, the variation $\Delta_{H_{(j)}}$ of the scaling exponents computed for overlapping scaling intervals having the same d_1 but different lengths was analyzed. [R1]

~~The median, Interquantile Range (IR), and quantiles of order 0.1 and 0.9 of the H distribution across stations, are presented in Fig. 3(i) for each 6-duration scaling interval. Figures 3(ii) - (iv) show the distribution of the scaling exponent variation $\Delta_{H_{(j)}}$ observed over stations when each scaling interval is lengthened from 6 to 12, 18, and 24 durations.~~

For each station and each d_1, $\Delta_{H_{(j)}}$ was defined as:

$$\Delta_{H_{(j)}} = H_{(j)} - H_{(6)} \tag{12}$$

where $j = 12, 18$, or 24 represent the number of durations considered in the specified scaling interval, $H_{(j)}$ is the corresponding scaling exponent, and $H_{(6)}$ is the scaling exponent estimated for the corresponding 6-duration scaling interval (i.e., the 6-duration interval having the same d_1). If SS is appropriate over a range of durations, $\Delta_{H_{(j)}}$ is expected to be small for scaling intervals defined within this range.[R1]

The median, Interquantile Range (IQR), and quantiles of order 0.1 and 0.9 of the H distribution across stations, are presented in Fig. 3(i) for each 6-duration scaling interval. Figures 3(ii)-(iv) show the distribution over valid SS stations of $\Delta_{H_{(j)}}$ for all relevant scaling intervals. ~~Figures 3(ii)-(iv) represent the changes observed in H values when the scaling interval length and d_1 increased~~. Median $\Delta_{H_{(j)}}$, as well as its IQR ~~IR~~[R1], increased with the number of durations added to the scaling interval for all d_1. For the 24-duration scaling interval "1h - 24h" (ID dataset), for instance, median $\Delta_{H_{(24)}} = 0.047$ was observed. For the interval "15min - 6h" (SD dataset), $\Delta_{H_{(24)}}$ was even larger, with a median scaling exponent variation approximately equal to 0.087 and with 25% of stations having $\Delta_{H_{(24)}} \geq 0.11$. These results indicate that, for some stations, a dramatic difference

could exist in IDF estimations obtained with the different definitions of the scaling interval. Changes in H values were also important when comparing 6- and 12-duration scaling intervals when $d_1 \leq 1$ h (SD and ID datasets) and in LD dataset [Fig. 3 (ii)].

Nonetheless the median scaling exponent variation was generally smaller than 0.05, except for a relatively small proportion of stations. Equally important, $|\Delta_{H_{(j)}}|$ was generally centered on 0 and for all $d_1 \geq 1$ h more than 50% of stations had $|\Delta_{H_{(12)}}| \leq 0.025$ (SD dataset) and $|\Delta_{H_{(18)}}| \leq 0.03$ (ID dataset) [Fig. 3 (ii)-(iii)].

The smallest median H values were observed for the shortest d_1 ($d_1 \leq 30$ min) in Fig. 3 (a-i), and for the longest d_1s in Fig. 3 (c-i). Scaling intervals beginning at 15 and 30 min also displayed the smallest variability across stations. Although fewer stations were available for these intervals (only 15PD stations were used and the number of valid SS stations was smaller), this result is consistent with previous reports in the literature demonstrating that H values are spatially more homogeneous for short durations.

A larger dispersion of H values was observed when d_1 ranged between approximately 1 h and 5 h, in particular in the SD dataset, for which the 10^{th}-90^{th} percentile difference almost covered the entire range of observed H values [Fig. 3 (i)]. This result could be in part explained by the fact that, if the scaling interval length is fixed, then the variance $V[\ln(d)]$ of the MSA regression covariate decreases as d_1 increases. In fact, the use of a logarithmic scale for the MSA regression implies that the mean distance between durations in the scaling interval decreases as d_1 increases Thus, regression errors of the same magnitude in short and long d_1 scaling intervals differently affect the OLS variance of H, especially when scaling intervals are short. This may result in larger uncertainty of H for longer d_1 scaling intervals of SD. Moreover, as showed in next sections, H variability across stations may be effectively larger due to the greater spatial variability of the scaling exponent for d_1 longer than a few hours.

Largest median H were observed for d_1 greater than 10 hours [Fig. 3 (b-i)] and lower than 2 days [Fig. 3 (c-i)], with approximately half of the stations having $H \geq 0.8$. This means that a stronger scaling (i.e., larger H values) is needed to relate extreme precipitation distributions at approximately 12-hours to distributions at daily and longer scales. It may therefore be expected that the stations characterized by H closer to 1 are located in geographical areas where differences in precipitation distributions are important among temporal scales included in these scaling intervals.

Examples of the spatial distributions of the scaling exponent are given in Fig. 4 and 5 for the first and last d_1 for each interval length and dataset, respectively. Since only one 24-duration scaling interval was defined for both the SD and ID datasets, only scaling intervals containing 6, 12, and 24 (Fig. 4) or 18 (Fig. 5) durations are presented. This avoids the redundancy of showing twice the "15min - 6h" (SD dataset) and "1h - 24h" (ID dataset) scaling intervals.

Generally, the scaling exponent displayed a strong spatial coherence and varied smoothly in space, although a more scattered distribution of H characterizes maps in Fig. 5. In this last figure, the local variability of H may be attributed to the larger estimation uncertainties, as previously mentioned. Meaningful spatial variability and clear spatial patterns emerged for $d_1 \geq 1$ h. In fact, for stations located in the interior and southern areas of the continent, a shift from weaker scaling regimes (smaller H) to higher H values was observed as d_1 increases [e.g., second and third rows of Fig. 4]. On the contrary, a smoother evolution

of H over the scaling intervals characterized the northern coastal areas, especially in north-western regions, and the Rockies, where $H > 0.75$ values were rarely observed even for greater d_1 values.

5 Regional analysis

Regional differences in scaling exponents were investigated. Only the results for the 6-duration scaling intervals are presented, similar results having been obtained for longer scaling intervals [see the supplementary material, Fig. S5 and S6 for 12- and 18-duration scaling intervals]. Stations were pooled into six climatic regions based on a previous classification suggested by Bukovsky (2012) [see Fig. 6]. Stations outside the domain covered by the Bukovsky regions were attributed to the nearest region. Regions with less than 10 stations were not considered (regions without colored borders in Fig. 6) and region A1 (W_Tun) was kept separated from region A2 (NW_Pac) because only 14 stations were available in region A1 (W_Tun) for ID and LD datasets.

To provide deeper insights about regional features of precipitation associated with specific scaling regimes two variables related to the precipitation events observed within AMS were also analyzed: the mean number of events per year, $\bar{N}_{eve}$, and the mean wet time per event, $\bar{T}_{wet}$, contributing to AMS within each scaling interval. For a given year and station, annual maxima associated to different durations of a given scaling interval were considered to belong to the same precipitation event if the time intervals over which they occurred overlapped. ~~[see Fig. 7 (g); in this example 3, 4, and 5 h annual maxima are associated with the first event while 1, 2, and 6 h annual maxima are associated to the second event]. The mean number of events at each station was then computed:~~

~~$$\bar{N}_{eve} = \frac{1}{n}\sum_{i}^{n} N_{eve,i} \quad (13)$$~~

~~with $N_{eve,i}$ the number of non-overlapping time intervals, i.e. the number of different events contributing to AMS during the i^{th} year of record. The distribution of $\bar{N}_{eve}$ values within each region is presented in Fig. 7.~~

[R1] The mean wet time per event contributing to AMS, $\bar{T}_{wet}$, was defined as the mean number of hours with non-zero precipitation within each event. Details on the calculation of $\bar{N}_{eve}$, [R1] $\bar{T}_{wet}$, and the corresponding results are presented in the supplementary material [Sect. S2 and Fig. S4 and S5[R1]].

5.1 Regional variation of the scaling exponents.

Figure 7 shows the distribution of H within each region. Three types of curves can be identified. First, curves in Fig. 7 (a) to (c) have a characteristic smooth S shape. Conversely, Fig. 7 (d) displays a rapid increase of H for scaling intervals defined in ID and LD datasets until $d_1 = 2$ days, preceded and followed by two plateaus, one for the longest d_1 with remarkably high H values, and one for the shortest d_1 with small H values. Finally, an inverse-U-shaped curve can be seen in Fig. 7 (e) and (f), with globally high H values already reached at sub-daily durations in dry regions (E).

The difference between Fig. 7 (a) and (e)-(f) can be partially explained by the weaker impact of convection processes in gen-

erating very short duration extremes in regions A1 and A2 with respect to southern areas (regions E and F). For northern regions, the transition between short and long duration precipitation regimes may be smoothed by cold temperatures which moderate short-duration convective activity, especially for W_Tun (region A1). The topography characterizing the northern pacific coast may explain the smoothing effect for the curve of region NW_Pac (A2): precipitation rates at daily and longer scales are enhanced by the orographic effect acting on synoptic weather systems coming from the Pacific Ocean (Wallis et al., 2007).

Similarly, mountainous regions in C [Fig. 7 (c)] displayed the smallest variations of H over d_1, indicating that analogous scaling regimes characterize both short- and long-duration scaling intervals. Again, this may be related to the important orographic effects of precipitation in these regions that are involved in the generation of extremes for both sub-daily and multi-daily time scales. The mean number of events per year in regions A and C was higher than in regions E-F, in particular for SD scaling intervals, and displayed steeper decreases with increasing d_1 [Fig.S4 ~~7~~[R1] (a) and (c) in the supplementary material[R1]].

Main differences between regions B and A were the stronger scaling regimes observed in B, which were mainly due to contributions from stations located in the south-eastern part of the E_Bor region (not shown). For scaling intervals in the ID dataset, region B was also characterized by the highest mean number of events per year, with most of the stations presenting $\bar{N}_{eve} > 2$ for $d_1 = 1$ h and $d_1 = 2$ h and sharp decreases of $\bar{N}_{eve}$ with increasing d_1 [Fig. S4 ~~7~~[R1] (b) in the supplementary material[R1]]. Moreover, a remarkably large range of $\bar{N}_{eve}$ was observed for 1 h $\leq d_1 \leq$ 6 h, suggesting that B may be highly heterogeneous.

Two distinct scaling regimes can be observed for SW_Pac (region D) at, respectively, $d_1 \leq 3$ h (SD dataset) and $d_1 \geq 2$ days (ID dataset) [region D in Fig. 7 (d)]. These plateaus may be interpreted by recalling that $H = 1 - H_{depth}$ ~~$H_{intensity} = 1 - H_{depth}$~~[R2]. On the one hand, the low and constant H observed for $d_1 \leq 3$ h indicates that the average precipitation depth increases with duration at the same growth rate for all these intervals. On the other hand, H approximately equal to 0.9 at daily and longer durations demonstrates that the average precipitation depth associated with long-duration annual maxima remained roughly unchanged when the duration increased from 1.5 to 7 days ($\lambda^{H_{depth}} \approx 1$ in Eq. (3)). This, along with the fact that the scaling exponent increased almost monotonically for 1 h $\leq d_1 \leq$ 24 h (ID and LD datasets), suggests that extremes at durations shorter than $\sim$ 3 h (SD dataset) drive annual maxima precipitation rates at longer scales, with the rapid and continuous decay in mean intensity caused by the increasing size of the temporal scale of observation.

For SW_Pac (region D), the relative absence of long-lasting weather systems able to produce important extremes for long durations, was confirmed by the analysis of $\bar{N}_{eve}$ and $\bar{T}_{wet}$ [see Fig. S4 and S5 of the supplementary material] ~~[for results on $\bar{T}_{wet}$ see Fig. S4 of the supplementary material]~~[R1]. In fact, the mean number of events per year was relatively high for short durations (the median $\bar{N}_{eve}$ is equal to 1.82 for $d_1 = 15$ min and to 1.4 for $d_1 = 1$h), while it rapidly decreased below 1.1 events per year for $d_1 \geq 6$ h (ID dataset) and for $d_1 \geq 18$ h (LD dataset). With the exception of $d_1 = 6$ h (LD dataset), at least 90% of SW_Pac stations had $\bar{N}_{eve} \leq 1.25$ for all $d_1 > 3$ h. In other regions, median $\bar{N}_{eve}$ were never smaller than 1.1 for the SD and ID datasets, except for $d_1 \geq 12$h in region E.

These results suggests that both the distinctive topography of the west coast and the characteristic large-scale circulation of the south-west areas of the continent are crucial factors determining the transition between the two scaling regimes in region D. [R1]

Median H values displayed inverse-U shapes for the remaining regions with very small IQR ~~IR~~[R1], despite the high number of valid SS stations: a slow transition from lower to higher H is observed approximately between 1 h and 12 h (region E) or 30 h (region F). The strongest scaling regimes were observed for 1 h $\leq d_1 \leq$ 2 days in arid western regions [Fig. 7 (e)], while median H values greater than 0.8 were only observed for approximately 6 h $\leq d_1 \leq$ 2 days in more humid areas [7 (f)]. In both region E and F, very short-duration extremes are typically driven by convective processes, while a transition to different precipitation regimes may be expected between 1 h and a few hours. However, the smoother increase of H visible in Fig. 7 (f) with respect to (e) may also indicate that, in eastern areas, the occurrence of sub-daily duration extremes are more likely associated to embedded convective and stratiform systems, or to mesoscale convective systems less active in western dry areas (Kunkel et al., 2012). On the contrary, for south-western dry regions [Fig. 7 (e)], where less intense summer extremes are expected compared to eastern areas [see supplementary material, Fig. S1], differences between short- and long-duration extreme precipitation intensity seem stronger since H tended to scatter in a range of higher values: precipitation intensity moments strongly decrease as the duration increases for approximately 1 h $\leq d_1 \leq$ 12 h.

In summary, these results suggest a regional effect on precipitation scaling of both local geographical characteristics, such as topography or coastal effects, and general circulation patterns. In general, the weakest ~~Weak~~[R1] scaling regimes were observed for short d_1 and along the west coast of the continent and seem to be connected to scaling intervals and climatic areas characterized by homogeneous weather processes. Low H values correspond in fact to small variations in AMS distribution moments. On the contrary, stronger scaling regimes, which indicate important changes occurring in AMS moments across duration and, thus, in extreme precipitation features, were observed for longer d_1 in the other regions of the study area. According to these results, it would be important to take into account the climatological information included in the scaling exponent to improve SS and IDF estimation. Even more important, these results could help for the definition of IDF relationships at non-sampled locations by the construction of spatial models for the IDF parameter H.

6 SS GEV etimation

Results presented in this section are limited to a descriptive analysis of GEV parameter estimates, and to an assessment of the potential improvements carried out by SS GEV models with respect to PWM estimates of non-SS GEV models, for 6-duration scaling intervals. Similar results were generally obtained 12-, 18-, and 24-duration intervals [see supplementary material, Fig. S9 to S15].

In our study, the PWM procedure was applied to estimate SS-GEV parameters μ_*, σ_*, and ξ_* [Eq. (7)] from $\boldsymbol{x}_{d^*}$ [Eq. (8)]. For each duration d, PWM were also used to estimate non-SS parameters μ_d, σ_d, and ξ_d from each of the non-SS samples $\boldsymbol{x}_d$. ~~In our application, GEV parameters μ_*, σ_*, and ξ_* [Eq. (7)] were estimated for $\boldsymbol{x}_{d^*}$ [Eq. (8)] using the PWM procedure.~~[R2]

Preliminary comparisons ~~of among~~ several estimation methods [~~e.g.,~~ PWM, classical ML estimators, and GAM-ML; see Sect. 2.2 ~~Generalized Additive Model ML - see Coles (2001), Katz (2013)- in which the joint estimation of H, μ_*, σ_*, and ξ_* is obtained by the introduction of the duration as model covariate~~[R2]], showed that PWM slightly outperformed the other methods.

Quantiles estimated from the SS and the non-SS GEV were compared with empirical quantiles. Global performance measures, such as RMSE, were computed to evaluate the overall fit of the estimated GEV to the empirical X_d distributions. In particular, mean errors between SS and non-SS quantile estimates and empirical quantiles were compared using the relative total RMSE ratio, $R_{\overline{rmse}}$, defined as:

$$R_{\overline{rmse}} = \frac{[\overline{R}_{ss} - \overline{R}_{non-ss}]}{\overline{R}_{non-ss}} \tag{13}$$

where

$$\overline{R}_{mod} = \sum_{d=d_1}^{D} \frac{\epsilon_{d,mod}}{\bar{x}_d} \tag{14}$$

~~$\overline{R}_{mod} = \sum_{d=d_1}^{D} \frac{r_{d,mod}}{\bar{x}_d}$ (14)~~[R1]

represents the normalized mean square difference between model and empirical quantiles of order $p > 0.5$ for all the durations included in the scaling interval.

6.1 Estimated SS GEV parameters

Figure 8 presents the distributions over valid SS stations of the SS GEV parameters rescaled ~~scaled~~[R1] at $d_* = 1$ h [Fig. 8 (a) and (b)] and $d_* = 24$ h [Fig. 8 (c)]. For the SD dataset, even for scaling intervals which did not include the reference duration d^*, the μ_* and σ_* distributions appeared to be similar to the non-SS μ_d and σ_d distributions [Figure 8, first row]. Conversely, in the ID and LD datasets, both μ_* and σ_* distributions were more positively skewed than the corresponding non-SS distributions. Moreover, the relative differences $\Delta_\mu = (\mu_* - \mu_d)/\mu_d$ and $\Delta_\sigma = (\sigma_* - \sigma_d)/\sigma_d$ were estimated for each station, duration, and scaling interval. Two important results came out of this analysis [see Figures S10 and S11 of the supplementary material]. On the one hand, median values of Δ_μ and Δ_σ were generally smaller than $\pm 5\%$ and $\pm 10\%$, respectively. On the other hand, Δ_σ showed large positive values when $\xi_d = 0$ (i.e. Gumbel distributions), while small $\Delta_\sigma < 0$ were estimated when $\xi_d \neq 0$ [not shown for conciseness]. These results are interesting since, while non-SS μ_d values are generally considered to be accurate estimates of the X_d location parameter, small uncertainties are expected for the scale parameter only when the ξ_d value is correctly assessed. In fact, the scale parameter σ_d may be strongly biased when the shape parameter is spuriously set to zero ($\xi_d = 0$).

In addition, μ_* and σ_* displayed strong coherence in their spatial distributions, which were characterized by an obvious North-West to South-East gradient [Fig. 9 shows examples for the scaling intervals 15min - 1.5h, 1h - 6h, and 6h - 36h].

Notable differences between SS GEV and non-SS estimates were observed for the shape parameter [third column of Fig. 8]. Firstly, SS ξ_* were closer to 0 than non-SS ξ_d, for both positive and negative shape values. Secondly, ξ_* distributions were generally more peaked around their median value than non-SS estimates.

Note that, the majority of stations had non-SS shape parameters ξ_d non-significantly different from zero according to asymptotic test proposed by Hosking et al. [1985] for PWM GEV estimators applied at level 0.05. In particular, for each duration,

non-SS models estimated light-tailed distributions (i.e., $\xi_d = 0$) for more than 85% of the stations, except that for $d = 15$ min and $d = 30$ min [Fig. 10, first col.]. Conversely, for all scaling intervals with $d_1 > 15$ min, SS GEV shape parameters were significantly different from zero for 40% to 45% of valid SS stations [Fig. 10, second col.]. Moreover, when using scaling intervals of 12 durations or more, the proportion of $\xi_* > 0$ was always important (greater than 35%) for all 18- and 24-duration scaling intervals [see the supplementary material, Fig. S9].

The previous results suggest that pooling data from several durations may effectively reduce the sampling effects impacting the estimation of ξ, allowing more evidence of non-zero shape parameters, and, in many cases, of heavy tailed ($\xi > 0$) AMS distributions. This conclusion is consistent with previous reports, namely that 100- to 150-year series are necessary to unambiguously assess the heavy-tailed character of precipitation distributions (e.g., Koutsoyiannis, 2004b; Ceresetti et al., 2010). In general, typical values of $\xi \approx 0.15$, close to the estimated ξ_* for cases in which $\xi_* > 0$, have also been reported (e.g., Koutsoyiannis, 2004b).

However, uncertainties on ξ_* estimates remain important. Support for this comes from the spatial distribution of ξ_*, which was still highly heterogeneous, with local variability dominating at small scales [e.g., Fig. 9, third col.].

6.2 Improvement with respect to Non-SS models

The proportion of series for which the SS model RMSE, $\epsilon_{d,ss}$ ~~$r_{d,ss}$~~[R1], was smaller than the non-SS GEV RMSE, $\epsilon_{d,non-ss}$ ~~$r_{d,non-ss}$~~[R1], was analyzed [see the supplementary material, Fig. S12]. For cases with non-zero ξ_*, the fraction of stations with $\epsilon_{d,ss} < \epsilon_{d,non-ss}$ ~~$r_{d,ss} < r_{d,nonss}$~~[R1] was higher than 60% for most of the scaling intervals and durations. On the contrary, $\epsilon_{d,ss} > \epsilon_{d,non-ss}$ ~~$r_{d,ss} > r_{d,nonss}$~~[R1] was observed for the majority of stations (generally more than 70%) with $\xi_* = 0$.

Figure 11 presents the $R_{\overline{rmse}}$ distribution over valid SS stations. When the SS distribution shape parameters were not significantly different from zero [Fig. 11, second col.], the relative increases in total RMSE were usually smaller than 0.1 in SD dataset, with only scaling intervals with $d_1 < 1$ h having greater $R_{\overline{rmse}}$. For the ID and LD datasets, the medians of the total relative RMSE ratio distributions were smaller than 0.05 for $d_1 \geq 4$ h and $d_1 \geq 24$ h, respectively. Furthermore, more than 90% of stations had $R_{\overline{rmse}} < 0.125$ for $d_1 \geq 6$ h (ID dataset) and $d_1 \geq 30$ h (LD dataset). When $\xi_* \neq 0$, an increase of the mean error in high order quantile estimates was observed for $d_1 = 15$ min (SD dataset) and $d_1 = 1$ h (ID dataset) for at least half of the stations [Fig. 11, first col.; note the different scale on the y-axis]. However, for all other d_1, negative $R_{\overline{rmse}}$ values were observed for the majority of stations for all scaling intervals, with a median reduction up to 30% of the mean error. Note that also for 12- and 18-duration scaling intervals the median $R_{\overline{rmse}}$ where generally negative for $d_1 > 1$ h and $\xi_* \neq 0$ [Fig. S13 and S14 of the supplementary material]. Conversely, $R_{\overline{rmse}}$ increased for the majority of stations in all 24-duration scaling intervals having $d_1 < 12$ h [Fig. S15 of the supplementary material]. Note also that no particular spatial pattern characterized the $R_{\overline{rmse}}$ estimates.

7 Discussion and conclusion

This study investigated simple scaling properties of extreme precipitation intensity across Canada and the United States. The ability of SS models to reproduce extreme precipitation intensity distributions over a wide range of sub-daily to weekly durations was evaluated. The final objective was to identify duration intervals and geographical areas for which the SS model can be used for the production of IDF curves.

The validity of SS models was empirically confirmed for the majority of the scaling intervals. In particular, the hypothesis of a scale-invariant shape of the X_d distribution held for all duration intervals spanning from 1 h to 7 days based on the comparison of SS distributions to empirical quantiles. Less convincing results were obtained for durations shorter than 1 h, especially for the longest scaling intervals (24-duration intervals). One possible explanation is that the coarse instrument ~~measurement~~[R1] resolution of the available 15 min series may strongly impact both the validation tools (for instance, GOF tests) and SS estimates. These results provide important operative indications concerning the inner and outer cut-off durations for AMS scaling and show the importance of a deeper analysis to evaluate the impact of dataset characteristics (e.g., their temporal and measurement resolutions, or the series length) on the scale invariant properties of extreme precipitation[R2].

The majority of the estimated scaling exponents ranged between 0.35 and 0.95, showing a smooth evolution over the scaling intervals and a well-defined spatial structure. Six geographical regions, initially defined according to a climatological classification of North America into 20 regions, displayed different features in terms of scaling exponent values. Specifically, distinct median values of H were observed for the various geographical regions, each characterized by a different precipitation regime. This is consistent with results reported in the literature for some specific regions and smaller observational datasets (e.g., Borga et al., 2005; Nhat et al., 2007; Ceresetti et al., 2010; Panthou et al., 2014, and references therein). Moreover, while small and smooth changes of H over the scaling intervals were observed in regions containing the majority of stations, one region, SW_Pac, displayed two dramatically distinct scaling regimes separated by a steep transition occurring between a few hours and 24 h. These results limit the applicability of SS models in SW_Pac, and were connected to the local features of intense precipitation events by the analysis of the mean number of events per year and the mean wet time of these events.

Weak scaling regimes, characterized by relatively small H values (H close to 0.5), were generally observed for scaling intervals containing very short durations (e.g, less than 2 h) and for regions on the west coast of the continent [regions A1, A2, and D; see Fig. 7]. For these scaling intervals and regions, we can expect that extreme precipitation events observed at various durations will have similar statistical characteristics, being governed by homogeneous weather processes.

The interpretation of high H values (e.g., $H > 0.8$), observed between 1 and several days, depending on the region, is more complex. These scaling regimes correspond to mean precipitation depth that varies little with duration. This suggests an important change in precipitation regimes occurring at some durations included in the scaling interval. One interesting example was region SW_Pac (region D) for scaling intervals of durations longer than 1 day . In this case, the analysis of the mean number of events per year sampled in AMS suggested that very few long-duration extreme events were produced by large-scale dynamic precipitation systems.

For scaling intervals of durations longer than 4 days, scaling exponents seemed to converge to approximately 0.7 for all regions,

except west coast regions (regions A1, A2, and D).

These results suggest that SS represents a reasonable working hypothesis for the development of more accurate IDF curves. Besides, the spatial distribution of the scaling exponent and its dependency on climatology should be taken into account when defining SS duration intervals since the accuracy of the SS approximation may depend on the range of considered temporal scales. Equally critical, estimated H values were found to gradually evolve with the considered scaling intervals. In this respect, interesting extensions of the analysis should consider methods for the quantification of the uncertainty in H estimations as well as the possibility of modeling the scaling exponent as a function of both the observational duration and the AMS distribution quantile/moment order, i.e. by the use of a multiscaling (MS) framework for IDFs. Equally important, the events sampled by the AMS also showed different statistical features within different geographical regions and some specific results [e.g., for the SW_Pac region] stimulate the interest for an analysis of the scaling property of extreme precipitation by the use of a temporal stochastic scaling approach.

The evaluation of SS model performances under the assumption of GEV distributions for AMS intensity was then performed. Results indicate that the proposed SS GEV models may lead to a more reliable statistical inference of extreme precipitation intensity than that based on the conventional non-SS approach. In particular, a better assessment of the GEV shape parameter seems possible when pooling data from several durations under the scaling hypothesis. The use of the SS approximation may introduce biases in high quantile estimates when AMS distributions move drastically away from perfect scale invariance (short durations and/or longest scaling intervals). Nonetheless, decreases in the SS GEV $RMSE$ with respect to non-SS GEV models for d_1 longer than a few hours and/or scaling intervals shorter than 24 durations indicate that quantile errors in IDF estimates can be generally reduced.

Caution is advised when interpreting these results due to the fact that high order empirical quantiles were used as reference estimates of true X_d quantiles, which could be a misleading assumption especially when available AMS are short. Considering this limitation and our general results, any future extension of this study should investigate the possibility of introducing spatial information in scaling models as well as improvements of scaling GEV estimation procedures.

8 Data availability

The 15min Precipitation Data (15PD) and Hourly Precipitation Data (HPD) were freely obtained from NOAA/Climate Prediction Center (CPC) [http://www.ncdc.noaa.gov/data-access/land-based-station-data]. Hourly Canadian Precipitation Data (HCPD) ~~(H)~~[R2] and Daily Maxima Precipitation Data (DMPD) ~~(DM)~~[R2] ~~data~~[R2] for Canada were acquired from Environment and Climate Change Canada (ECCC) and from the MDDELCC of Québec [data available upon request by contacting *InfoClimat@mddelcc.gouv.qc.ca*].

Acknowledgements. Silvia Innocenti's scholarship was partly provided by the Consortium OURANOS. Financial support for this project was also provided by the Collaborative Research and Development Grants program from the Natural Sciences and Engineering Research Council of Canada (AM). They also thank Guillaume Talbot for preliminary data processing and Dikra Khedhaouiria for useful discussions.

References

Alila, Y.: Regional rainfall depth-duration-frequency equations for Canada, Water Resources Research, 36, 1767–1778, doi:10.1029/2000WR900046, http://onlinelibrary.wiley.com/doi/10.1029/2000WR900046/abstract, 2000.

Asquith, W. H. and Famiglietti, J. S.: Precipitation areal-reduction factor estimation using an annual-maxima centered approach, Journal of Hydrology, 230, 55–69, doi:10.1016/S0022-1694(00)00170-0, http://www.sciencedirect.com/science/article/pii/S0022169400001700, 2000.

Bara, M., Kohnová, S., Gaál, L., Szolgay, J., and Hlavcová, K.: Estimation of IDF curves of extreme rainfall by simple scaling in Slovakia, Contributions to Geophysics and Geodesy, 39, 187–206, 2009.

Bendjoudi, H., Hubert, P., Schertzer, D., and Lovejoy, S.: Interprétation multifractale des courbes intensité-durée-fréquence des précipitations, Comptes Rendus de l'Académie des Sciences-Series IIA-Earth and Planetary Science, 325, 323–326, http://www.sciencedirect.com/science/article/pii/S1251805097813791, 1997.

Bernard, M. M.: Formulas For Rainfall Intensities of Long Duration, Transactions of the American Society of Civil Engineers, 96, 592–606, http://cedb.asce.org/cgi/WWWdisplay.cgi?276728, 1932.

Blanchet, J., Ceresetti, D., Molinié, G., and Creutin, J. D.: A regional GEV scale-invariant framework for Intensity-Duration-Frequency analysis, Journal of Hydrology, 540, 82–95, doi:10.1016/j.jhydrol.2016.06.007, http://www.sciencedirect.com/science/article/pii/S0022169416303584, 2016.

Blöschl, G. and Sivapalan, M.: Scale issues in hydrological modelling: a review, Hydrological Processes, 9, 251–290, http://onlinelibrary.wiley.com/doi/10.1002/hyp.3360090305/abstract, 1995.

Borga, M., Vezzani, C., and Dalla Fontana, G.: Regional rainfall depth-duration-frequency equations for an alpine region, Natural Hazards, 36, 221–235, http://link.springer.com/article/10.1007/s11069-004-4550-y, 2005.

Bougadis, J. and Adamowski, K.: Scaling model of a rainfall intensity-duration-frequency relationship, Hydrological Processes, 20, 3747–3757, doi:10.1002/hyp.6386, 2006.

Bukovsky, M. S.: Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO. Downloaded (2015): 05-08., pp. 06–18, 2012.

Burlando, P. and Rosso, R.: Scaling and multiscaling models of DDF for storm precipitations, Journal of Hydrology, 187, 45–64, doi:10.1016/S0022-1694(96)03086-7, 1996.

Ceresetti, D.: Structure spatio-temporelle des fortes précipitations: application à la région Cévennes-Vivarais, Ph.D. thesis, Université de Grenoble, 2011.

Ceresetti, D., Molinié, G., and Creutin, J.-D.: Scaling properties of heavy rainfall at short duration: A regional analysis, Water Resources Research, 46, n/a–n/a, doi:10.1029/2009WR008603, http://dx.doi.org/10.1029/2009WR008603, w09531, 2010.

Coles, S., Heffernan, J., and Tawn, J.: Dependence Measures for Extreme Value Analyses, Extremes, 2, 339–365, doi:10.1023/A:1009963131610, 1999.

Coles, S. G.: An Introduction to Statistical Modeling of Extreme Values, Springer, London, 2001.

CSA: Development, interpretation and use of rainfall intensity-duration-frequency (IDF) information: Guideline for Canadian water resources practitioners, Tech. Rep. Canadian Standard Association, Tech. Rep. PLUS 4013, Mississauga, Ontario, 2nd ed., http://shop.csa.ca/en/canada/infrastructure-and-public-works/plus-4013-2nd-ed-pub-2012/invt/27030802012, 2012.

Cunnane, C.: A particular comparison of annual maxima and partial duration series methods of flood frequency prediction, Journal of Hydrology, 18, 257–271, doi:10.1016/0022-1694(73)90051-6, http://www.sciencedirect.com/science/article/pii/0022169473900516, 1973.

De Michele, C., Kottegoda, N. T., and Rosso, R.: The derivation of areal reduction factor of storm rainfall from its scaling properties, Water Resources Research, 37, 3247–3252, doi:10.1029/2001wr000346, 2001.

Devine, K. A. and Mekis, E.: Field accuracy of Canadian rain measurements, Atmosphere-Ocean, 46, 213–227, doi:10.3137/ao.460202, http://dx.doi.org/10.3137/ao.460202, 2008.

Dubrulle, B., Graner, F., and Sornette, D.: Scale Invariance and Beyond, EDP Sciences, Les Ulis, France, les Houches Workshop, march 10-14, 1997 edn., http://www.springer.com/physics/complexity/book/978-3-540-64000-4, 1997.

ECCC: Environment Climate Change Canada. Historical Climate Data Canada; editing status 2016-08-09; re3data.org - Registry of Research Data Repositories. last accessed: 2016-11-09, doi:10.17616/R3N012, http://doi.org/10.17616/R3N012.

Eggert, B., Berg, P., Haerter, J. O., Jacob, D., and Moseley, C.: Temporal and spatial scaling impacts on extreme precipitation, Atmos. Chem. Phys., 15, 5957–5971, doi:10.5194/acp-15-5957-2015, http://www.atmos-chem-phys.net/15/5957/2015/, 2015.

Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Springer Science & Business Media, 2013.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R.: Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form, Water Resources Research, 15, 1049–1054, 1979.

Gupta, V. K. and Waymire, E.: Multiscaling properties of spatial rainfall and river flow distributions, Journal of Geophysical Research: Atmospheres, 95, 1999–2009, doi:10.1029/JD095iD03p01999, 1990.

Hartmann, D. L., Klein Tank, A. M. G., Rusicucci, M., Alexander, L. V., Broenniman, B., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., Zhai, P. M., and Kent, E. C.: Observations: Atmosphere and Surface, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., pp. 159–254, Cambridge University Press, Cambridge, 2013.

Hosking, J. R. M. and Wallis, J. R.: Regional Frequency analysis: an approach based on L-moments, Cambridge University Press, 1997.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, Technometrics, 27, 251–261, 1985.

Hubert, P. and Bendjoudi, H.: Introduction à l'étude des longues séries pluviométriques, XIIème journées hydrologiques de l'Orstom, pp. 10–11, http://hydrologie.org/ACT/ORSTOMXII/VENDREDI/HUBERT/HUBERT.DOC, 1996.

Katz, R. W.: Statistical Methods for Nonstationary Extremes, in: Extremes in a Changing Climate, edited by AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., and Sorooshian, S., no. 65 in Water Science and Technology Library, pp. 15–37, Springer Netherlands, 2013.

Katz, R. W., Parlange, M., and Naveau, P.: Statistics of extremes in hydrology, Advances in Water Resources, 25, 1287–1304, 2002.

Koutsoyiannis, D.: Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation, Hydrological Sciences Journal, 49, doi:10.1623/hysj.49.4.575.54430, 2004a.

Koutsoyiannis, D.: Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records, Hydrological Sciences Journal, 49, doi:10.1623/hysj.49.4.591.54424, 2004b.

Koutsoyiannis, D., Kozonis, D., and Manetas, A.: A mathematical framework for studying rainfall intensity-duration-frequency relationships, Journal of Hydrology, 206, 118–135, doi:10.1016/S0022-1694(98)00097-3, http://www.sciencedirect.com/science/article/pii/S0022169498000973, 1998.

Kunkel, K. E., Easterling, D. R., Kristovich, D. A. R., Gleason, B., Stoecker, L., and Smith, R.: Meteorological Causes of the Secular Variations in Observed Extreme Precipitation Events for the Conterminous United States, Journal of Hydrometeorology, 13, 1131–1141, doi:10.1175/JHM-D-11-0108.1, http://journals.ametsoc.org/doi/abs/10.1175/JHM-D-11-0108.1, 2012.

Langousis, A., Carsteanu, A. A., and Deidda, R.: A simple approximation to multifractal rainfall maxima using a generalized extreme value distribution model, Stochastic Environmental Research and Risk Assessment, 27, 1525–1531, http://link.springer.com/article/10.1007/s00477-013-0687-0, 2013.

Lovejoy, S. and Mandelbrot, B. B.: Fractal properties of rain, and a fractal model, Tellus A, 37, 209–232, http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0870.1985.tb00423.x/abstract, 1985.

Lovejoy, S. and Schertzer, D.: Generalized Scale Invariance in the Atmosphere and Fractal Models of Rain, Water Resources Research, 21, 1233–1250, doi:10.1029/WR021i008p01233, 1985.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., and others: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, Reviews of Geophysics, 48, http://onlinelibrary.wiley.com/doi/10.1029/2009RG000314/pdf, 2010.

Mascaro, G., Deidda, R., and Hellies, M.: On the nature of rainfall intermittency as revealed by different metrics and sampling approaches, Hydrology and Earth System Sciences, 17, 355–369, doi:10.5194/hess-17-355-2013, 2013.

MDDELCC: Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques, 2016. Donneées du Programme de surveillance du climat, Direction geéneérale du suivi de l'état de l'environnement, Queébec.

Menabde, M., Seed, A., and Pegram, G.: A simple scaling model for extreme rainfall, Water Resources Research, 35, 335–339, 1999.

Nhat, L. M., Tachikawa, Y., Sayama, T., and Takara, K.: A Simple Scaling Charateristics of Rainfall in Time and Space to Derive Intensity Duration Frequency Relationships, Ann. J. Hydraul. Eng, 51, 73–78, http://hywr.kuciv.kyoto-u.ac.jp/publications/papers/2007AJHE_LeMinh.pdf, 2007.

NOAA: Climate Data Online; editing status 2016-06-17; re3data.org - Registry of Research Data Repositories. last accessed: 2016-11-09, doi:10.17616/R32059, http://doi.org/10.17616/R32059.

Olsson, J., Singh, V. P., and Jinno, K.: Effect of spatial averaging on temporal statistical and scaling properties of rainfall, Journal of Geophysical Research: Atmospheres, 104, 19 117–19 126, doi:10.1029/1999JD900271, http://onlinelibrary.wiley.com/doi/10.1029/1999JD900271/abstract, 1999.

Overeem, A., Buishand, A., and Holleman, I.: Rainfall depth-duration-frequency curves and their uncertainties, Journal of Hydrology, 348, 124–134, doi:10.1016/j.jhydrol.2007.09.044, 2008.

Panthou, G., Vischel, T., Lebel, T., Quantin, G., and Molinié, G.: Characterising the space–time structure of rainfall in the Sahel with a view to estimating IDAF curves, Hydrology and Earth System Sciences, 18, 5093–5107, doi:10.5194/hess-18-5093-2014, http://www.hydrol-earth-syst-sci.net/18/5093/2014/, 2014.

Papalexiou, S. M. and Koutsoyiannis, D.: Battle of extreme value distributions: A global survey on extreme daily rainfall, Water Resources Research, 49, 187–201, doi:10.1029/2012WR012557, http://onlinelibrary.wiley.com/doi/10.1029/2012WR012557/abstract, 2013.

Papalexiou, S. M., Koutsoyiannis, D., and Makropoulos, C.: How extreme is extreme? An assessment of daily rainfall distribution tails, Hydrology and Earth System Sciences, 17, 851–862, doi:10.5194/hess-17-851-2013, 2013.

Rodriguez-Iturbe, I., Gupta, V. K., and Waymire, E.: Scale considerations in the modeling of temporal rainfall, Water Resources Research, 20, 1611–1619, http://www.hydro.washington.edu/pub/lettenma/cee_599/wgr_papers/rodriguez_1984.pdf, 1984.

Sivakumar, B.: Fractal analysis of rainfall observed in two different climatic regions, Hydrological Sciences Journal, 45, 727–738, doi:10.1080/02626660009492373, http://dx.doi.org/10.1080/02626660009492373, 2000.

Sivapalan, M. and Blöschl, G.: Transformation of point rainfall to areal rainfall: Intensity-duration-frequency curves, Journal of Hydrology, 204, 150–167, doi:10.1016/S0022-1694(97)00117-0, http://www.sciencedirect.com/science/article/pii/S0022169497001170, 1998.

Tessier, Y., Lovejoy, S., and Schertzer, D.: Universal Multifractals: Theory and observations for rain and clouds, J. Appl. Meteorol., 32, 223-250, 32, 223–250, 1993.

Veneziano, D. and Furcolo, P.: Multifractality of rainfall and scaling of intensity-duration-frequency curves, Water Resources Research, 38, 1306, doi:10.1029/2001WR000372, http://onlinelibrary.wiley.com/doi/10.1029/2001WR000372/abstract, 2002.

Veneziano, D. and Iacobellis, V.: Multiscaling pulse representation of temporal rainfall, Water Resources Research, 38, 13–1, doi:10.1029/2001WR000522, http://onlinelibrary.wiley.com/doi/10.1029/2001WR000522/abstract, 2002.

Veneziano, D. and Langousis, A.: Scaling and fractals in hydrology, in: Advances in data-based approaches for hydrologic modeling and forecasting., World Scientific, Singapore, Sivakumar, Bellie and Berndtsson, Ronny edn., http://www.itia.ntua.gr/getfile/1024/2/documents/Pages_from_ScalingFractals.pdf, 2010.

Veneziano, D. and Yoon, S.: Rainfall extremes, excesses, and intensity-duration-frequency curves: A unified asymptotic framework and new nonasymptotic results based on multifractal measures, Water Resources Research, 49, 4320–4334, doi:10.1002/wrcr.20352, 2013.

Veneziano, D., Lepore, C., Langousis, A., and Furcolo, P.: Marginal methods of intensity-duration-frequency estimation in scaling and nonscaling rainfall, Water Resources Research, 43, n/a–n/a, doi:10.1029/2007wr006040, 2007.

Venugopal, V., Roux, S. G., Foufoula-Georgiou, E., and Arnéodo, A.: Scaling behavior of high resolution temporal rainfall: New insights from a wavelet-based cumulant analysis, Physics Letters A, 348, 335–345, http://www.sciencedirect.com/science/article/pii/S0375960105013253, 2006.

Wallis, J. R., Schaefer, M. G., Barker, B. L., and Taylor, G. H.: Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington State, Hydrology and Earth System Sciences, 11, 415–442, doi:10.5194/hess-11-415-2007, http://www.hydrol-earth-syst-sci.net/11/415/2007/, 2007.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, Reviews of Geophysics, 52, 2014RG000464, doi:10.1002/2014RG000464, http://onlinelibrary.wiley.com/doi/10.1002/2014RG000464/abstract, 2014.

Willems, P., Arnbjerg-Nielsen, K., Olsson, J., and Nguyen, V. T. V.: Climate change impact assessment on urban rainfall extremes and urban drainage: Methods and shortcomings, Atmospheric Research, 103, 106–118, doi:10.1016/j.atmosres.2011.04.003, http://www.sciencedirect.com/science/article/pii/S0169809511000950, 2012.

Yu, P.-S., Yang, T.-C., and Lin, C.-S.: Regional rainfall intensity formulas based on scaling property of rainfall, Journal of Hydrology, 295, 108–123, doi:10.1016/j.jhydrol.2004.03.003, 2004.

Table 1. List of available datasets and their main characteristics.

Dataset	Region	N. of stations	Operational period[b]	Temporal resolution	Prevalent[c] resolution [mm]
Daily Maxima Prec. Data[R2] [a] (DMPC[R2])	Canada	370	1964-2007	1, 2, 6, 12 h	0.1 (82.25%)
Hourly Canadian Prec. Data (HCPC) ~~(H)~~[R2]	Canada	665	1967-2003	1 h	0.1 (70%)
Hourly Prec. Data (HPD)	USA	2531	1948-2013	1 h	0.254 (82.5%)
15-Min Prec. Data (15PD)	USA	2029	1971-2013	15 min	2.54 (80.42%)

[a] Daily maxima depth series over a 24-hour window beginning at 8:00 AM.

[b] Main station network operational period corresponding to 25^{th} percentile of the first recording year and the 75^{th} percentile of the last recording year of the stations.

[c] Prevalent instrument~~measurement~~[R1] resolution, estimated by the lowest non-zero value for each series, and corresponding percentage of stations with this resolution.

Table 2. Final datasets used in scaling analysis and corresponding AMS characteristics.

Scaling dataset	Durations	N. of Stations	Mean series length [yr]	Max series length [yr]
SD[a]	15min, 30min, ..., 6h	1083	20	36
ID	1h, 2h, ...,24h	2719	37.4	66
LD	6h, 12h, ..., 168h	2719	37.4	66

[a] Only 15PD series.

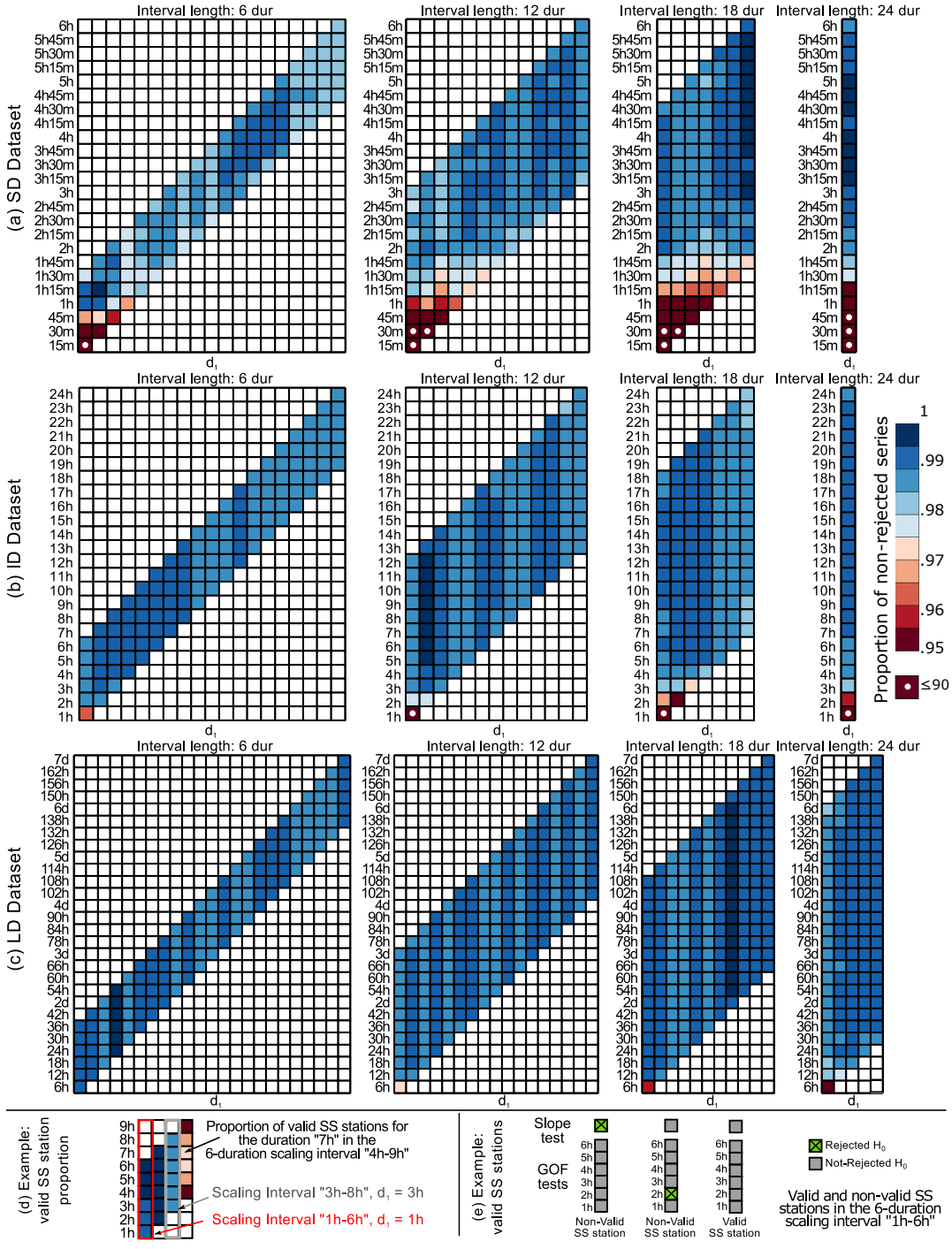

Figure 1. a) - c) Proportion of stations satisfying both the Slope and GOF tests applied at the 0.95 confidence level, for each duration (vertical axis) and scaling interval (horizontal axis) for the SD, ID, and LD datasets [row a), b), and c) respectively]. White circles indicate proportions between 0.25 and 0.90; d) Example of valid SS station proportion values and identification of durations and scaling intervals within each matrix; e) Examples of valid and non-valid SS stations.

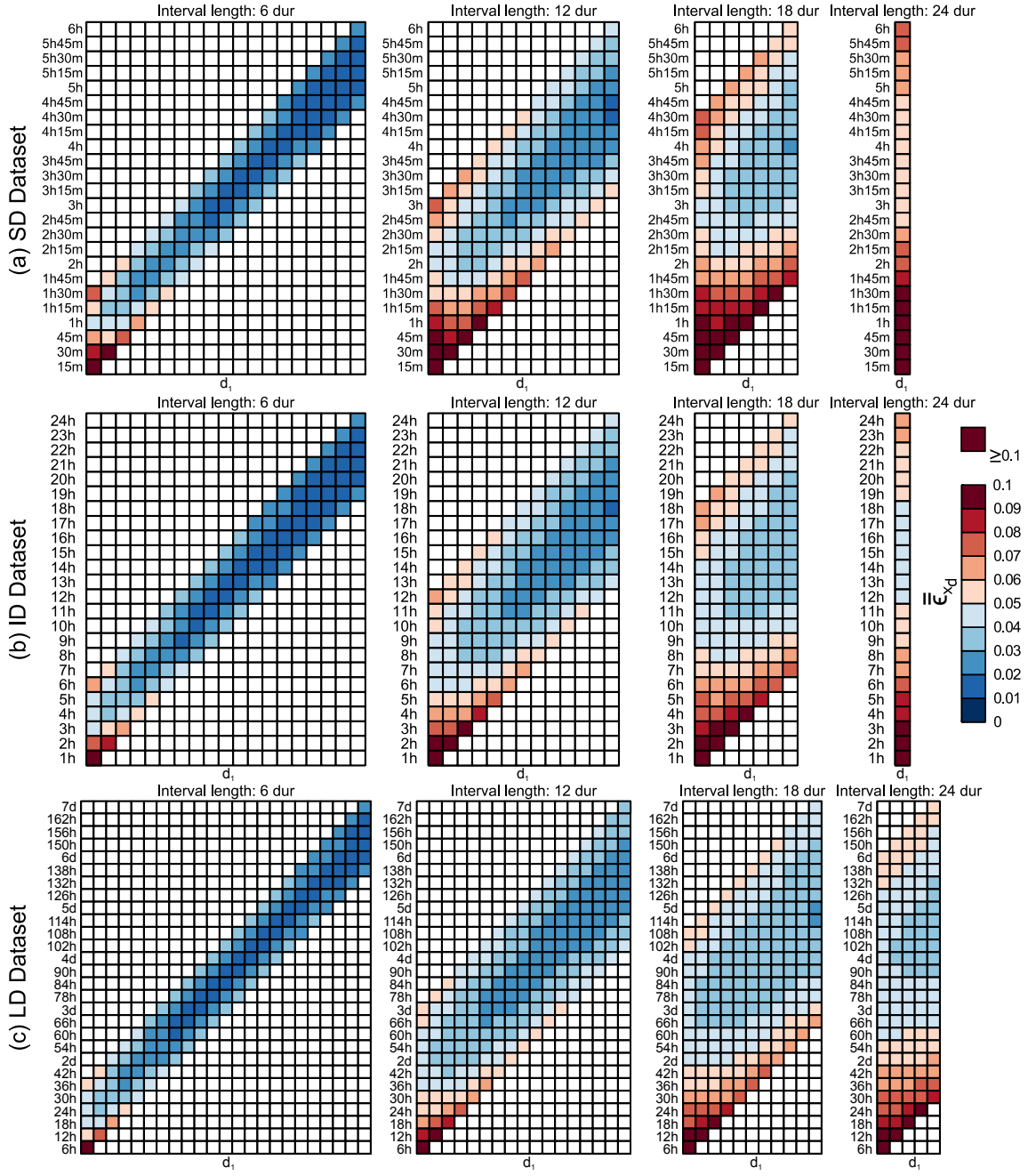

Figure 2. Cross-Validation Normalized RMSE averaged over all valid SS stations $(\bar{\bar{\epsilon}}_{x_d})/(\bar{\bar{r}}_{x_d})^{R1}$ for each duration (vertical axis) and scaling interval (horizontal axis) in the SD, ID, and LD datasets [row a), b), and c) respectively]. See Fig. 1 (d) for the identification of durations and scaling intervals within each matrix. White circles indicate values between 0.15 and 0.3.

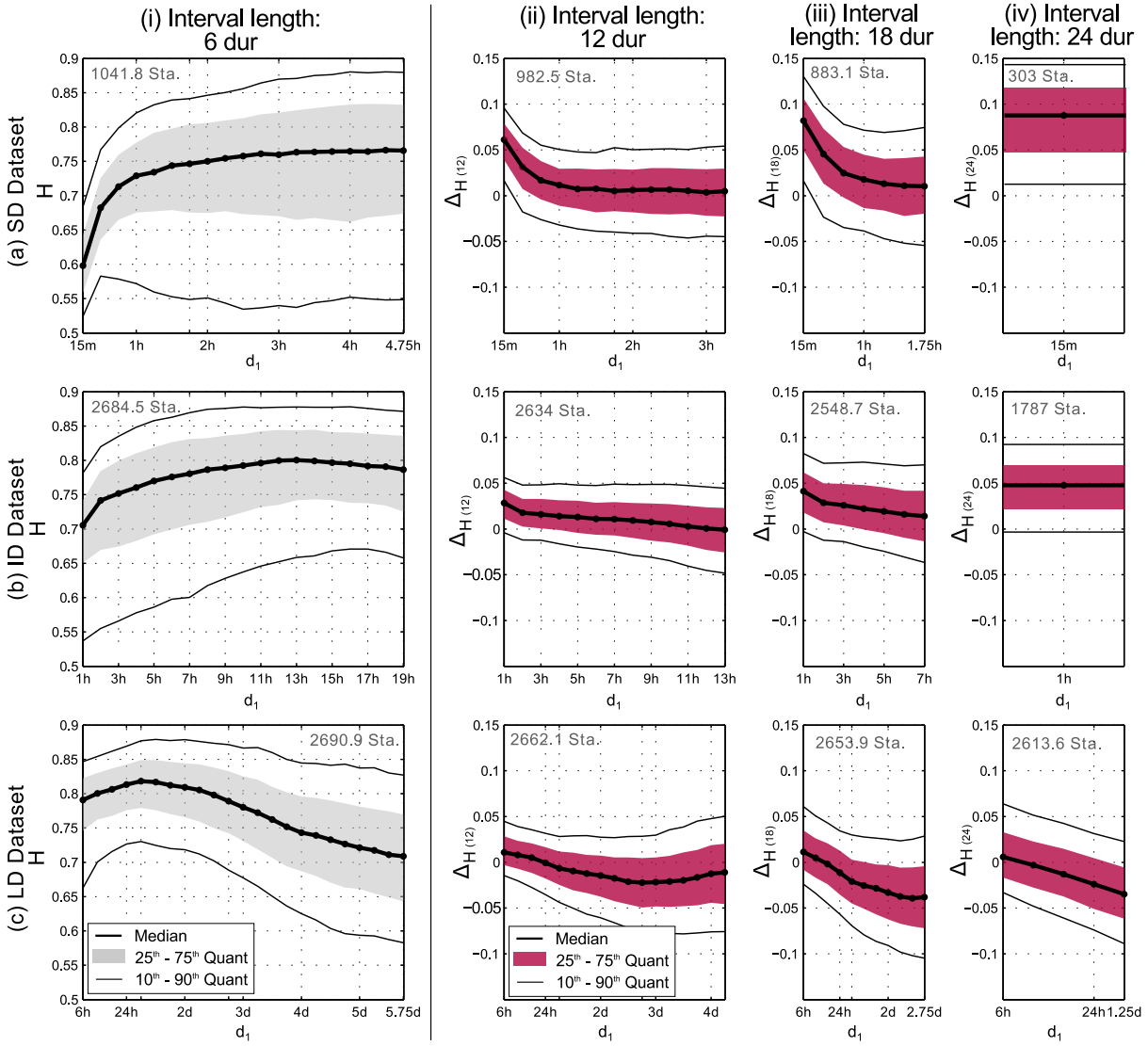

Figure 3. Col. (i): Median and relevant quantiles of the scaling exponent distribution over all valid SS stations for each 6-duration scaling interval. Col. (ii)-(iv): Median and relevant quantiles of the distribution of the scaling exponent deviation $\Delta_{H_{(j)}}$ [defined in Eq. (12)]. The average number of valid SS stations over the scaling intervals (identified by their first duration, d_1) is indicated at the top of each graph.

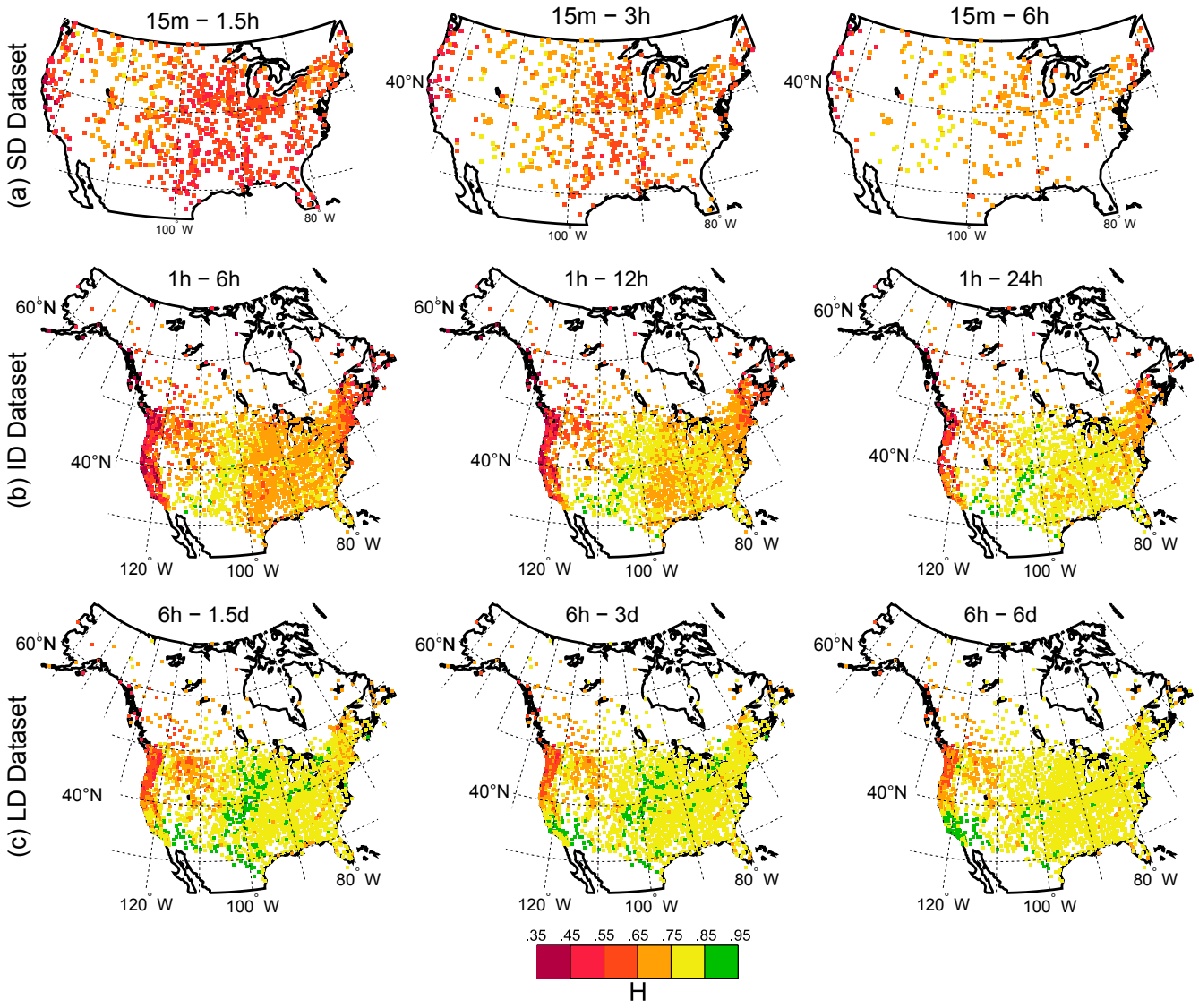

Figure 4. Spatial distribution of the scaling exponent for the first (i.e. with minimum d_1) 6-, 12-, and 24-duration scaling intervals (first, second, and third col., respectively) for SD, ID, and LD datasets (first, second, and third row, respectively). These scaling intervals correspond to the first column of matrices in Fig. 1 and 2.

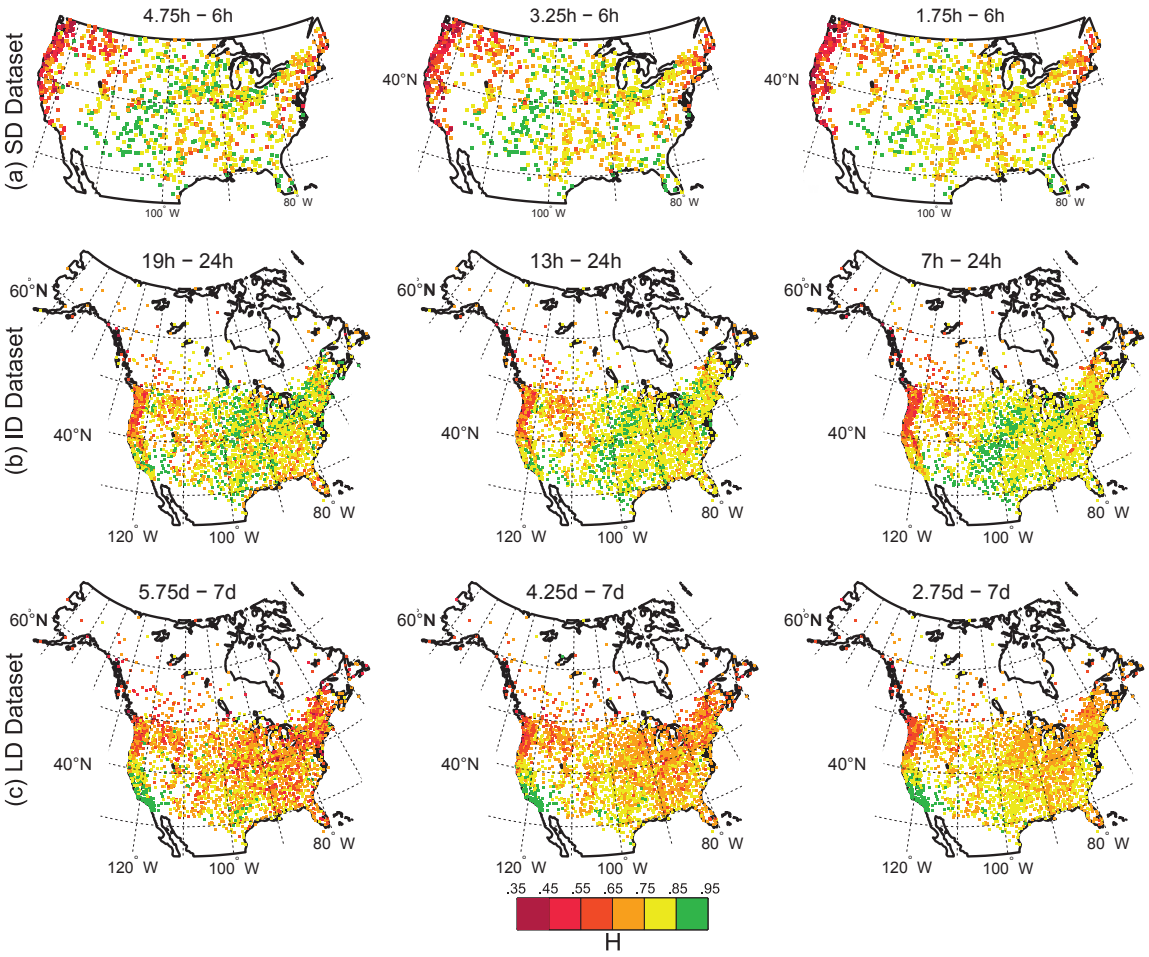

Figure 5. Spatial distribution of the scaling exponent for the last (i.e. with maximum d_1) 6-, 12-, and 18-duration scaling intervals (first, second, and third col., respectively) for SD, ID, and LD datasets (first, second, and third row, respectively). These scaling intervals correspond to the last column of matrices in Fig. 1 and 2.

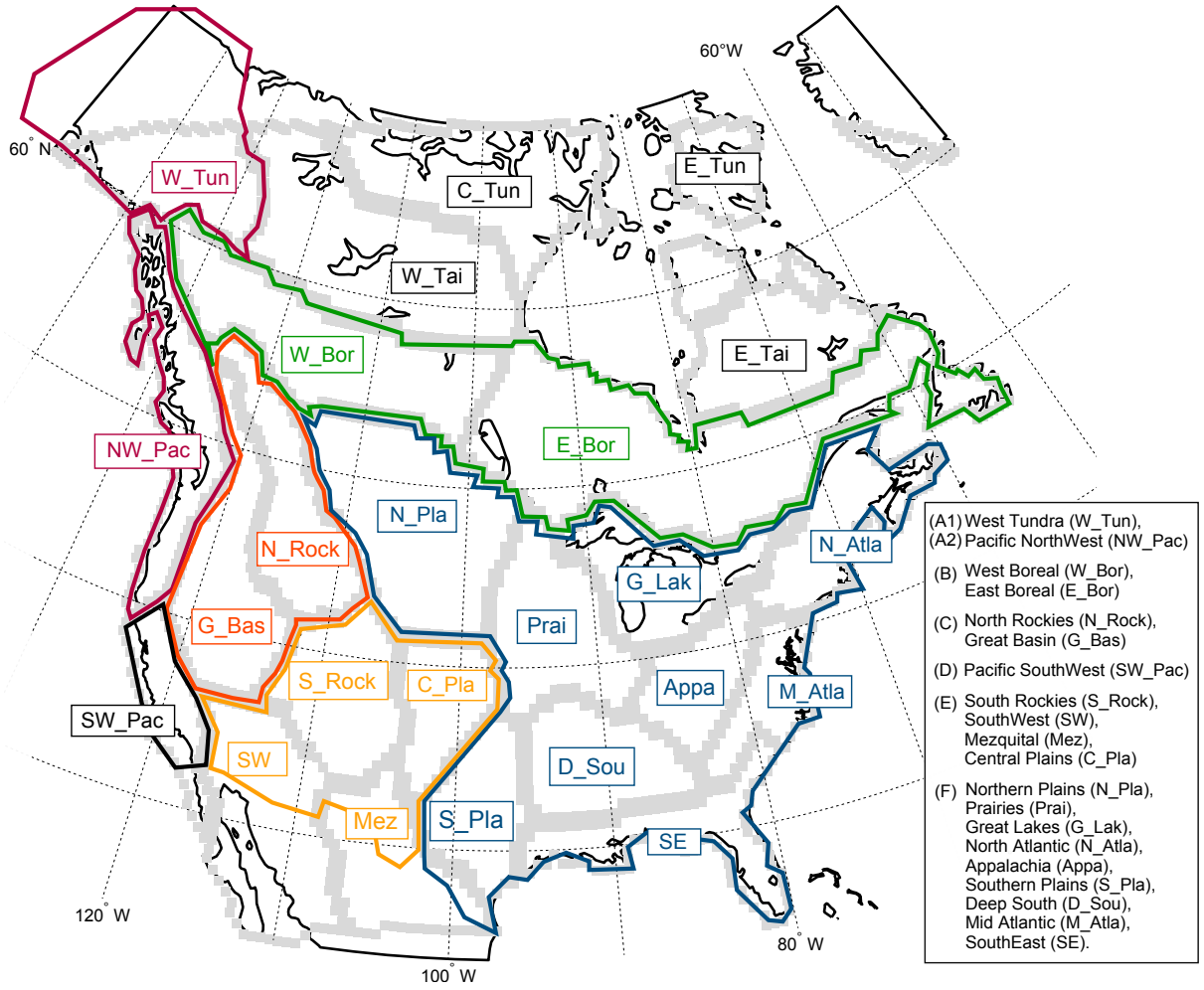

Figure 6. Climatic regions of Bukovsky (2012) [grey borders] and regions defined for this analysis [regions A1 to F in the legend; colored borders]. Abbreviations for each region are in parenthesis.

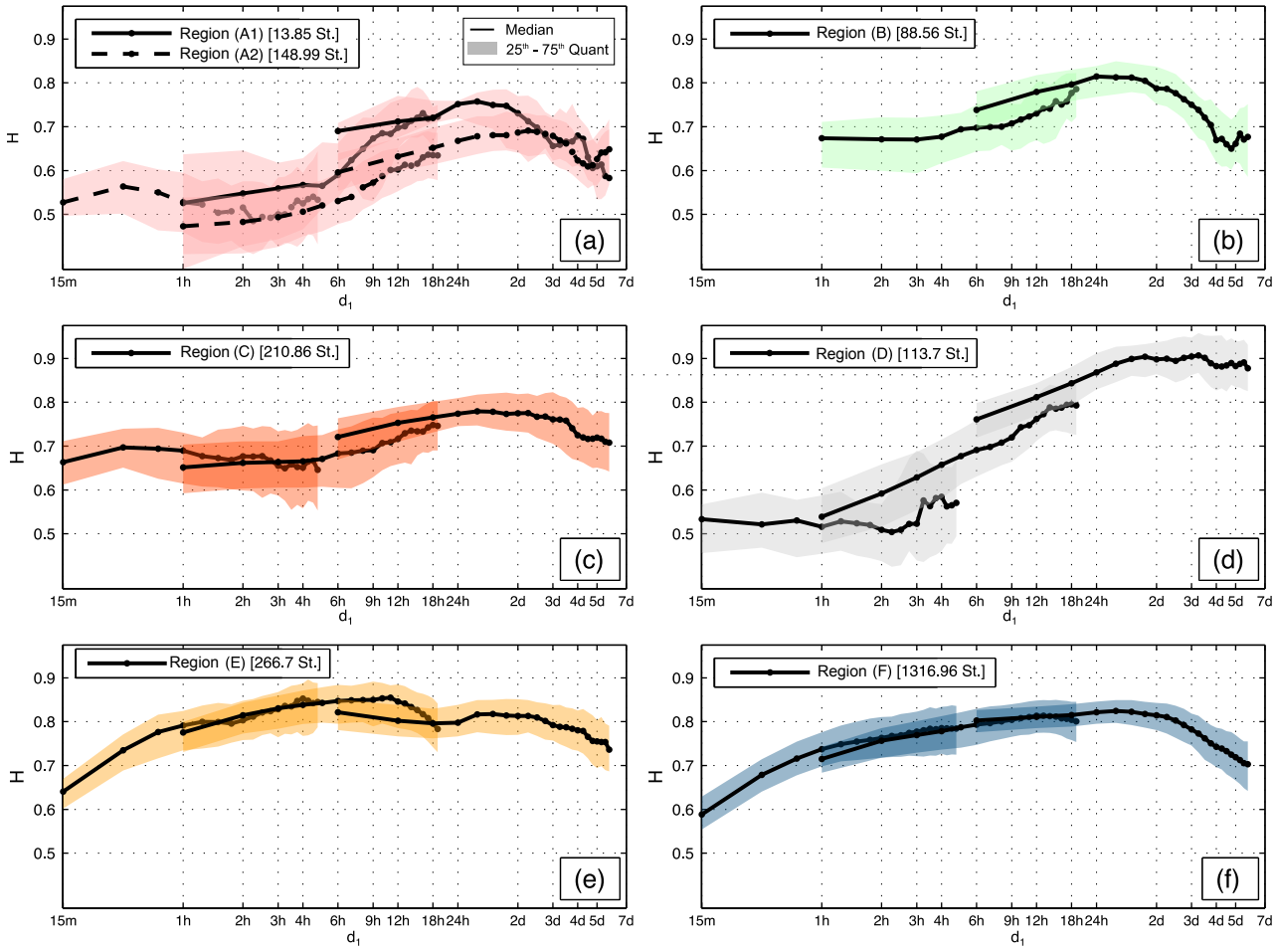

Figure 7. Median and Interquantile Range (IQR) ~~(IR)~~[R1] of the scaling exponent distribution over valid SS stations within each region of Fig. 6 for 6-duration scaling intervals for the SD (left curve), ID (central curve), and LD (right curve) datasets. For each region, the mean number of valid SS stations over the scaling intervals is indicated in brackets in the legend. See Fig. 6 for region definition.

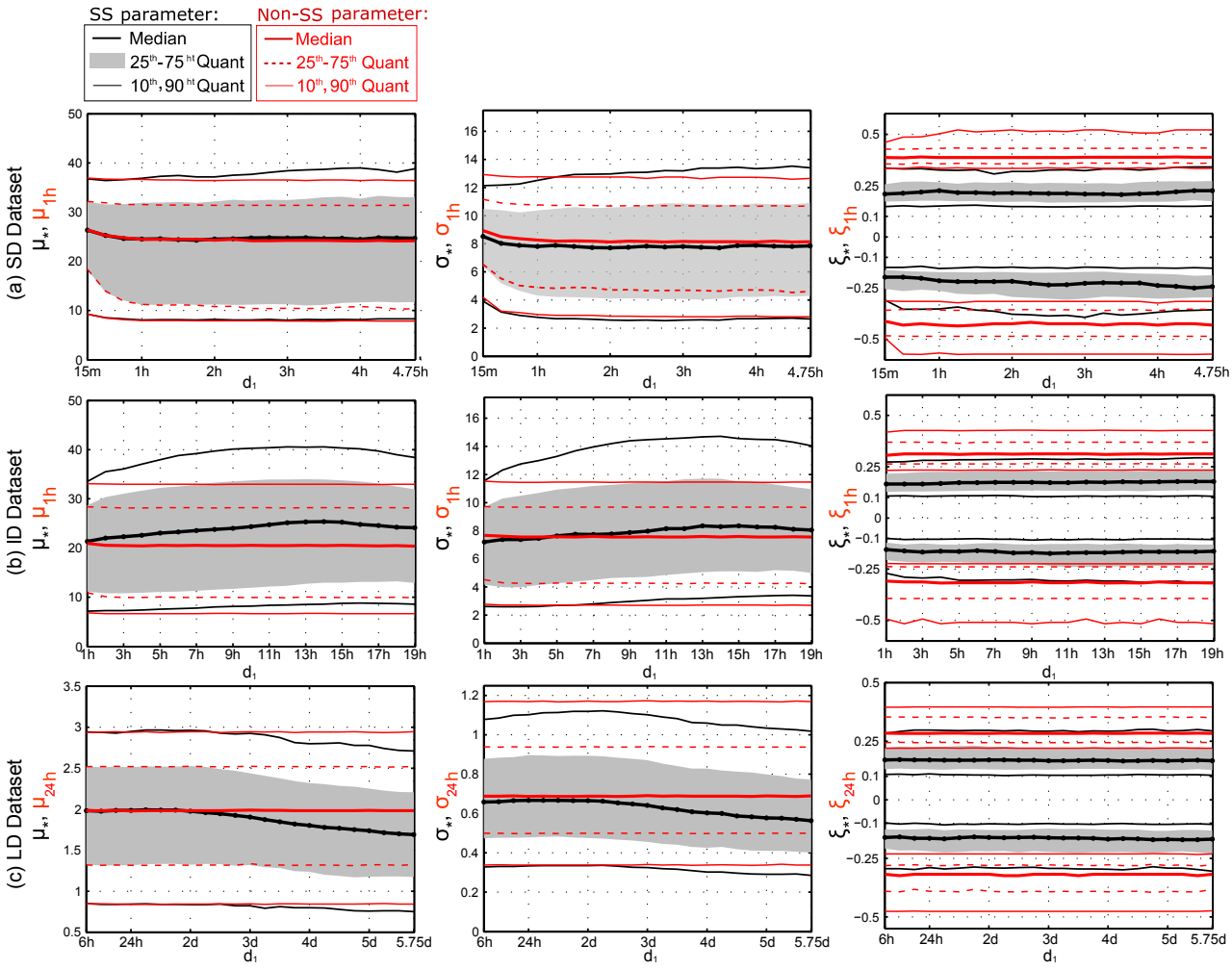

Figure 8. Distribution over valid SS stations of SS GEV (gray and black lines) and non-SS GEV (red solid and dashed lines) parameters for 6-duration scaling intervals. Location and scale parameters (first and second col., respectively) are scaled at $d_* = 1$h (SD and ID datasets) and $d_* = 24$h(LD dataset). Distributions for the shape parameter (third col.) are presented for $\xi > 0$ and $\xi < 0$, excluding cases where $\xi = 0$ (Gumbel distribution).

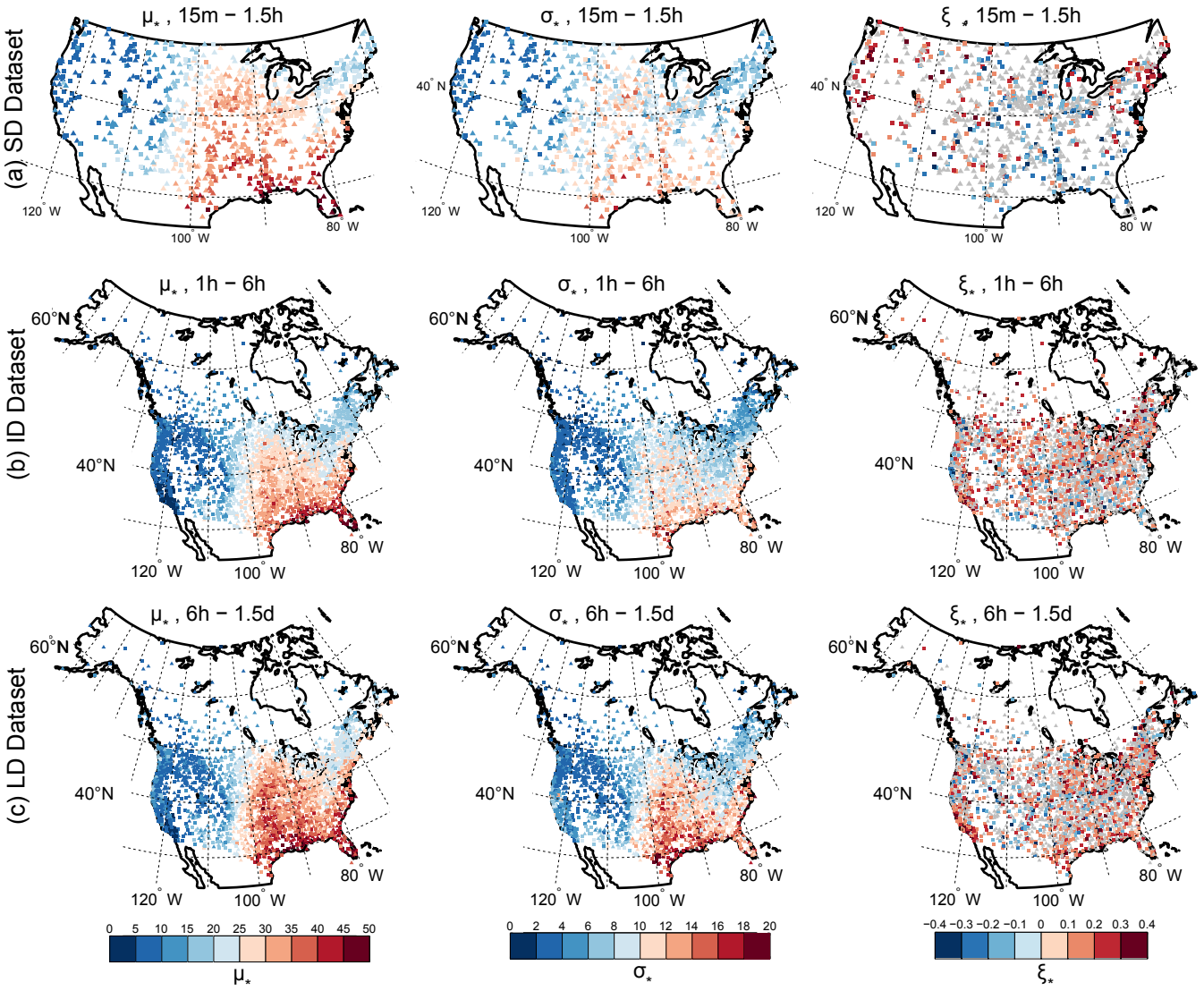

Figure 9. Spatial distribution over valid SS stations of SS GEV position (first col.), scale (second col.), and shape (3^{rd} col.; gray symbols indicate Gumbel distributions, $\xi_* = 0$) parameters scaled at $d_* = 1$h for the first 6-duration scaling interval (i.e. interval with minimum d_1) of: resolution SD (a), ID (b), and LD (c) datasets.

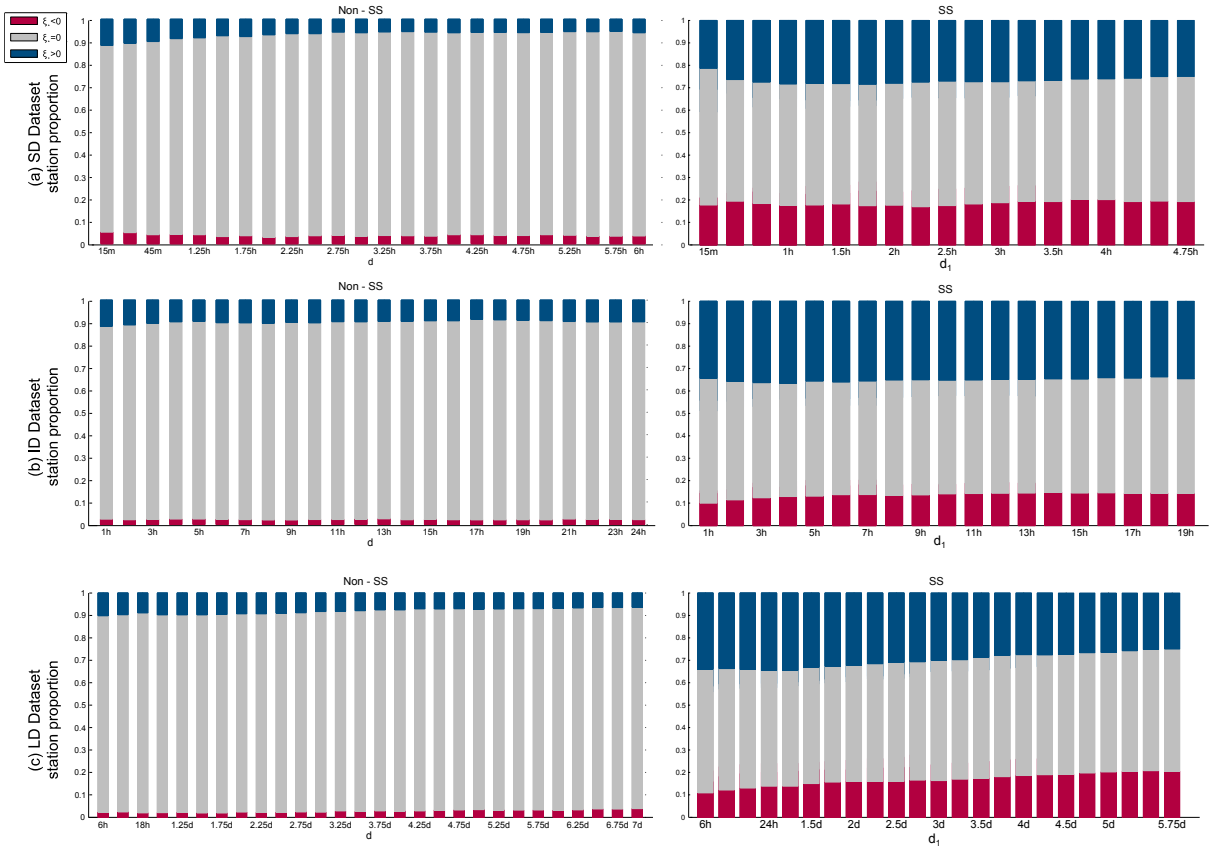

Figure 10. Stacked histograms of the fractions of valid SS stations with $\xi < 0$ (in red), $\xi = 0$ (in grey), and $\xi > 0$ (in blue) resulting from the Hosking test applied at the 0.95 confidence level [R2] for each duration (non-SS GEV, first col.) and each 6-duration scaling interval (SS GEV, second col.) for: SD (a), ID (b), and LD (c) datasets.

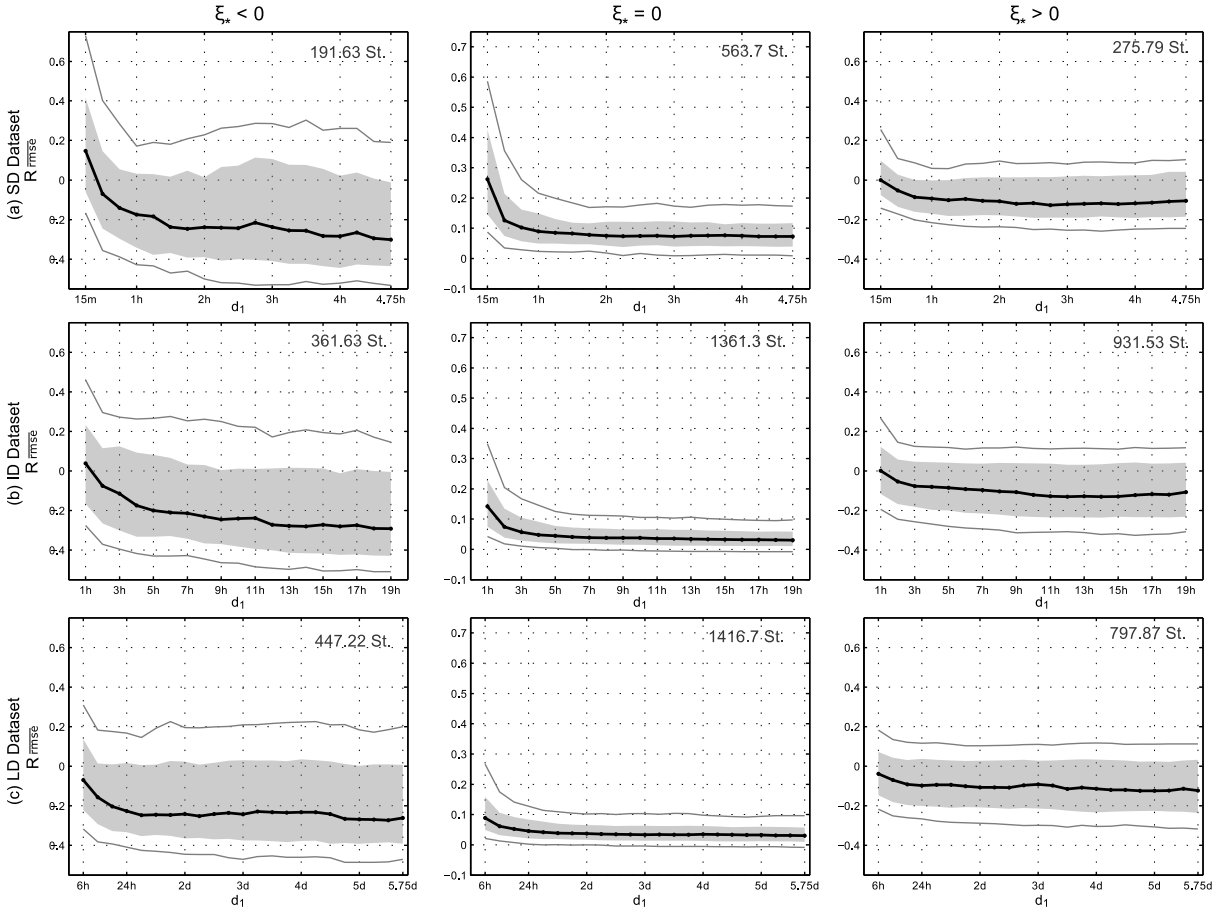

Figure 11. Distribution of the relative total RMSE ratio, $R_{\overline{rmse}}$, for $\xi_* < 0$ (first col.), $\xi_* = 0$ (second col.), and $\xi_* > 0$ (third col.) for 6-duration scaling intervals in SD (a), ID (b), and LD (c) datasets. The average number of valid SS station over the scaling intervals is indicated in the right-top corner of each graph.