Dear Editor,

Please find enclosed the second revision of the manuscript **"Simple Scaling of Extreme Precipitation in North America"** by Innocenti et al. to be considered for publication in HESS.

According to comments from the first and third reviewers, we modified several paragraphs resulting in a clearer and more comprehensive description of methodology and results, in particular for Sect 4 to 6. Methodological choices for the regional analysis (Section 4.2) are discussed and reviewed in our reply to the first reviewer. In particular, we explained and motivated in details our methodology for the definition of geographical regions in North America in our reply to comment 15 a). Some additional analysis has been also presented in the reply to this specific comment. Finally, important considerations and further explanations concerning the results presented for Sect. 6 (SS GEV models) have been added.

We provide below a detailed response to each comment and a copy of the revised manuscript in "track changes" mode. The following specific colors are used to link corrections and reviewers' comments:

- blue^{R1} is used to underline changes related to first referee comments,
- red^{R2} is used for changes related to the second referee comments,
- purple^{R3} is used for changes related to the third referee comments,
- and gray is used for all other changes.

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

Sincerely,

Silvia Innocenti, on behalf of the co-authors.

Authors' response to 1st referee's comments

This article is too long, very difficult to read, and has almost convinced me that Simple Scaling does work: e.g. Figure 3i and Figure 7, the scaling coefficient H is all over the places, whereas one would expected it to be constant over different durations, being a scale-invariant parameter. An other example is my last comment (just citing your text). Also, it is not clear to me why the authors chose the Bukovsky regions for their regional analysis, when they have a clerar clustering in their Figures 4 and 5 (and S1) which suggests a different regionalization. Moreover, the physical interpretation in Section 5 is not clear (how do you link the behaviour of H across different durations with the topography, as an example). Often the exposition does not follow a logical order. I doubt several results in Section 6. My overall comment is that the authors dwelve too much into the details, but they fail to convey the main message (at least to me ...). I regret to say that my recommendation is a rejection: I leave however the decision to the Editor, since I am not sure I have fully understood the study.

My numerous comments:

1. Equation 2 (and through out the article): I suggest changing the notation by replacing above the equality sign (=) the letter “d” with “pdf”. This avoid confusion with “d” denoting the duration.

   Thank you for the suggestion but the relationship in Eq. 2 is more general and does not only concern the pdf. To avoid confusion, we replaced $\overset{d}{=}$ by $\overset{dist}{=}$ in **Eq. 2**, at **Line 25, Page 3**, and at **Line 23, Page 5**.

2. Page 5, the sentence at line 12-13-14 should be stated earlier, after line 9.

   The paragraph has been moved to **Lines 6 to 10, Page 5**.

3. Page 5, lines 24-25: The equalities at line 24 pertains a dilatation of the (values of the) GEV distributions by a factor lambda, whereas Equation 2 pertains to a change of distribution due to a sampling on different durations, with a dilatation factor = lambda. The implication stated at line 25 (the GEV family satisfies Equation 2) is not a direct consequence of the equality at line 24. Rephrase (you might have to add a specific reference, or show explicitely the implication).

   We understand the reviewer perspective and we agree that we did not clearly explain these points in Section 2. In particular, we did not explicitly mention that SS models assume statistical scale-invariance with respect to the sampling scale [as stated, for instance, at **Lines 9 to 12, Page 2**].
   In order to explicitly mention this point in the text we rewrote **Lines 20 to 22, Page 3** as:

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

   $X_d \overset{dist}{=} \lambda^H X_{\lambda d}$"

   We also agree that the paragraph describing the scale-invariant nature of the GEV distribution may be misleading since equations at **Line 23, Page 5** only imply the validity of Eq. (1) for the GEV cdf with respect to a constant multiplicative factor λ. Then, Eq. (2) follows if we consider the GEV cdf scale invariance being valid when changing the observational scale from d to λd. To be more precise, we thus rephrased **Lines 24 to 27, Page 5** as:

"This means that the GEV family described by Eq. (5) and (6) satisfies Eq. (1) and thus complies with statistical scale invariance for any constant multiplicative transformation of X. When this scale invariance is further assumed for the change of observational scale from duration d to λd [as in Eq. (2)], the wide sense SS definition [Eq. (3)] gives: ... [Eq. (7)] ".

4. Equation 7: I would still prefer you replace d with lambda (since the protagonist here is the dilatation factor)! Here and throughout the article, wherever consistency is required.

We appreciate the reviewer's suggestion but, in our opinion, it is important, here, to highlight the fact that the AMS distribution for any duration d can be always reparametrized in terms of the GEV distribution parameters of an unit scale used as reference duration ($d^* = 1$). This notation allows in fact to underline that GEV parameters μ_d, σ_d, and ξ_d can be thought almost as adimensional parameters for any arbitrary set of durations in the range of scale for which SS is valid.

In our application [Sect. 6], the use of $d_* = 1h$ and, thus, of $\lambda = d$ [**Line 3, Page 4** and **Line 3, Page 4**] allows to compare the parameters of different scaling intervals without the specification of a particular λ for each interval, and to underline the physical interpretation of the SS GEV parameters in the different scaling intervals.

5. Page 6, line 19: I appreciate your clarification wrt my previous comment. In light of that explanation, I suggest adding at the end of this sentence " .. from May to October in the south, and from June to September in the north. Specifically, the 'year' from which the annual maxima was sampled, was limited to the observed summer season, and was defined as ... ".

To improve the text we modified **Lines 23 to 28, Page 6** to:

"For this reason, the *year* from which the annual maxima was sampled was limited to the recording season going from June to September for northern stations [stations located north of the 52^{nd} Parallel] and from June to September for the southern stations. As a result, 122 days a year were used for northern stations and 184 days a year for remaining stations."

6. Page 7, line 4: write " ... at least 85% of valid observations for each summer season, otherwise ... "

Thank you for the suggestion. Since the May to October or June to September periods don't correspond to the summer season, we changed the sentence for the following [**Line 10, Page 7**]:

"at least 85% of valid observations for each May to October (or June to September) period, otherwise the corresponding year was considered as missing."

7. Page 6, line 23: substitute 'cheked' with 'quality controlled'.

Done.

8. Page 7, line 13: typo: "finstance".

Corrected, thank you.

9. Page 7, line 30: use the new notation 1h30 rather than 1.5h ...

   Done: the paragraph has been modified after the addition of Fig. 1. [see our reply to your comment 12 a)].

10. Page 7, line 23-24: you need to better state what MSA does. Line 23 is fine, “for intensity AMS” is too concise: you could write something like “for inferring annual maxima / extremes / GEV parameters for durations not sampled by using SS on sampled durations” or something similar.

    The expression has been replaced by *"for modeling AMS empirical distributions"* [**Line 30, Page 7**].

    For seek of precision, we did not add "for durations not sampled by using SS on sampled durations" since SS models can be used for modeling sampled duration distributions, and not only for non-sampled ones. Accordingly, in fact, we estimated SS and evaluated its performances in both "calibration" and "cross-validation" mode [see also reply to comment 12 e) - (Section 4)].

11. Page 7, line 27: write “scaling intervals” in italic, so that the reader understand this is a definition (valid from hereafter throughout all the article).

    Done.

12. **Section 4:**

    a) I do still believe that the article will gain in adding a figure showing i) the linear regression between log(duration) and log(moments) and ii) the (previously obtained) regession coefficients K_q and q (as Figure 4a in Panthou et al, 2014). These figures were fundamental for me to understand page 8, lines 8 to 16 (i.e. what you do in the MSA regression and in the slope test).

    Following your suggestion we added a new figure [Fig. 2 on **Page** 27] which explains the various steps of the methodology used for non-parametric SS estimation [Section 4]. In particular, the new Fig. 1 contains six panels representing the following steps:

    Panel a): Definition of the SD, ID and LD scaling datasets.

    Panel b): Example of the definition of five scaling intervals for the ID dataset. This panel is intended to help interpreting Fig.2 and 3.

    Panel c): MSA regression: estimation of K_q slope coefficients.

    Panel d): Slope test: testing linearity of coefficient K_q on q.

    Panel e): Definition of Valid SS stations.

    Panel f): Example of Valid SS station proportion and Normalized RMSE ($\overline{\overline{r}}_{x_d}$) as presented, respectively, in Fig. 2 and 3 [Fig. 1 and 2 in the previous version of the manuscript].

    The following references were also added in the text:

    - at **Line 9, Page 7**: *"[see Figure 1(a)]"*;
    - at **Line 7, Page 8**: *"More schematically, Fig. 1(b) shows an example of the first five 6-duration scaling intervals for the ID dataset [i.e. 1h - 6h, 2h - 7h, . . . , 5h - 10h, containing six contiguous durations defined with an increment of 1h]."* ;

- at **Line 17, Page 8**: *"[see Fig. 1(c) for a graphic example]"*;
- at **Line 20, Page 8**: *"[see Fig. 1 (d)]"*;
- in the **captions of Fig. 2 and 3**: *"See Fig. 1 (b) and (f) for the identification of durations and scaling intervals within each matrix."*.

b) In the text regarding the MSA regression, you need to add a sentence which explains that the slope Kq that you are evaluating is equal to -Hq (it took me a while to figure this out), and that this is from Equation 4. Also, at line 8 write “for each q ... ”, so that it is clear that you have a K_q for each individual separate q.

Thank you for the suggestion but the equality $K_q = H\ q$ is what we want to evaluate with slope test, while the MSA regression should not assume this linear relationship: as explained at **Lines 7 to 8, Page 4**, SS models can be considered valid only if $K_q \approx Hq$.

As you suggested, we added "each" at **Line 15, Page 8** which now reads: *"for each $q = 0.2, 0.4, \ldots, 2.8, 3$, the slopes K_q of the log-log linear relationships between the empirical $q-$moments ...."*

c) In the slope test text (page 8, lines 12-16) you need first to say that you aim to find the scaling coefficient H, which is the slope of the line you obtain while regressing Kq versus q. Then you say that you use a OLS regression and estimate H, and you call such estimate β_1 (I would call it $\hat{H}$). Finally you can say that you perform a t-test on the regression. I suggest stating that the null hypothesis was $\hat{H} = K_1$, and that if this null hypothesis is not rejected then you set $H = \hat{H} = K_1$. I would avoid whiting the relation $Kq = \beta_0 + \beta_1 q$ (otherwise the reader will ask where β_0 is gone, for the equality of $K_1 = \beta_1$).

We apologize for this lack of clarity. To improve the description of this methodological step, which aim at testing the linearity of the K_q scaling exponents estimated at the previous step, we modified **Lines 19 to 25, Page 8** to:

"To verify the SS assumption that the estimated K_q exponents vary linearly with the moment order q, i.e. $K_q \approx Hq$, an OLS regression between the MSA slopes K_q and q was applied [see Fig. 1 (d)]. For the regression line $K_q = \hat{h}_0 + \hat{h}_1 q$, a Student's t-test was then used to test the null hypothesis $\mathbf{H}_0$: $\hat{h}_1 = K_1$. If $\mathbf{H}_0$ was not rejected at the significance level $\alpha = 0.05$, the SS assumption was considered appropriate for the scaling interval and the simple scaling exponent $H = K_1$ was retained."

d) The Goodness of Fit test (page 8 lines 19-26) can be explained more clearly, following the cronologiocal order of the calculations you perform: first, you consider the AMS for each duration d_j, AMS $d_j = \{xd_j, 1, x_{d_j,2}, ... x_{d_j,n}\}$, and you rescale it as $d_j H \cdot$ AMS d_j, which can be considered as a sample of the reference duration (the choice of notation x'_{d_j} is confusing, since one would think it is a sample for the duration d_j, whereas it is a sample for the reference duration). You then pool together these rescales samples for all dj. Then you invert equation 2 and obtain, from this large sample, a sample for each duration dj, under SS hypohesis.

Following your suggestion we changed notation for the description of the SS sample construction and we rephrased **Lines 28 to 13, Page 8** to:

"To this end, each AMS, $\boldsymbol{x}_{d_j} = \left(x_{d_j,1}, x_{d_j,2}, \ldots, x_{d_j,i}, \ldots x_{d_j,n}\right)$, recorded at duration d_j was rescaled at the reference duration d by inverting Eq. (2):*

$$\boldsymbol{x}^*{}_{d_j} = \left({d_j}^H x_{d_j,1}, {d_j}^H x_{d_j,2}\ , \ldots\ , {d_j}^H x_{d_j,i}, \ldots {d_j}^H x_{d_j,n}\right) \tag{8}$$

where n represents the number of observations (years) in $\boldsymbol{x}_{d_j}$. Then, the pooled sample, $\boldsymbol{x}_{d^*}$, of the D rescaled AMS, $\boldsymbol{x}^*{}_{d_j}$, was used to define X_{d^*} under the SS assumption:

$$\boldsymbol{x}_{d^*} = \left(\boldsymbol{x}^*{}_{d_1}, \ldots, \boldsymbol{x}^*{}_{d_j}, \ldots, \boldsymbol{x}^*{}_{d_D}.\right) \quad (9)$$

Since, in Eq. (9), D represents the number of durations d_j in the scaling interval, $n \times D$ rescaled observations were included in $\boldsymbol{x}_{d^*}$."

e) I suggest performing the GOF test also in a cross validation way (e.g. for the ID dataset, when you test the duraton of 3h, you rescale to the duration of 1h all durations excluded the 3h; then you invert Eq.2 and you apply the test to the obtained -slightly smaller- samples).

Thank you for pointing out our omission. The GOF and Slope test were also applied during the cross-validation experiment, which consists of repeating steps 1 to 3 described on **Page** 8 [see below], but we did not present the relative results since these are not substantially different to those of Fig. 2 [Fig. 1 in the previous version]. For completeness, we added these results (Slope and GOF tests in cross-validation) in Fig. S4 of the supplementary material.

To clarify that the steps 1 to 3 were repeated in a cross-validation setting, **Lines 24 to 29, Page 9** were modified to:

"The SS model validity and the mean error resulting from approximating the X_d distribution by the SS model were then evaluated in a cross-validation setting. For this analysis, each duration was iteratively excluded from each scaling interval and the scaling model re-estimated at each station by repeating steps 1 to 3 [MSA regression, Slope test, and GOF tests]".

Then, we added the following reference to cross-validation results presented in Fig. S4 [**Line 13, Page 11**]:

"These findings were also confirmed by cross-validation experiments. The proportion of valid SS stations resulting from cross-validation Slope and GOF tests were similar, event if slightly lower, to proportions displayed in Fig. 2 [see Fig. S4 of the supplementary material].

Moreover, the following explanation has been added in the introduction of the supplementary material:

"Figure S4 presents the results for the cross-validation experiment for the SS model (Slope and GOF tests) for each duration and scaling interval. "

f) Page 9, lines 21-23: this is not entirely clear, do you repeat all steps associated to the GOF part only (page 8, line 7 onwards), or also for the MSA regression and slope test? Only after reading page 11 line 1 I understood that you repeat all (points 1,2,3 at page 8). Make the sentence clearer.

The sentence has been modified specifying that steps 1 to 3 of the methodology [MSA regression, Slope test, and GOF tests] were repeated in cross-validation experiments. Please, see our reply to the previous comment.

g) Text from page 9 line 27 to page 10, line 4: similar to what suggested for the GOF test, describe what you calculate in order of calculation, i.e. define first the normalized RMSE for each station (Eq 11) and after the average over all stations (Eq 10).

Following your suggestion **Lines 1 to 13, Page 10** were modified to:

"For each station s, the normalized RMSE, $\overline{\epsilon}_{x_{d,s}}$, was estimated:

$$\overline{\epsilon}_{x_{d,s}} = \frac{\epsilon_{x_{d,s}}}{\overline{x}_{d,s}} \qquad (8)$$

where $\epsilon_{x_{d,s}}$ and $\overline{x}_{d,s}$ are, respectively, the RMSE and the mean value of all X_d quantiles of order $p > 0.5$. Then, the average over all stations of the normalized RMSE, $\overline{\overline{\epsilon}}_{x_d}$, was computed for each scaling interval and duration:

$$\overline{\overline{\epsilon}}_{x_d} = \frac{1}{n_s}\sum_{s=1}^{n_s}\overline{\epsilon}_{x_{d,s}} \qquad (9)$$

where n_s is the number of valid SS stations in the dataset. Note that $\overline{\overline{\epsilon}}_{x_d}$ 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. ".

13. **Section 4.1:**

a) Figure 1 shows results of slope test and GOF test together. However a reader is intrigued in disentageling the two. You have actually looked at the results seperately, and decided to put them together, with the knowledge that solely the GOF test has a signal. The reader, however, does not know this and remains in the doubt up unti reading at the end of page 10 (lines 25-27). I suggest you to move the sentence at lines 25-27 at the very beginning of the description of these results, after line 10. After this sentence, you should state that the differences in station rejections for the separate durations as shown in Figure 1 are due to the results of GOF test only. And then you keep describing Figure 1 as in your lines 11-25 (which I assume refers solely to the GOF test: this should be made more explicit). Eliminate the [not shown] at line 13 (I think you show this, with the SD dataset)

We agree with the reviewer and we modify the first paragraph of section 4.1 to [see **Line 15, Page 10**]:

"Figure 2 presents the results of steps 1 to 3 of the methodology for evaluating the SS validity. For all the three scaling datasets, no particular pattern was observed for slope test results, with at most 2% of the stations within each scaling interval displaying a non linear evolution of the scaling exponent with the moment order. For this reason, Fig. 2(a)-(c) show, for each scaling interval and duration, the proportion of valid SS stations without differentiating for slope or GOF test results. "

b) Figure 2: Page 11, lines 2-4: rewite this as “On the other hand, the extrapolation under SS of the Xd distribution is generally less accurate for durations at the boundaries of the scaling intervals (especially for the short durations)”. You do not show/perform extrapolations of Xd by estimating H with durations outside the scaling interval (as far as I can see), and I believe you meant to rewrite the sentence deleted at line 5.

Thank you fo the suggestion. Apply the cross-validation for durations at the boundaries of the scaling intervals results in the estimation of H and X_d using non-recorded durations (i.e. durations outside the scaling interval since one boundary is moved). For instance, the first cross-validation for the interval 1h-6h in the ID dataset consists in the estimation of the SS model on durations 2h, 3h, 4h, 5h, and 6h. Then, this model is evaluated for 1h which technically is outside of the scaling interval "2h-6h" used for the estimation. For this reason, we modified **Line 20, Page 11** to:

" Conversely, the extrapolation under SS of the X_d distribution is generally less accurate for durations at the boundaries or outside the scaling interval used to estimate H. "

c) Figure 2: I would be curious to see the slope and GOF test results also for the cross validation experiment (as in Figure 1, to be able to compare them). This actually is related also to my previous comments e) and f) for Section 4. In alternative, you could reproduce Figure 2 for the non-cross- validation calculation.

Please, refer to reply to comment 12 e): Fig. S4 showing Slope and GOF test results for cross-validation experiments has been added to the supplementary material.

14. **Section 4.2:** After your introductory sentence (page 11, lines 10-12) I suggest describing first the results pertaining to ΔH (Figure 3 ii, iii, iv) and after the results pertaining to the spatial variability of H (Figure 3 i, and Figures 4 and 5). My suggestions for improvement are:

a) Join the paragraph at lines 13-16 with the paragraph at lines 20-25 (page 11).

The paragraphs are now contiguous.

b) Eliminate (move) the sentence at page 11 lines 26-27 to page 12 line 11, where you will start the description of the spatial variability of H.

Done. **Lines 7 to 10, Page 12** now reads:

"Figures 4(ii)-(iv) show the median, Interquantile Range (IQR), and quantiles of order 0.1 and 0.9 of the $\Delta_{H_{(j)}}$ distribution over valid SS stations for all relevant scaling intervals."

c) The text at lines 29-33 is difficult to follow, give (for each sentence) a precise reference to the figure panels.

Thank you for the suggestion. The references to figures have been added and the paragraphs has been rewritten. Please, see our reply to your following comment.

d) You might want to discuss first the results at page 12 lines 3-10 (which are positive, showing near zero $DeltaH$) and after the results at page 11, line 29 to page 12 line 2 (which are more detailed and negative).

The paragraphs have been inverted [**Lines 10 to ??, Page 12**]:

"Adding new durations to the scaling intervals, median $\Delta_{H_{(j)}}$, as well as its IQR, increased for all d_1. 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. 4 (ii)-(iii)].
For some stations, a dramatic difference could exist in IDF estimations obtained with the different definitions of the scaling interval. For instance, for the 24-duration scaling interval "1h - 24h" (ID dataset), the median $\Delta_{H_{(24)}}$ was equal to 0.047 [Fig. 4(iv) b)]. 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 [Fig. 4(iv) a)]. Finally, 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. 4 (ii)]."

e) Eliminate lines 13-17 and the related sentence at lines 30-31 of page 12 (this is too technical and does not add meat to the article, but rather distracts the reader).

Thank you for the suggestion. The sentence needed to be rephrased as also pointed out by Reviewer 3 [see comment

3 of the 3rd Reviewer]. However, we consider that the paragraph is important for the interpretation of Fig. 4 and 5 and following results since it highlights an important issue on the uncertainty of the H estimator. In this regard, note that Reviewer 3 did not suggest to eliminate the sentence but to rephrase it. To simplify the text without eliminating all the information, **Lines 34 to 12, Page 12** have been modified to:

"This result could be partially explained by the use of scaling intervals having equally spaced durations. This implies that the mean distance between the logarithms of durations in the scaling interval decreases as d_1 increases. Hence, the OLS estimator of H used in the MSA regression may have larger variance for longer d_1, especially when scaling intervals include few durations. Larger uncertainty may thus have an impact on the H estimation for the longest d_1 scaling intervals of SD. However, as showed in next sections, H spatial distribution may also explain the greater variability of the scaling exponent for d_1 greater than a few hours."

15. Section 5:

a) Figure 4 and 5 (and Figure S1) show clear spatial clustering in the behaviour of H (as commented in my previous revisions): I still think that the authors should apply a cluster analysis to their own data. Climatology and extremes are different, and extremes might not follow the Bukovsky regions. You can develop a spatial model for SS and IDF estimation (as you state at page 15, lines 16-18) solely considering a regionalization based on the extreme behavior (rather than pre-set regions).

The reviewer raised an interesting question regarding the construction of homogeneous region for extremes, which, as underlined by the reviewer, generally display a different spatial distribution than climatology. Two distinct issues must be discussed to correctly address the reviewer's comment: on the one hand, the choice of a methodology for the definition of geographical regions which is consistent with our analysis, and, on the other hand, the difference between "regionalization and regional estimation" evoked by the reviewer and the more general concept of "spatial model" for IDF parameters mentioned in our text.

To address the problem of the identification of homogeneous regions for extremes, a panoply of algorithms based on different approaches has been developed in the literature [e.g., Grimaldi et al, 2011; Hosking and Wallis, 1997; among others]. According to several authors [e.g., Wazneh et al., 2015; and references therein], the definition of homogeneous regions involves a great amount of subjectivity due to the choice of :

i) the definition of *'homogeneity'* and the subsequent choice of the site *characteristics* or *statistics* to be used for the classification [for instance, should one use the geographical location, elevation, climatological features, and/or precipitation extreme indexes for the observed sites?];

ii) the number of regions and some basic classification criteria [e.g., should be the regions be geographically contiguous?] which depend on several factors such as the spatial scale of the analysis and on the subsequent use of the classification [e.g., should one use the regionalization for approximating unknown quantities at un-gaged sites or it is only a descriptive tool?];

iii) the algorithm performing the classification based on the selected classification variables [site characteristic(s) and/or site statistic(s)];

Points i)-iii) are also closely interconnected. For instance, if one uses a cluster analysis based on site extreme precipitation statistics, stations from different geographical areas could be pooled in the same group [see figure below]. This non-spatially-contiguous classification may lack of physical consistency and may be hardly interpreted in term of climatology.

An example is provided in Figure 1 which shows the results of a basic *k-means cluster analysis* performed on H (scaling exponent) and $\bar{P}_{24h}$ (climatic median of station AMS for the duration $d = 24h$) for the scaling interval presented in Fig. 5 [Fig. 4 of the previous version of the paper mentioned by the reviewer]. The number of groups

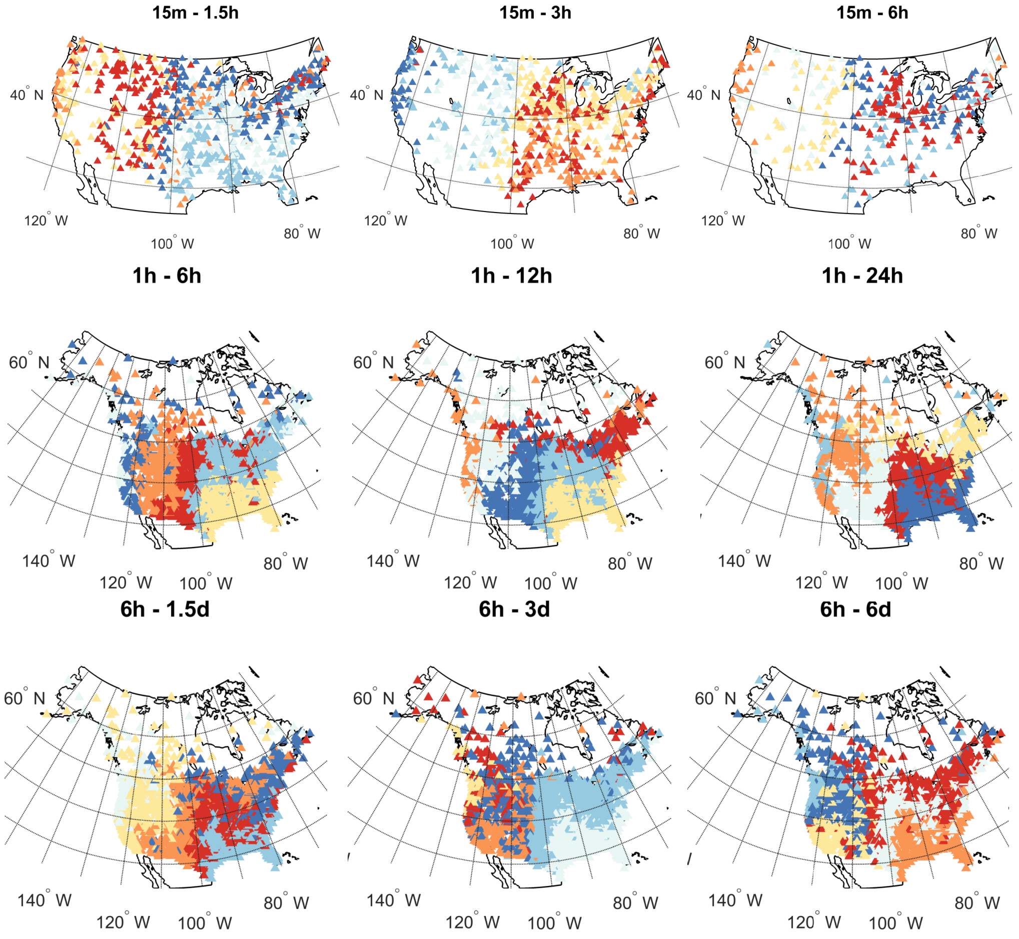

Figure 1. Example of regions obtained after applying a k-mean cluster analysis applied on H (scaling exponent) and $\bar{P}_{24h}$ (climatic median of station AMS for the duration $d = 24h$) for scaling interval presented in Fig. 5 of the paper. Stations of the same color belong to the same cluster.

has been arbitrarily set to $N_{reg} = 6$, since six geoclimatic regions were used in the paper based on Bukovsky (2012) classification [see Fig. 7]. However, more objective criteria should be defined to guide the selection of an appropriate number of clusters.

Even if, at a first sigh, the clustering in Fig 1 does not seem to significantly differ with the one used in our analysis, it is important to highlight that:

a) As observed by the reviewer in its previous revision, some regions used in the paper may be split in various subregions; however, these differences seem to be mainly driven by the NW-to-SE gradient of $\bar{P}_{24h}$ (see Fig. S1 of the supplementary material).

b) As previously noted, the method produces non-contiguous regions which limits the possibility of interpreting the physical meaning of the values taken by the scaling exponent at a regional scale in terms of geoclimatic characteristics of the regions. To improve the performances of the clustering method and allow for the definition of physically sound regions, one should probably include other classification variables, such as the geographical coordinates, the mean Total Annual Precipitation and Temperature, or other climatological indexes.

c) The classification is based on the H and $\bar{P}_{24h}$ which are at-site statistics of rainfall extremes estimated on available AMS. The estimation of this unknown quantities entails an amount of uncertainty which may impact the classification in an unpredictable way, especially if observed series are short. Conversely, climatic statistics such as the mean Total Annual Precipitation and Temperature can be generally estimated with lower uncertainty.

Essentially, more sophisticated methods should be used to provide an accurate and physically consistent clustering in order to rigorously define the spatial structure of H and, eventually, to approximate AMS distributions with a regional approach, as recommended by the reviewer. Although interesting, we sincerely believe that these analyses deserve to be comprehensively done and are out of scope of this study. In that perspective, it is important to remind that :

Firstly, our study did not intend, at this point, to use a regional approach to estimate SS models and IDF curves; the main objective of Section 5, is to investigate the relationship between the spatial distribution of H and the geographical and climatological features of the study region. This qualitative analysis was merely descriptive and did not aim at defining a regional and strictly homogeneous value of H.

Secondly, to our opinion, H can be realistically expected to be a climatological characteristic of extremes with a smoother behavior in space than other extreme statistics [e.g.,high order moments or quantiles of extreme precipitation distributions]. This is also confirmed by maps in Fig. 5 [Fig. 4 in the previous version].

Finally, the Bukovsky (2012) regions correspond to a subdivision of North America territory into homogeneous climatological regions that was already been used in the literature [e.g., Separovic et al., 2013; Prein et al., 2016;]. Our analysis could hardly reproduce the accuracy of this classification with the available data. Hence, Bukovsky regions appeared to be a convenient and reasonable choice for our study.

We realized that some confusion may arise by the expression "spatial model" used in the following sentence [**Line 12, Page 16**]: *"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"*. Note that this sentence did not refer to to regional estimation approaches such as RFA (Regional Frequency Analysis). These regionalization approaches effectively rely on an accurate definition of homogeneous regions and the validation of homogeneous regions in order to pool the series from stations with similar characteristics. However, **Line 12, Page 16** referred to "spatial models" aiming at explicitly modeling the spatial distribution of the GEV and/or IDF parameters such as in Begueria, and Vicente-Serrano (2006), Blanchet and Lehning (2010), Panthou et al (2012), and Davison et al. (2012). To avoid any confusion we rephrased the sentence at **Line 12, Page 16** as:

"Even more important, these results give useful guidelines for modeling the spatial distribution of H, which could help for the definition of IDF relationships at non-sampled locations."

- *Begueria, S and SM Vicente-Serrano (2006). Mapping The Hazard Of Extreme Rainfall By Peaks Over Threshold Extreme Value Analysis And Spatial Regression Techniques. J Appl Meteorol. Vol. 45. no. 1, pp. 108–124.*
- *Blanchet, J, Marty, C, and M Lehning (2009). Extreme value statistics of snowfall in the Swiss Alpine region. Water Resour Res. Vol. 45. no. 5.*
- *Blanchet, J. and M. Lehning (2010). Mapping snow depth return levels : smooth spatial modeling versus station interpolation. Hydrol Earth Syst Sc. Vol. 14. no. 12, pp. 2527–2544.*
- *Davison, Anthony C, SA Padoan, M Ribatet, et al. (2012). Statistical modeling of spatial extremes.*
- *Grimaldi S, Kao S-C, Castellarin A, Papalexiou S-M, Viglione A, Laio F, Aksoy H and Gedikli A (2011) Statistical Hydrology. In: Peter Wilderer (ed.) Treatise on Water Science, vol. 2, pp. 479–517 Oxford: Academic Press.*
- *Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-moments. Cambridge, UK, 244pp*
- *Panthou, G., T. Vischel, T. Lebel, J. Blanchet, G. Quantin, and A. Ali (2012). Extreme rainfall in West Africa : A regional modeling. Water Resour Res. Vol. 48. no. 8, n/a–n/a.*
- *Prein, A. F., Holland, G. J., Rasmussen, R. M., Clark, M. P., and Tye, M. R. 2016. Running dry: The US Southwest's drift into a drier climate state. Geophysical Research Letters, 43(3), 1272-1279.*
- *Separovic L, A Alexandru, R Laprise, A Martynov, L Sushama, K Winger, K Tete, M Valin. 2013. Present climate and climate change over North America as simulated by the fifth-generation Canadian regional climate model. Clim Dyn 41:3167-3201. DOI 10.1007/s00382-013-1737-5.*

- *H. Wazneh, F. Chebana, T.B.M.J. Ouarda, Delineation of homogeneous regions for regional frequency analysis using statistical depth function, Journal of Hydrology, Volume 521, February 2015, Pages 232-244, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2014.11.068.*

b) page 13, lines 9-23: I am not sure the two statistics described here add too much to the article (unless they help the physical interpretation of Section 5.1, which is obscure to me -see following comment-): in fact, I believe that their behaviour is expected, from their definition: shorter duration have a larger number of events, which decay the longer is d_1 (S4); conversely, the mean wet time per event decay as d1 grows (averaged on longer duration) ... I suggest moving all this to the supplementary material (also the related text at page 14 lines 7-8, 10-13, 23-32). The implications at page 14, line 13 is not clear.

Thank you for the suggestion but we consider that $\bar{N}_{eve}$ and $\bar{T}_{wet}$ concretely show some important features of the extreme events sampled by the AMS observed for the set of durations included in the scaling intervals. Therefore, we didn't move this paragraph to the supplemtary material. We agree that their overall behavior and patterns over d_1 are generally expected. However, the differences among panels of Fig. S5 and S6 (Fig. S4 and S5 of the previous version of the supplementary material) are crucial and support the physical interpretation of H and the regional analysis given at **Page 14**.

16. **Section 5.1:**

a) From page 13 line 30 to page 14 line 3: The link between the behaviour of H and the physical characteristics of the precipitation in the region is missing / not clear. Similarly, at page 14 lines 4-7, the implication is not clear at all.

In order to improve the discussion of Fig. 8 we recalled that *"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."* [**Line 28, Page 4**] This was also stated at **Line 29, Page 15**, **Line 7, Page 16**, and **Line 27, Page 19** ,and at **Line 3, Page 15** for $H_{depth} = 1 - H$. In this regard, note that, as reported in the literature, 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); this was mentioned at **Line 20, Page 4**.

Hence, we modified **Lines 17 to 25, Page 14** to:

"For $d_1 \leq 24$ h, Fig. 8 (a) displays lower values of H than Fig. 8 (e)-(f), meaning that smaller variation in AMS moments are observed in A1 and A2 when the scale is changed. This difference can be partially explained by the weaker impact of convection processes in generating very short duration extremes in North-West coastal regions with respect to southern areas (regions E and F). For northern regions, in fact, the transition between short and long duration precipitation regimes may be smoothed out by cold temperatures which moderate short-duration convective activity, especially for W_Tun (region A1). The topography characterizing the northern pacific coast may then explain the smoothing effect for the curve of region NW_Pac (A2). In this case, in fact, the 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)."

b) Maybe you need to explain beforehand what does it means (physically) when H is small and when H is large (as at page 15, lines 13-15), when H increase and when H decreases with d_1.

Please, see our reply to the previous comment. As you mentioned, **Lines 15 to 29, Page 4** already list the results in the literature that reports the physical interpretation of H and gives details about its spatial distribution over several regions. In the initial version of the paper this explanation was part of the introduction of Sect. 5, just before the interpretation of our results. However, they have been moved following Editor's suggestion to improve the separation between the methods, results and discussions. In this way, the practical interpretation of high and low values of the scaling exponent had been logically linked to its statistical meaning [Eq. (2) and lines below].

In the actual version of the manuscript, Section 5.1 is structured as follow: the section first describes the regional distribution of H; then the connection between H spatial patterns to the geographical and climatological characteristics of each region is made; finally a general interpretation of the results fis given at **Lines 2 to 11, Page 16**.

c) Page 14, lines 14-22: good interpretation of the behaviour in Region D.

Thank you.

d) Page 15 lines 2-5: clear, whereas you loose me at lines 6-9.

Thank you for pointing out our lack of clarity. **Lines 24 to 1, Page 15** have been rephrased to:

"However, H shows a smoother increase in Fig. 7 (f) with respect to Fig. 7(e). This may indicate that in eastern areas [region F] sub-daily duration extremes are more likely associated to embedded convective and stratiform systems or to mesoscale convective systems, which are less active in western dry areas of region E (Kunkel et al., 2012). On the contrary, differences between short- and long-duration extreme precipitation intensity seem stronger for south-western dry regions [Fig. 8 (e)], where less intense summer extremes are expected compared to eastern areas [see supplementary material, Fig. S1]. In particular, H tended to scatter in a range of higher values for approximately 1 h $\leq d_1 \leq$ 12 h indicating that precipitation intensity moments strongly decrease as the duration increases."

e) page 15, lines 10-11: I disagree with this sentence, I have not seen any evidence that the material illustrated here supports these results (you have not proven this).

In Section 5.1 we showed that different geographical regions generally displayed distinct distribution of H for different scaling intervals [panels (a)-(f) of Fig. 8 show different curves]. These regions are characterized by different climates and precipitation regimes [please, see also our reply to comment 15 a) for a discussion of the importance of choosing geo-climatological regions for this analysis]. The suggestion we made at **Line 2, Page 16** is thus that one can suppose a link (even qualitative) between the difference observed for H values [e.g, for the black lines in Fig. 8 representing the median of H in each region] and the differences between the specific climates and weather regimes characterizing each region. We think that the results presented in the paper qualitatively support the hypothesis that there exist a link between scaling and climatology even if it is not proven in a statistical or mathematical way. That was we use the term "suggest". However, to prevent any confusion we rephrased this sentence as:

"In summary, these results suggest that both local geographical characteristics, such as topography or coastal effects, and general circulation patterns may influence precipitation scaling at a regional scale."

17. **Section 6:**

a) From page 15 line 30 to page 16 line 7: since you are assessing a distribution, why don't you use a KS statistics, rather than inventing an engeneered metric which compares the quantiles? Recall Equation 11 for clarity, please.

Thank you for the suggestion but we preferred to limit the used of the GOF test (AD and KS) to the non-parametric estimation of SS models (Section 4) for several reasons. Primarily, these tests have a low statistical power so it is always preferable to limit their use to situations in which no alternative exists; this was the case in Section 4. Secondly, the use of a discrepancy measure such as RMSE allows to numerically evaluate the mean errors on quantile estimates; conversely, GOF tests can only state the statistical significance of the results according to a specified significance level. For this reason, in Sect. 4 we also estimated the Normalized RMSE after the evaluation of SS validity via the use of the KS and AD tests. Finally, the use of GOF tests is not strictly necessary to validate the use of the SS-GEV models considering that i) the use of the SS hypothesis for approximating the AMS distributions has been legitimated by results in Section 4 and that ii) the GEV distribution is the only natural distribution to use for Annual Maxima Series. Hence, no direct logical alternative exists.

As you suggested, the following reference to Eq. (11) has been added at **Line 4, Page 17**:

"See Eq. 10 for the definition of $\epsilon_{d,mod}$ for each station."

b) Page 16, line 13: for the ID and LD datasets, the behaviour of the SS parameters versus non-SS parameters is opposite, they cannot be both more right skewed for the SS estimation.

Thank you for pointing out this error. We agree with the reviewer that, while for the ID dataset "both μ_* and σ_* distributions were more positively skewed than the corresponding non-SS distributions", this is not true for LD. Hence, we corrected the sentence to [**Lines 9 to 13, Page 17**]:

"Similarly, for 6 h $\leq d_1 \leq$ 2 days in the LD dataset, the SS location and scale parameter distributions are in relatively close agreement with the corresponding non-SS parameter distributions. Conversely, in the ID dataset, both μ_* and σ_* distributions are more positively skewed than the corresponding non-SS distributions. Finally, for $d_1 \geq$ 2 days in the LD dataset, μ_* and σ_* had distributions shifted toward lower values than μ_{24h} and σ_{24h}."

c) Page 16, line 16: it seems to me that the discerepancies between SS and non-SS parameters for the LD dataset and long durations (the $\Delta\mu$ and $\Delta\sigma$ in the supporting material) and quite big.

Thank you for pointing out our mistake: the colorbar label was expressed in percent scale, while $\Delta\mu$ and $\Delta\sigma$ where defined in relative scale. The error has been corrected by changing the colorbar labels of Fig. S12 and S13 to, respectively, $\Delta\mu\%$ and $\Delta\sigma\%$.

Once these corrections made, one can observed that the relative biases in μ and σ estimates were small in most cases since the *"median values of Δ_μ and Δ_σ were generally smaller than $\pm 5\%$ and $\pm 10\%$"* [as stated at **Line 15, Page 17**].

d) Page 16, lines 14-21: there is something strange about your results: you state that the estimation of the scale parameter has a small uncertainty when the shape parameter is correctly estimated. However, from Figure 8. I can see that the shape parameter is not correctly estimated (red and black curves do not match). What is the implication of the mismatch of the shape parameter for the non-SS and SS estimation on your results? The shape parameter is usually set to zero because it is all over the places, and often not-significantly different from zero (as you find in your Figure 9c: the estimated shape parameter is quite noisy)!!!

Figure 9 presents the distributions of SS and non-SS GEV parameters across all stations (these are not at-site differences). For the shape parameter (Fig. 9, 3^{rd} col.), one has to remind (as specify in the caption) that the case $\xi = 0$ was not considered [see also our reply to comment 17 h)]. Hence, the fact that the red and black curves don't match simply indicate that the shape parameters estimated using the SS hypothesis are different than the non-SS estimates when the shape parameter is significantly different from zero.

We argue that the SS model allows a better assessment of the the shape parameter values than the non-SS model.

In our application, in fact, the majority of stations have non-SS shape parameters ξ_d which are non-significantly different from zero while these fractions are substantially reduced under the SS model [see Fig. 11]. In this regard, note that many authors have shown that the fact of setting $\xi = 0$ *"may lead to important underestimations of the extreme quantiles quantiles (e.g., Koutsoyiannis, 2004a, b; Overeem et al., 2008; Papalexiou et al., 2013; Papalexiou and Koutsoyiannis, 2013)"* [see **Line 17, Page 5**]. Moreover, we observed more evidence of heavy tailed AMS distributions for SS GEV models ($\xi_* > 0$) with values of ξ_* mostly in the $(0.10, 0.25]$ interval [please, see **Line 8, Page 18**, Fig. 11, and Fig. 9, 3rd col.]. These results are consistent with previous studies, which reported typical values of $\xi \approx 0.15$ for cases in which AMS are long enough to reduce estimation uncertainty and provide non-zero estimates of the GEV shape parameter (e.g., Koutsoyiannis, 2004b). [see **Line 9, Page 18**]. Hence, the shape parameter ξ_* seemed to be better assessed by SS GEV models, since under the SS hypothesis, the statistical information from several durations can be pooled.

Concerning the scale parameter estimation, we observed that the estimation of the scale parameter, σ, may be biased when the ξ is spuriously set to zero [**Line 17, Page 17**], since the σ may tends in this case to increase to fit the distribution to the more extreme events. Hence, large uncertainties in ξ could imply large uncertainties and biases in σ.

In other words, our interpretation of Fig. 9 [Fig. 8 in the previous version] is that SS-hypothesis provides a framework under which the GEV parameters (especially the shape parameter) are more accurately estimated than the non-SS estimates, even if we recognize that strong uncertainties still affect the ξ_* estimation [**Line 14, Page 18**]. Furthermore all our results are consistent with what has been previously reported in the literature.

To better stress these points, several changes have been made in Section 6.1 [**Page** 17]:

- We moved the sentence *"the scale parameter σ_d may be biased when the shape parameter is spuriously set to zero ($\xi_d = 0$)"* from the end of the paragraph to **Line 17, Page 17**.

- **Lines 25 to 1, Page 17** has been rewritten as:

  "Notable differences between SS GEV and non-SS GEV estimates were observed for the shape parameter [Fig. 9, third col., and Fig. 11] Firstly, for cases having shape parameters strictly different from zero [third column of Fig. 9], ξ_* absolute values were smaller than non-SS ξ_d absolute values. Secondly, the distributions of ξ_* across stations were generally more peaked around their median value than the corresponding non-SS distributions. Finally, for the non SS model, the majority of stations had shape parameter ξ_d non-significantly different from zero, while the fraction of SS GEV shape parameters $\xi_* \neq 0$ was always greater than 39% [asymptotic test for PWM GEV estimators applied at level 0.05; Hosking et al., 1985]. "

e) Page 16, lines 22-23: I agree with this sentence, nice spatial coherence.

f) Page 16, lines 25-26: why SS shape parameter is nearer to zero and exhibit less spread? Is this an effect of the assumptions of SS?

Yes, as stated at **Line 7, Page 18**, the fact that ξ_* show less variability than ξ_d is one of the results which 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."*. In other words, the improvement in GEV estimation induced by the SS models allows to 1) decrease the number of $\xi = 0$ cases, and, 2) reduce the dispersion of shape parameters among the station which manifest itself into shape parameter values closer to zero.

g) Page 16, line 27: very difficult sentence to read (essentially the shape parameter is zero).

As reported in reply to comment 16 d), the sentence has been rewritten as [**Line 30, Page 17**]:

"Finally, for the non SS model the majority of stations had shape parameter ξ_d non-significantly different from zero, while the fraction of SS GEV shape parameters $\xi_* \neq 0$ was always greater than 39% [asymptotic test for PWM GEV estimators applied at level 0.05; Hosking et al., 1985]."

h) Page 16, lines 27-31: these results are not intuitive. Figure 10 suggests that the shape parameter for the non-SS estimates is predominantely zero, whereas for the SS estimates there is a large proportions of positive and negative shape parameters. However, the right column of Figure 8 shows exactely the opposite (SS estimate are nearer to zero than non-SS estimates). I assume the difference is due to the very different width of confidence interval associated to the estimate of the shape parameter, for non- SS and SS samples. Are the sample sizes very different (I believe so ... given that x* in equation 8 is obtained from a very large sample). Then maybe these results are artificial ... (as you conclude afterwards, at page 17, lines 3-5: but this link is not explicit)!!!

The reviewer must remind that Fig. 9, 3^{rd} col, excludes cases for which $\xi = 0$ (Gumbel distribution) [see the caption of Fig. 9]. Therefore, Figures 9 and 11 are not inconsistent since a larger fraction of stations have significant non-zero shape parameters for the SS-model. The reviewer might be right when he mentioned that the difference between shape parameter distributions is due to the very different width of confidence interval associated to the estimate for the SS- and non-SS models. Sample sizes are indeed very different since the SS model pools the AMS from all the durations included in each scaling interval. These results are not "artificial" and reflect the fact that, under SS model, the sampling error is reduced. Corrections made due to previous comments should have clarified these points [for instance, see also our reply to comment 16 d)].

d) Page 17, lines 7: eliminate (not clear what this refer too).

To make the text clearer, we modify the sentence to [**Line 11, Page 18**]:

"These studies typically reported values of $\xi \approx 0.15$ (e.g., Koutsoyiannis, 2004b), which are close to ξ_* values estimated in the present analysis for cases with $\xi_* > 0$. "

e) Overall, the text starting at page 16 line 27 and ending at page 17 line 10 should be all reorganized ina more coherent single paragraph.

We agree with the reviewer and corrections made in response to previous comments should have improved the text between **Line 30, Page 17** and **Line 15, Page 18**.

f) Page 17, lines 12-15: Figure S13 shows that for shape parameter equal to 0 (the majority of the stations) the error of quantiles estimated by the non-SS is smaller than that for SS estimates.

Yes, as reported at **Line 3, Page 18**, for many stations (55% to 60% for 6-duration scaling intervals) the SS shape parameter is not-significantly different from zero ($\xi_* = 0$). In these cases, the SS GEV model allows a reduction of the mean error on quantiles only for a small proportion of stations (generally lower than 0.4), as showed in Fig. S14 [Fig. S13 in the previous submission]. Conversely, for cases with non-zero $\xi_* = 0$, the fraction of stations with decreasing errors was higher than 60% for most of the scaling intervals and durations.

To better describe these results **Lines 18 to 23, Page 18** have been modified to:

"For cases with non-zero ξ_*, more than 60$\%$ of stations had $\epsilon_{d,ss} < \epsilon_{d,non-ss}$ over most scaling intervals and durations. The 6-duration scaling intervals "15 min - 1 h 30 min" (SD dataset) and "1 h - 6 sih" (ID dataset) showed the largest fractions of stations with increasing errors. On the contrary, increasing errors ($\epsilon_{d,ss} > \epsilon_{d,non-ss}$) were observed for all scaling intervals and durations for most stations (generally more than 70%) having $\xi_* = 0$. "

Authors' response to 2nd referee's comments

I checked the article and author's responses. I think the authors did a good job to account for both reviews. A few technical points need to be addressed but this is minor.

Minor comments:

1. p 3 l 23 : the fact that X_d^q and $\lambda^{Hq} X_{\lambda d}^q$ have the same distribution comes directly from (2). You don't need for that to have finite moments (please note by the way that exponent 'q' is missing in lambda). However (3) needs finite moments.

   Corrected, thank you. The paragraph now reads [**Lines 26 to 29, Page 3**]:

   "An important consequence of the SS assumption is that X_d and $\lambda^H X_{\lambda d}$ have the same distribution. Hence, if X_d and $X_{\lambda d}$ have finite moments of order q, $E[X_d^q]$ and $E[X_{\lambda d}^q]$, these moments are thus linked by the following relationship ... "

2. p7 l 14 : I don't think the relation $x_{d_2} \leq x_{d_1}$ for $d1 < d2$ is always valid. For example, let consider the hourly series with values 10-2-10 mm/h. Then the maximum 2h-intensity is 12/2=6mm/h, while the maximum 3h-intensity is 22/3>6 mm/h. So $x_{d_2} > x_{d_1}$ for d_2=3h and d_1=2h. Also $x_{d_2}/x_{d_1} < d_1/d_2$.

   We apologize for the error made in the sentence at **Line 20, Page 7** and we agree that relationship $x_{d_2} \leq x_{d_1}$ is not always valid. The numerical algorithm used for data screening and AMS construction considered the following criterion for *precipitation depth [mm]*:

   $$\frac{x_{d_2}}{x_{d_1}} \geq \frac{d_2}{d_1}$$

   When the rainfall depth is used, in fact, the relationship must be respected for each couple of durations $d_1 < d_2$ (for your example we have: $x_{d_1} = 12 < 22 = x_{d_2}$ and $\frac{x_{d_2}}{x_{d_1}} = \frac{22}{11} \geq \frac{3}{2} \frac{d_2}{d_1}$). The confusion arose when converting this relationship in term of rainfall intensity. The error has been corrected by changing **Line 20, Page 7** to:

   "For instance, each pair of DMPD rainfall intensities [mm/h] (x_{d_1}, x_{d_2}) observed at durations $d_1 < d_2$ must respect the condition $x_{d_2}/x_{d_1} \geq d_1/d_2$ derived from the definitions of daily maximummaxima rainfall intensity and depth; "

3. p 7 l 29 : IS → ID ?

   Corrected.

4. p 7 l 30: the first matrix on the left of Fig. 1(a) → the top left matrix of Fig. 1(a)

   Corrected, thank you.

5. p 8 : Does the SS sample $x_{d,ss}$ comprises the non-SS sample x_d ? I guess it should not (for independence testing) but it is not clear to me on (8)

We agree that for GOF tests applied in Sect 4 it would be worth considering SS samples $x_{d,ss}$ that do not contain data from duration d. For this reason, the steps 1-3 of the methodology described at **Page 9** has been repeated in a cross validation settings [see also the replies to comments 11 e), 11 f), and 12 c) of the 1^{st} referee].

However, in the mathematical definition of SS sample, $x_{d,ss}$, [**Line 15, Page 9**] one should include the observations of the rescaled AMS for duration d (i.e., observations from sample sample x_d). This sample, in fact, is not only used for the SS validation presented in Section 4, but also for other analyses which require to include the rescaled x_d observations in $x_{d,ss}$. For instance, for the estimation of the SS GEV models [Sect. 6, **Lines 24 to 26, Page 16**] $x_{d,ss}$ should contain observations from AMS observed at d and rescaled at d^*.

6. p 15 l 25 : obtained 12 → obtained for 12

Corrected.

Authors' response to 3rd referee's comments

Thank you for the opportunity to review this manuscript. I agree with the two reviewers on that the first two sections of the paper are very well written and the rest of the sections need improvement in order to clearly deliver your messages. Although still a little bit difficult to understand (I had to read it a few times. But it could be due to my lack of background.), I can see that the manuscript has improved significantly by addressing the comments from previous reviewers. I recommend the manuscript to be accepted with minor revisions.

1. There are a lot of symbols and abbreviations in this manuscript. Maybe the authors can redefine them when first used in each major section (as well as in figure captions) to help readers to understand the methods and results better. Or adding a glossary table in the supplementary documents may help.

   Following the reviewer suggestion we added Table S1 to the supplementary material; this table lists the relevant and recurrent acronyms. A reference to Table S1 has then been added in the introduction [**Line 17, Page 3**]. Then, additional definitions of some acronyms have also been added to the text [see, for instance, acronym added at **Line 21, Page 3**, AMS definition added at **Line 13, Page 6**, section title of Sect. 6, and reference to Figure 1 a) added at **Line 9, Page 7**].

2. What is the colour scheme in Figure 1? Can you please add a legend of the colour scheme?

   We apologize for the technical problem, the color bar has been added to Fig. 2 [corresponding to Fig. 1 in the previous version of the manuscript].

3. To improve readability, this reviewer recommends the authors to break some long sentences into shorter ones. For example, the sentence starting on Line 11 on Page 12 includes at least three messages, which should be broken down into shorter sentences. There are a number of similar sentences in the results and discussion sections.

   We agree with the reviewer and we rephrase various paragraphs of the manuscript in order to increase the readability. See, for instance, the following paragraphs :

   - Sect. 4: **Line 27, Page 10**, **Line 7, Page 8**, and **Line 34, Page 12** [cited by the reviewer];
   - Sect. 5: **Line 24, Page 15**, **Line 1, Page 14**, **Line 14, Page 14**, and **Line 7, Page 16**;
   - Sect. 6: **Line 18, Page 16**, **Line 22, Page 17**, and **Line 30, Page 17**;
   - Conclusion: **Line 8, Page 20**

4. Section 7 on discussion and conclusion seems to be a summary of what you have written so far. Can you please add some discussion on why this research is important? How can the outcomes be used in practical sense? For example, do they imply any changes to current flood risk estimation/infrastructure design guidelines? What are the limitations of this study and recommended future research?

   Several part of Sect. 7 introduce the mentioned issues. In particular recommended extensions of the research or study limitations were briefly discussed at **Line 6, Page 20**, **Line 29, Page 20**, and **Line 11, Page 20**. However, to make more explicit references to practical implications and limitations of our study, the following modifications have been made:

   - We modified **Lines 6 to 8, Page 20** to :

     "These results suggest that SS represents a reasonable working hypothesis for the development of more accurate IDF curves. This may have important implications for infrastructure design and risk assessment for natural

ecosystems, which would benefit from a more accurate estimation of precipitation return levels. Besides, the spatial distribution of the scaling exponent and its dependency on climatology should be taken into account when defining SS duration intervals for practical estimation of IDF. The accuracy of the SS approximation may in fact depend on the range of considered temporal scales. Equally critical, estimated H values were found to gradually evolve with the considered scaling intervals."

- We modified **Lines 26 to 31, Page 20** to :

  "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. Moreover, a more comprehensive assessment of the scaling exponent uncertain and of the influence of dataset characteristics on the estimation of AMS simple scaling is recommended. Considering these limitations 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. "

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 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* ~~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 focuses 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]. Table S1 of the supplementary material lists in alphabetic order the recurrent acronyms used in text.[R3]

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 sampled ~~considered~~[R1] at two different durations d and λd, the Simple Scaling (SS)[R3] can be expressed as (Gupta and Waymire, 1990; Menabde et al., 1999):

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

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

where $H \in \mathbb{R}$ and $\overset{dist}{=}$ ~~$\overset{d}{=}$~~[R1] means that the same probability distribution applies for X_d and $X_{\lambda d}$, up to a dilatation or contraction of size λ^H. An important consequence of the SS assumption is that, ~~if X_d has finite moments $E[X_d^q]$ of order q, then~~[R2] X_d and $\lambda^H X_{\lambda d}$~~$\lambda^H X_{\lambda d}^q$~~[R2] have the same distribution. Hence, if X_d and $X_{\lambda d}$ have finite moments of order q, $E[X_d^q]$ and $E[X_{\lambda d}^q]$, these ~~Their~~[R2] moments are ~~thus~~ linked by the following relationship (Gupta and Waymire, 1990; Menabde et al., 1999):

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

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:

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

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.

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). 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). If $\xi = 0$ (light-tailed shape, Gumbel distribution), Eq. (5) reduces to:[R1]

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

where $-\infty < x < +\infty$. In Eq. (5) and (6), the parameters [R1] $\mu \in \mathbb{R}$, $\sigma > 0$ and ξ respectively represent the location, scale, and shape parameters of the distribution. The shape parameter describes the characteristics of the distribution tails. 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:~~

~~$F(x) = \exp\{-\exp-\{\frac{x-\mu}{\sigma}\}\}$ (6)~~

~~where $-\infty < x < +\infty$.~~[R1]

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 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 $X \overset{dist}{=}$ ~~$X \overset{d}{=}$~~[R1] $GEV(\mu, \sigma, \xi)$ then $\lambda X \overset{dist}{=}$ ~~$\lambda X \overset{d}{=}$~~[R1] $GEV(\lambda\mu, \lambda\sigma, \xi)$ for any $\lambda \in \mathbb{R}$. This means that the GEV family described by Eq. (5) and (6) satisfies Eq. (1) ~~(2))~~[R1] and thus complies with statistical[R1] scale invariance for any constant multiplicative transformation of X. Hence, when the scale invariance is further assumed for the change of observational duration from d to λd [as in Eq. (2)], ~~Under this assumption~~[R1] the wide sense SS definition [Eq. (3)] gives:

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

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. In this case, 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).

3 Data and study region

Four station datasets were used for the construction of intensity Annual Maxima Series (AMS) ~~AMS~~^{R3} at different durations: the Daily Maxima Precipitation Data (DMPD) and the Hourly Canadian Precipitation Data (HCPD) 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 and HCPD 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 and HCPD 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* from which the annual maxima was sampled was limited to the recording season going from June to September for northern stations [stations located north of the 52^{nd} Parallel] and from June to September for the southern stations. As a result, 122 days a year were used for northern stations and 184 days a year for remaining stations.

~~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].~~^{R1}

Data were collected through a variety of instruments [e.g., standard, tipping-bucket, and Fischer-Porter rain gauges] and precipitation values were processed and quality-controlled ~~checked~~^{R1} 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 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 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 [see Figure 1 (a)][R3].

(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 for each May to October (or June to September) period~~each year~~[R1], otherwise the corresponding[R1] 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 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.

(iii) A moving window was applied to 15PD, HCPD, and HPD series to estimate aggregated series at each duration. For DMPD 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 instance ~~finstance~~[R1], each pair of DMPD intensity (x_{d_1}, x_{d_2}) observed at durations $d_1 < d_2$ must respect the condition $x_{d_2}/x_{d_1} \geq d_1/d_2$ ~~the conditions $x_{d_2}/x_{d_1} \leq 1$ and $d_1 x_{d_1} \leq d_2 x_{d_2}$~~[R1] derived from the definitions of daily maximum~~maxima~~ rainfall intensity and depth; otherwise all DMPD values recorded that day were discarded and assimilated to missing data.

(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.

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 modeling AMS empirical distributions ~~for intensity AMS~~[R1]. 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.

In order to identify possible changes in the SS properties of AMS distributions, various *scaling intervals* ~~scaling intervals~~[R1] were defined for the MSA. In particular, all possible subsets with 6, 12, 18 and 24 contiguous durations were considered within each dataset. Figure 2 and Figure 3 show the 136 scaling intervals thereby defined: 40 scaling intervals for SD and ID ~~IS~~[R2], and 56 scaling intervals for LD. For instance, the top left matrix ~~the first matrix on the left~~[R2] of Fig. 2(a) presents the 6-duration scaling intervals 15 min - 1 h 30 min, 30 min - 1 h 45 min, … 4 h 45 min - 6 h~~15min - 1.5h, 30min- 1.75h, …, 4.75 h - 6 h~~[R1] defined for the SD dataset [i.e. the 19 scaling intervals containing six contiguous durations defined with a 15min increment] .~~,~~[R3] More schematically, Fig. 1(b) shows an example of the first five 6-duration scaling intervals for the ID dataset [i.e. 1h - 6h, 2h - 7h, …, 5h - 10h, containing six contiguous durations defined with an increment of 1h]. ~~while Fig. 2(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].

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 each[R1] $q = 0.2, 0.4, \ldots, 2.8, 3$, 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 $[d_1, d_D]$ were estimated by Ordinary Least Squares (OLS) [see Fig. 1 (c) for a graphic example][R1]. Order $q \geq 3$ were not considered because of the possible biases affecting empirical high order moment estimates.

2. *Slope test:* to verify the SS assumption that the estimated K_q exponents vary linearly with the moment order q, i.e. $K_q \approx Hq$, an OLS regression between the MSA slopes K_q and q was applied [see Fig. 1 (d)]. For the regression line $K_q = \hat{h}_0 + \hat{h}_1\ q$, a Student's t-test was then used to test the null hypothesis $\boldsymbol{H}_0$: $\hat{h}_1 = K_1$. ~~Regressing the MSA slopes 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 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 $K_q = \beta_0 + \beta_1 q$.~~[R1] If $\boldsymbol{H}_0$ was not rejected at the significance level $\alpha = 0.05$, the SS assumption was considered appropriate for the scaling interval 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, each AMS, $\boldsymbol{x}_{d_j} = \left(x_{d_j,1}, x_{d_j,2}, \ldots, x_{d_j,i}, \ldots x_{d_j,n}\right)$, recorded at duration d_j was rescaled at the reference duration d^* by inverting Eq. (2):[R1]

$$\boldsymbol{x^*}_{d_j} = \left({d_j}^H x_{d_j,1}, {d_j}^H x_{d_j,2}, \ldots, {d_j}^H x_{d_j,i}, \ldots {d_j}^H x_{d_j,n}\right) \qquad (8)$$

where n represents the number of observations (years) in $\boldsymbol{x}_{d_j}$. Then, the pooled sample, $\boldsymbol{x}_{d^*}$, of the D rescaled AMS, $\boldsymbol{x}^*{}_{d_j}$, was used to define X_{d^*} under the SS assumption:[R1]

$$\boldsymbol{x}_{d^*} = \left(\boldsymbol{x}^*{}_{d_1}, \dots, \boldsymbol{x}^*{}_{d_j}, \; \dots \; , \boldsymbol{x}^*{}_{d_D}\right) \tag{9}$$

Since, in Eq. (9), D represents the number of durations d_j in the scaling interval, $n \times D$ rescaled observations were included in $\boldsymbol{x}_{d^*}$

~~To this end, the pooled sample~~

~~$\boldsymbol{x}_{d^*} = \left(\boldsymbol{x}'_{d_1}, \dots, \boldsymbol{x}'_{d_j}, \; \dots \; , \boldsymbol{x}'_{d_D}\right)$ (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 $j = 1, \dots, D$, in the scaling interval. Each rescaled sample $\boldsymbol{x}'_{d_j}$ of the annual maxima $x_{d_j,i}$, $i = 1, \dots, n$, observed for the duration d_j was obtained by simply inverting Eq. (2) for d^*:~~

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

~~where n represents the number of observations (years) available for each duration. In this way, $n \times D$ rescaled observations were included in $\boldsymbol{x}_{d^*}$~~[R1].

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, $\boldsymbol{x}_d$. 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 the SS and non-SS samples. Then, the test statistic distribution under the null hypothesis of equality of the $X_{d,ss}$ 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 SS model validity and the mean error resulting from approximating the X_d distribution by the SS model were then evaluated in a cross-validation setting. For this analysis, each duration was iteratively excluded from each scaling interval and the scaling model re-estimated at each station by repeating steps 1 to 3 [MSA regression, Slope test, and GOF tests]. ~~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.~~[R1] Predictive ability indices, such as the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) between empirical and SS distribution quantiles, were then estimated for highest quantiles for valid SS stations.. 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).

For each station s, the normalized RMSE, $\overline{\epsilon}_{x_{d,s}}$, was estimated: [R1]

$$\overline{\epsilon}_{x_{d,s}} = \frac{\epsilon_{x_{d,s}}}{\overline{x}_{d,s}} \tag{10}$$

where $\epsilon_{x_{d,s}}$ and $\overline{x}_{d,s}$ are, respectively, the RMSE and the mean value of all X_d quantiles of order $p > 0.5$. Then, the average over all stations of the normalized RMSE, $\overline{\overline{\epsilon}}_{x_d}$, was computed for each scaling interval and duration: [R1]

$$\overline{\overline{\epsilon}}_{x_d} = \frac{1}{n_s}\sum_{s=1}^{n_s}\overline{\epsilon}_{x_{d,s}} \tag{11}$$

where n_s is the number of valid SS stations in the dataset. Note that $\overline{\overline{\epsilon}}_{x_d}$ 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.

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

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

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

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

~~and $\epsilon_{x_{d,s}}$ 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 2 presents the results of steps 1 to 3 of the methodology for evaluating the SS validity~~of points 1 to 3 of the procedure for the evaluation of the SS validity.~~[R1] For all the three scaling datasets, no particular pattern was observed for slope test results, and at most 2% of the stations within each scaling interval displaying a non linear evolution of the scaling exponent with the moment order. For this reason, Fig. 2(a)-(c) show, for each scaling interval and duration, the proportion of valid SS stations without differentiating for slope or GOF test results.~~It shows, for each scaling interval and duration, the proportion of valid SS stations [Fig. 2(a)-(c)].~~[R1] As showed in Fig. 1(e)~~the example in Fig. 2(e)~~[R1], 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]~~[R1], due to the 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. However, several factors can impact GOF test results when shorter d_1 are considered. First, GOF tests are particularly sensitive to the presence of very large values in short-duration samples~~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~~[R3]. 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 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).

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.~~^{R1} 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.

These findings were also confirmed by cross-validation experiments. The proportion of valid SS stations resulting from cross-validation Slope and GOF tests were similar, event if slightly lower, to proportions displayed in Fig. 2 [see Fig. S4 of the supplementary material].^{R2}

Figure 3 presents, for each scaling interval and duration, the station average, $\overline{\epsilon}_{x_d}$, 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. 3, first col.]. Greater 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 for durations at the boundaries or outside the scaling interval ~~if d is outside the range of durations~~^{R1} used to estimate H. Moreover, 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^{R1}. However, for more that 70% of 12-, 18-, and 24-duration scaling intervals, $\overline{\overline{\epsilon}}_{x_d} \leq 0.1$ for each duration included in the scaling interval. $\overline{\overline{\epsilon}}_{x_d} \geq 0.25$ 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 $\geq$ 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.

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

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

where $j = 12, 18$, or 24 represents 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 scaling 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.

Figures 4(ii)-(iv) show for all relevant scaling intervals the median, Interquantile Range (IQR), and quantiles of order 0.1 and 0.9 of the $\Delta_{H_{(j)}}$ distribution over valid SS stations.~~The median, Interquantile Range (IQR), and quantiles of order 0.1 and 0.9 of the H distribution across stations, are presented in Fig. 4(i) for each 6-duration scaling interval. Figures 4(ii)-(iv) show the distribution over valid SS stations of $\Delta_{H_{(j)}}$ for all relevant scaling intervals.~~[R1] Adding new durations to the scaling intervals, the median $\Delta_{H_{(j)}}$, as well as its IQR, increased for all d_1.~~Median $\Delta_{H_{(j)}}$, as well as its IQR, increased with the number of durations added to the scaling interval for all d_1.~~ 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. 4 (ii)-(iii)]. For some stations, a dramatic difference could exist in IDF estimations obtained with the different definitions of the scaling interval. [R1] For instance, for the 24-duration scaling interval "1h - 24h" (ID dataset), the median $\Delta_{H_{(24)}}$ was equal to 0.047.~~for instance,median $\Delta_{H_{(24)}} = 0.047$ was observed.~~ [Fig. 4(iv) b)][R1] 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 $\Delta_{H_{(24)}} \geq 0.11$ [Fig. 4(iv) a)][R1]. Finally, 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. 4 (ii)].[R1] ~~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. 4 (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. 4 (ii)-(iii)].~~[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. 4(i) for each 6-duration scaling interval.[R1] The smallest median H values were observed for the shortest $d_1 \leq 30$ min ~~d_1 ($d_1 \leq 30$ min)~~ in Fig. 4 (a-i), and for the longest d_1s in Fig. 4 (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. 4 (i)]. This result could be partially explained by the use of scaling intervals having equally spaced durations. This implies that the mean distance between the logarithms of durations in the scaling interval decreases as d_1 increases. Hence, the OLS estimator of H used in the MSA regression may have larger variance for longer d_1, especially when scaling intervals include few durations. Larger uncertainty may thus have an impact on the H estimation for the longest d_1 scaling intervals of SD. However, as showed in next sections, H spatial distribution may also explain the greater variability of the scaling exponent for d_1 greater than a few hours. [R1] ~~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.~~ [R1]

Largest median H were observed for d_1 greater than 10 hours [Fig. 4 (b-i)] and lower than 2 days [Fig. 4 (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. 5 and 6 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. 5) or 18 (Fig. 6) 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. 6. In this last figure, the local variability of H may be attributed to the larger estimation uncertainties affecting longer d_1 scaling intervals, 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. 5]. 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. S6~~S5~~[R1] and S7~~S6~~[R1] for 12- and 18-duration scaling intervals]. Stations were pooled into six climatic regions based on the~~a previous~~ classification suggested by Bukovsky (2012) [see Fig. 7]. 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. 7) ; regions A1 (W_Tun) and A2 (NW_Pac) were kept separated since ~~and region A1 (W_Tun) was kept separated from region A2 (NW_Pac) because~~R3 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. 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 $\bar{N}_{eve}$, $\bar{T}_{wet}$, and the corresponding results are presented in the supplementary material [Sect. S2 and Fig. S5 and S6 ~~S4 and S5~~R1].

5.1 Regional variation of the scaling exponents.

Figure 8 shows the distribution of H within each region. Three types of curves can be identified. First, curves in Fig. 8 (a) to (c) have a characteristic smooth S shape. Conversely, Fig. 8 (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 plateau~~, one~~R3 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. 8 (e) and (f), with globally high H values already reached at sub-daily durations in dry regions (E).

For $d_1 \leq 24$ h, Fig. 8 (a) displays lower values of H than Fig. 8 (e)-(f), meaning that smaller variation in AMS moments are observed in A1 and A2 when the scale is changed. This difference ~~The difference between Fig. 8 (a) and (e)-(f)~~R1 can be partially explained by the weaker impact of convection processes in generating very short duration extremes in North-West coastal regions ~~regions A1 and A2~~ with respect to southern areas (regions E and F). For northern regions, in fact,R1 the transition between short and long duration precipitation regimes may be smoothed out by cold temperatures which moderate short-duration convective activity, especially for W_Tun (region A1). The topography characterizing the northern pacific coast may thenR1 explain the smoothing effect for the curve of region NW_Pac (A2). In this case, in fact, the~~:~~R1 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. 8 (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. S5 ~~S4~~R1 (a) and (c) in the supplementary material].

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. S5~~S4~~R1 (b) in the supplementary material].

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. 8 (d)]. These plateaus may be interpreted by recalling that $1 - H = H_{depth}$. 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. S5 and S6~~S4 and S5~~R1 of the supplementary material]. 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.

Median H values displayed inverse-U shapes for the remaining regions with very small IQR, 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. 8 (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 [8 (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, H shows a smoother increase in Fig. 7 (f) with respect to Fig. 7(e). This may indicate that in eastern areas [region F] sub-daily duration extremes are more likely associated to embedded convective and stratiform systems, or to mesoscale convective systems, which are less active in western dry areas of region E~~However, the smoother increase of H visible in Fig. 8 (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~~R1 (Kunkel et al., 2012). On the contrary, differences between short- and long-duration extreme precipitation intensity seem stronger for south-western dry regions [Fig. 8 (e)], where less intense summer extremes are expected compared to eastern areas [see supplementary material, Fig. S1]. In particular, H tended to scatter in a range of higher values for approximately 1 h $\leq d_1 \leq$ 12 h indicating that precipitation intensity moments strongly decrease as the duration increases. ~~On the contrary, for south-western dry regions [Fig. 8 (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~~R1.

In summary, these results suggest that both local geographical characteristics, such as topography or coastal effects, and general circulation patterns may influence precipitation scaling at a regional scale. ~~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.~~R1 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. Low H values correspond in fact to small variations in AMS distribution moments. On the contrary stronger scaling regimes were observed for longer d_1 in the other regions of the study area. This indicates that important changes occur in AMS moments across duration and, thus, in extreme precipitation features.~~, 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.~~R3 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 give useful guidelines for modeling the spatial distribution of H, which could help for the definition of IDF relationships at non-sampled locations.~~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~~R1.

6 Simple Scaling GEV etimation

~~6 SS GEV estimation~~R3

Results presented in this section are limited to a descriptive analysis of GEV parameter estimates for 6-duration scaling intervals. Similar results were generally obtained for 12-, 18-, and 24-duration intervals [see supplementary material, Fig. S10 to S16]. An assessment of the potential improvements carried out by Simple Scaling GEV (SS GEV) models with respect to non-SS GEV models is also presented.~~ ,and to an assessment of the potential improvements carried out by SS GEV (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]~~R3.

In our study, the Probability Weighted Moment (PWM)~~ PWM~~R3 procedure was applied to estimate SS-GEV parameters μ_*, σ_*, and ξ_* [Eq. (7)] from $\boldsymbol{x}_{d^*}$ [Eq. (9)]. 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 $\boldsymbol{x}_d$. Preliminary comparisons of various estimation methods [PWM, classical ML estimators, and GAM-ML; see Sect.2.2], 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:

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

where

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

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. See Eq. 10 for the definition of $\epsilon_{d,mod}$ for each station.^{R1}

6.1 Estimated SS GEV parameters

Figure 9 presents the distributions over valid SS stations of the SS GEV parameters rescaled at $d_* = 1$ h [Fig. 9 (a) and (b)] and $d_* = 24$ h [Fig. 9 (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 9, first row]. Similarly, for 6 h $\leq d_1 \leq$ 2 days in the LD dataset, the SS location and scale parameter distributions are in relatively close agreement with the corresponding non-SS parameter distributions. Conversely, for the ID dataset, ~~Conversely, in the ID and LD datasets,~~^{R1} both μ_* and σ_* distributions are~~were~~ more positively skewed than the corresponding non-SS distributions. Finally, for $d_1 \geq 2$ days in the LD dataset, μ_* and σ_* had distributions shifted toward lower values than μ_{24h} and σ_{24h}.^{R1} 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 S11 and S12~~S10 and S11~~^{R1} of the supplementary material]. On the one hand, median values of Δ_μ and Δ_σ were generally smaller than ±5% and ±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 the estimation of the scale parameter σ of a GEV distribution may be biased when the shape parameter is spuriously set to zero ($\xi = 0$). Hence, ^{R1} while non-SS μ_d values can be~~are generally~~ considered to be accurate estimates of the X_d location parameter, small uncertainties should be~~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 ($\xi_d = 0$).~~^{R1} In addition, μ_* and σ_* displayed a^{R3} strong spatial^{R3} coherence . Their~~in their~~^{R3} spatial distributions ~~, which~~^{R3} were characterized by an obvious North-West to South-East gradient [Fig. 10 shows examples for the scaling intervals 15min - 1.5h, 1h - 6h, and 6h - 36h].

Notable differences between SS GEV and non-SS GEV estimates were observed for the shape parameter [Fig. 9, third col., and Fig. 11]. Firstly, for cases having shape parameters strictly different from zero [third column of Fig. 9], ξ_* absolute values were smaller than non-SS ξ_d absolute values. Secondly, the distributions of ξ_* across stations were generally more peaked around their median value than the corresponding non-SS distributions. ~~[third column of Fig. 9]. 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.~~^{R1} Finally, for the non SS model the majority of stations had shape parameter ξ_d non-significantly different from zero, while the fraction of SS GEV shape parameters $\xi_* \neq 0$ was always greater than 39% [asymptotic test for PWM GEV estimators applied at level 0.05; Hosking et al., 1985]. ~~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.~~$^{\mathrm{R1}}$ 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. 11, 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. 11, 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. S10~~S9~~$^{\mathrm{R1}}$].

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). These studies typically reported values of $\xi \approx 0.15$ (e.g., Koutsoyiannis, 2004b), which are close to ξ_* values estimated in the present analysis for cases with $\xi_* > 0$. ~~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).~~$^{\mathrm{R1}}$

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. 10, third col.].

6.2 Improvement with respect to Non-SS models

The proportion of series for which the SS model RMSE, $\epsilon_{d,ss}$, was smaller than the non-SS GEV RMSE, $\epsilon_{d,non-ss}$, was analyzed [see the supplementary material, Fig. S14~~S12~~$^{\mathrm{R1}}$]. For cases with non-zero ξ_*, more than 60% of stations have $\epsilon_{d,ss} < \epsilon_{d,non-ss}$ over most scaling intervals and durations. the fraction of stations with $\epsilon_{d,ss} < \epsilon_{d,non-ss}$ was higher than 60% for most of the scaling intervals and durations. The 6-duration scaling intervals "15 min - 1 h 30 min" (SD dataset) and "1 h - 6 sih" (ID dataset) showed the largest fractions of stations with increasing errors. $^{\mathrm{R1}}$ On the contrary, increasing errors ($\epsilon_{d,ss} > \epsilon_{d,non-ss}$) were observed for all scaling intervals and durations for most stations (generally more than 70%) having $\xi_* = 0$.~~$\epsilon_{d,ss} > \epsilon_{d,non-ss}$ was observed for the majority of stations (generally more than 70%) with $\xi_* = 0$.~~$^{\mathrm{R1}}$

Figure 12 presents the $R_{\overline{rmse}}$ distribution over valid SS stations. When the SS ~~distribution~~ shape parameters were not significantly different from zero [Fig. 12, second col.], the relative increases in total RMSE were usually smaller than 0.1 in SD dataset and only scaling intervals with $d_1 < 1$ h had greater $R_{\overline{rmse}}$ ~~, with only scaling intervals with $d_1 < 1$ h having greater $R_{\overline{rmse}}$~~$^{\mathrm{R3}}$. 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. 12, 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. S14 and S15~~S13 and S14~~$^{\mathrm{R1}}$ 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. S16~~S15~~R1 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 an efficient~~the~~R3 production of IDF curves.

The validity of SS models was empirically confirmed for the majority of the scaling intervals. In particular, based on the comparison of SS distributions to empirical quantiles, 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 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.

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. 8]. 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. This may have important implications for infrastructure design and risk assessment for natural ecosystems, which would benefit from a more accurate estimation of precipitation return levels^{R3}. Besides, the spatial distribution of the scaling exponent and its dependency on climatology should be taken into account when defining SS duration intervals for practical estimation of IDF. The~~since the~~^{R3} accuracy of the SS approximation may in fact^{R3} 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. Moreover, a more comprehensive assessment of the scaling exponent uncertain and of the influence of dataset characteristics on the estimation of AMS simple scaling is recommended. ^{R3} Considering these~~this~~^{R3} limitations^{R3} 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 15-Min 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]. Houly Canadian Precipitation Data (HCPD) and Maximum Daily Precipitation Data (DMPC) for Canada were acquired from Environment and Climate Change Canada (ECCC) and from the MDDELCC of Québec [data available upon request by contacting *Info-Climat@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, 2014RG000 464, 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[a] (DMPC)	Canada	370	1964-2007	1, 2, 6, 12 h	0.1 (82.25%)
Hourly Canadian Prec. Data (HCPD)	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 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.

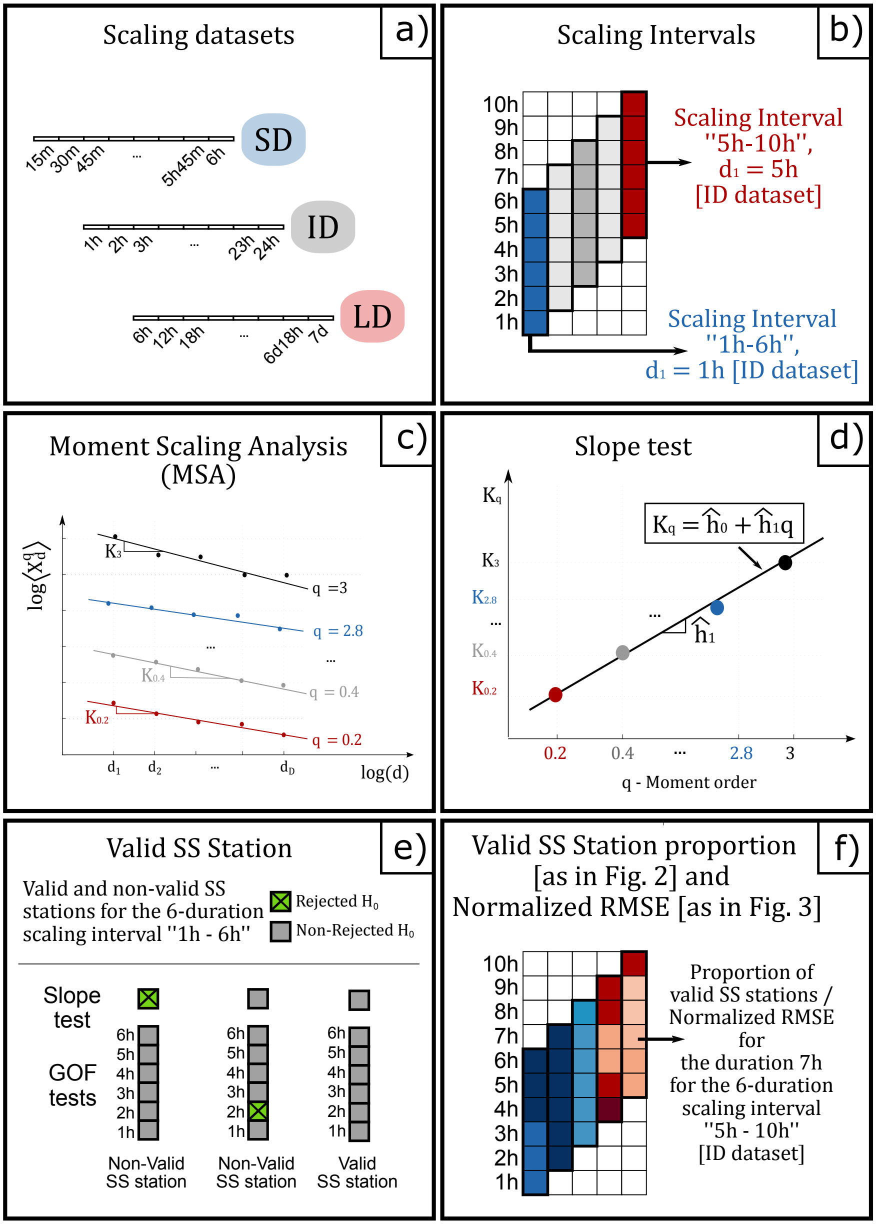
Scaling datasets
a)
15m
30m
45m
...
5h45m
6h
SD
1h
2h
3h
...
23h
24h
ID
6h
12h
18h
...
6d18h
7d
LD
Scaling Intervals
b)
10h
9h
8h
7h
6h
5h
4h
3h
2h
1h
Scaling Interval "5h-10h", d₁ = 5h [ID dataset]
Scaling Interval "1h-6h", d₁ = 1h [ID dataset]
Moment Scaling Analysis (MSA)
c)
$\log\langle X_d^q \rangle$
K_3
q =3
q = 2.8
...
...
$K_{0.4}$
q = 0.4
$K_{0.2}$
q = 0.2
d_1
d_2
...
d_D
log(d)
Slope test
d)
K_q
$K_q = \hat{h}_0 + \hat{h}_1 q$
K_3
$K_{2.8}$
...
...
$\hat{h}_1$
$K_{0.4}$
$K_{0.2}$
0.2
0.4
...
2.8
3
q - Moment order
Valid SS Station
e)
Valid and non-valid SS stations for the 6-duration scaling interval "1h - 6h"
☒ Rejected H_0
■ Non-Rejected H_0
Slope test
GOF tests
6h
5h
4h
3h
2h
1h
Non-Valid SS station
Non-Valid SS station
Valid SS station
Valid SS Station proportion [as in Fig. 2] and Normalized RMSE [as in Fig. 3]
f)
10h
9h
8h
7h
6h
5h
4h
3h
2h
1h
Proportion of valid SS stations / Normalized RMSE for the duration 7h for the 6-duration scaling interval "5h - 10h" [ID dataset]

Figure 1. Methodology steps: a) Definition of the SD, ID and LD scaling datasets. b) Identification of durations and scaling intervals within each matrix of Fig. 2 and 3; c) Moment Scaling Analysis (MSA) regression for the estimation of the slope coefficients K_q; d) Slope test: regression of K_q on the moment order q and Student's t-test for the null hypothesis $\boldsymbol{H}_0$: $\hat{h}_1 = K_1$; e) Examples of valid and non-valid SS stations according to the Slope and GOF tests; f) Example of valid SS station proportion values and Normalized RMSE values, $\overline{\overline{r}}_{x_d}$, as represented, in Fig. 2 and 3.

Figure 2. ~~a) - c)~~$^{\mathrm{R1}}$ 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. See Fig. 1 (b) and (f) for the identification of durations and scaling intervals within each matrix.~~ Example of valid SS station proportion values and identification of durations and scaling intervals within each matrix; c) Examples of valid and non-valid SS stations.~~$^{\mathrm{R1}}$

Figure 3. Cross-Validation Normalized RMSE averaged over all valid SS stations ($\overline{\overline{r}}_{x_d}$) for each duration (vertical axis) and scaling interval (horizontal axis) in the SD, ID, and LD datasets [row a), b), and c) respectively]. White circles indicate values between 0.15 and 0.3. See Fig. 1 (d) and (f) for the identification of durations and scaling intervals within each matrix.$^{\mathrm{R1}}$

Figure 4. 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 5. 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. 2 and 3.

Figure 6. 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. 2 and 3.

Figure 7. 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 8. Median and Interquantile Range (IQR) of the scaling exponent distribution over valid SS stations within each region of Fig. 7 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. 7 for region definition.

Figure 9. Distribution over valid SS stations of SS GEV parameters (gray and black lines) for 6-duration scaling intervals and non-SS GEV parameters (red solid and dashed lines) ~~parameters~~ for reference durations. ~~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 10. 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: SD (a), ID (b), and LD (c) datasets.

Figure 11. 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 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 12. 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.