Les Cahiers du GERAD ISSN: 0711–2440

Statistical testing of scaling models for precipitation Intensity-Duration-Frequency curves

A. Paoli, J. Carreau, and J. Jalbert

G–2024–34

May 2024

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2024
– Bibliothèque et Archives Canada, 2024

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2024
– Library and Archives Canada, 2024

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Statistical testing of scaling models for precipitation Intensity-Duration-Frequency curves

Auguste Paoli

Julie Carreau

Jonathan Jalbert

GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal, (Qc), Canada, H3T 1J4

auguste.paoli@polymtl.ca
julie.carreau@polymtl.ca
jonathan.jalbert@polymtl.ca

May 2024
Les Cahiers du GERAD
G–2024–34

Abstract : Producing accurate precipitation Intensity-Duration-Frequency (IDF) curves necessitates robust statistical methodologies. Employing a scaling model to combine information across the various precipitation accumulation durations is desirable for reducing uncertainty and facilitating interpolation to durations not observed. A variety of such scaling models exist, yet there is currently no formal goodness-of-fit testing procedure for selecting an appropriate one. In this paper, we develop a goodness-of-fit procedure to determine if a scaling model is suitable for precipitation IDF data. The proposed test extends the Anderson-Darling test and involves dividing the database into training and validation sets. The training set is used to estimate the parameters of the target model, while the validation set is utilized to compute the test statistic. The asymptotic distribution of the test statistic is established within a general framework, enabling analytical calculation of the critical region for the test. In our application to precipitation IDF curves, data corresponding to the shortest accumulation duration are chosen for the validation set. We validate the performances of the test through a simulation study, demonstrating that under the null hypothesis, the test maintains the nominal rejection rate even for small samples ranging from 5 to 20 years. Under an alternative hypothesis, the rejection rate increases with the discrepancy between the models as well as with the sample size. When applied to historical data, the test suggests the use of different scaling models for Montréal (QC) and Vancouver (BC).

Keywords: Precipitation, scaling models, intensity-duration-frequency curves, extremes, goodness-of-fit test, Anderson-Darling test

Résumé : Produire des courbes d'Intensité-Durée-Fréquence (IDF) précises pour les précipitations nécessite des méthodologies statistiques robustes. Il est souhaitable d'employer un modèle d'échelle pour combiner les informations des différentes durées d'accumulation des précipitations afin de réduire l'incertitude et de faciliter l'interpolation pour des durées non observées. Divers modèles d'échelle existent, mais il n'existe actuellement aucune procédure formelle de test d'adéquation pour sélectionner le modèle approprié. Dans cet article, nous développons une procédure de test d'adéquation pour déterminer si un modèle d'échelle est adapté aux données IDF des précipitations. Le test proposé étend le test d'Anderson-Darling et repose sur la séparation de la base de données en ensembles d'entraînement et de validation. L'ensemble d'entraînement est utilisé pour estimer les paramètres du modèle cible, tandis que l'ensemble de validation est utilisé pour calculer la statistique de test. La distribution asymptotique de la statistique de test est établie dans un cadre général, permettant le calcul analytique de la région critique pour le test. Dans notre application aux courbes IDF des précipitations, les données correspondant à la plus courte durée d'accumulation sont choisies pour l'ensemble de validation. Nous validons les performances du test par une étude de simulation démontrant que, sous l'hypothèse nulle, le test maintient le taux de rejet nominal même pour de petits échantillons allant de 5 à 20 ans. Sous une hypothèse alternative, le taux de rejet augmente avec la divergence entre les modèles ainsi qu'avec la taille de l'échantillon. Lorsqu'appliqué aux données historiques, le test suggère l'utilisation de différents modèles d'échelle pour Montréal (QC) et Vancouver (CB).

Mots clés : Precipitations, modèles d'échelle, courbes intensité-durée-fréquence, événements extrêmes, test d'adéquation, test d'Anderson-Darling

Acknowledgements: This work was supported by Natural Sciences and Engineering Research Council of Canada and le Fonds de Recherche du Québec – Nature et technologies.

1 Introduction

Extreme rainfall events can have catastrophic repercussions, as underscored by a recent incident in Libya. On September 10, 2023, the city of Derna suffered extensive devastation due to the failure of a dam precipitated by a severe storm Petley (2023). It has been postulated that the precipitation, measuring 200 mm, may have surpassed the intended capacity of the dam. Having reliable estimates of the magnitude and frequency of such extreme events is crucial for designing appropriate hydrological structures, such as dams, reservoirs, and sewer systems.

In hydrology and engineering, estimates of the precipitation intensity $i(d, T)$ in mm/h during the duration $d > 0$ that is expected to be exceeded once every T years are given by Intensity-Duration-Frequency (IDF) curves (see e.g. Koutsoyiannis et al., 1998, for a review). In Canada, IDF curves are estimated and provided by Environment and Climate Change Canada (ECCC). Figure 1 displays the IDF curves for the meteorological station located at the Pierre-Elliott-Trudeau International Airport in Montréal (QC). For each of the nine durations indicated on the x-axis, six crosses represent the estimated quantiles of the rain intensity distribution corresponding to return periods of 2, 5, 10, 25, 50, and 100 years, respectively. For example, this chart indicates that for a duration of 30 minutes, a rainfall intensity of 84 mm/h is expected to occur on average once every 100 years. In other words, rainfall accumulation during 30 minutes is expected to exceed 42 mm once in a hundred years.

Among existing methodologies, the procedure for estimating the IDF curves provided by ECCC is as follows. For a given duration $d \in \mathcal{D}$, where $\mathcal{D}$ corresponds to the set of 5-minute, 10-minute, 15-minute, 30-minute, 1-hour, 2-hour, 6-hour, 12-hour, and 24-hour durations, the return levels are estimated using the Gumbel distribution fitted to the annual maxima of precipitation intensity, independently for each duration. As interpolation at intermediate durations is often necessary in practice, Figure 1 displays six regression lines, each based on the nine return level estimations for a given return period. This modeling choice assumes that the relation between precipitation intensity and duration follows a power law (see e.g. Menabde et al., 1999, for a review). This leads to the following assumption:

$$i(T, d) = a(T) \times d^{b(T)}; \tag{A1}$$

where b represents the slope of the line on the log scale, and a measures the intensity of a rain event of unit duration. The regression parameters a and b are estimated independently for each return period.

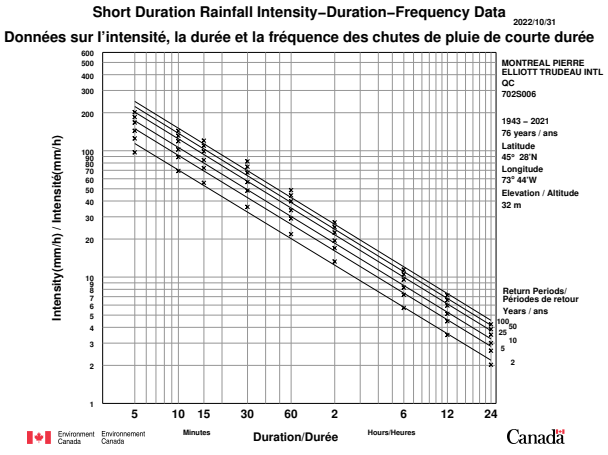


Figure 1: IDF curves for the Pierre-Elliott Trudeau International Airport. Source: ECCC.

The IDF curves estimation method used by ECCC has several drawbacks. Firstly, the use of the Gumbel distribution to model the series of annual maximum intensities is inconsistent with the extreme

value theory (see Coles, 2001, for a review). This is an important limitation, especially considering that precipitation maxima typically exhibit heavy-tailed behavior (see e.g. Jalbert et al., 2022, for precipitation in Canada). Secondly, since the power law exponent parameter $b(T)$ depends on the return period and is estimated independently for each, the regression lines for different return periods across durations may intersect each other, which is physically impossible. Finally, the functional dependence between the precipitation return levels across durations, as expressed in Assumption A1, is not taken into account in the estimation procedure. Incorporating this functional dependence in the estimation procedure enables the sharing of information on precipitation extremes between durations. It enables reducing parameter uncertainties and interpolating to intermediate durations.

A natural idea consists in transforming Assumption A1 into a so-called "scaling" model between the distributions of precipitation intensity maxima across durations (see e.g., Gupta & Waymire, 1990). This links the distributions of precipitation maxima across durations with a single parsimonious model, pooling information on extreme values across durations and avoiding potential crossover between return level curves. This scaling property can be expressed as a one-parameter relationship over the set of marginal distributions. Blanchet et al. (2016) and many others have associated this property with extreme value theory in order to compute IDF curves with a reduced number of parameters, hence resulting in narrower confidence intervals. Koutsoyiannis et al. (1998) generalized Assumption A1 into a 2-parameter relationship over the set of marginal distributions. When it comes to IDF curves, that second parameter allows concavity to appear in log-log scale for smaller durations. Lima et al. (2018), amongst others, used that flexible scaling property in order to build a regional framework for computing IDF curves.

Even though those scaling models have strong physical justifications for precipitation across durations, they are not theoretically proven to hold in every situation. In some practical cases, data shows evidence of diverging from the scaling assumption. Many authors have proposed other empirical alternatives to simple scaling. Bougadis & Adamowski (2006) highlighted the existence of two scaling regimes, one for short durations and another for long durations. Courty et al. (2019) proposed applying modifications directly to the scaling properties of the extreme value parameters. Haruna et al. (2023) combined those approaches to build a "data-driven" scaling model showing high flexibility.

The choice of which scaling IDF model to use is often made rather arbitrarily in the literature. In many cases, the simple scaling model is preferred as a default, as it is the most parsimonious and easy to implement, and its validity can be verified by graphical methods. Menabde et al. (1999) observed that, for two locations in Australia and South Africa, simple scaling seemed to hold for any rain duration larger than 30 minutes. As the climate is very different between those locations, they suggested that this was a general behavior. However, when Yeo et al. (2021) reproduced their methodology for data from the Pierre-Elliott-Trudeau International Airport (Canada) and Seoul (South Korea), they showed that the simple scaling property did not hold for the same duration intervals in both places. Furthermore, such geographical discrepancies may occur between nearby locations, as observed by Rodriguez-Sola et al. (2017) for a set of Spanish weather stations.

Thus, the decision of whether or not to use a given scaling property to build IDF curves must be made on a case-by-case basis. From a practical standpoint, this suggests the necessity of a procedure to assess the validity of the scaling assumption. Haruna et al. (2023) proposed a cross-validation scheme to choose the best IDF model at a given location; however, that procedure seems too cumbersome to be practical for engineering purposes and is not applicable for small samples (i.e., weather stations with few years of observations). Innocenti et al. (2017) established a testing procedure for simple scaling as part of a study over the whole set of North American weather stations. Yet, the discriminating power of the test wasn't studied, and preliminary analysis suggested that it wouldn't reject often enough. There exists no formal goodness-of-fit test to assess the scaling assumption.

The main objective of this paper is to propose a goodness-of-fit test procedure tailored for scaling models. Specifically, we first adapt the Anderson-Darling test to utilize a training set for estimating distribution parameters and a validation set for evaluating fit. We then apply this test to precipitation

IDF data, with the smallest duration data composing the validation set and the remaining durations forming the training set. The newly proposed Anderson-Darling family of goodness-of-fit tests is highly versatile and applicable in various scenarios involving data partitioning into a training and validation sets. In this paper, we demonstrate its application to precipitation scaling models.

The remainder of the paper is as follows: Section 2 describes the different scaling models for precipitation across durations. Section 3 reviews the existing Anderson-Darling goodness-of-fit tests for a known target distribution and for a target parametric family. Section 4 introduces the proposed Anderson-Darling test procedure using training and validation sets, applied in Section 5 for assessing the adequacy between a target scaling model and precipitation IDF data. The Type I error as well as the power of the proposed test procedure are studied with a simulation study in Section 6. The proposed test is applied to real precipitation data for estimating IDF curves in Section 7. Section 8 provides the general discussion and conclusion. The proposed test is provided in the Julia open-source package IDFCurves.jl, and the code reproducing the paper results and figures is available on this public repository: `https://github.com/jojal5/Publications`.

2 Scaling models for IDF curves

The scaling models presented in this section will be applied to the precipitation data recorded at the Pierre-Elliott-Trudeau International Airport (QC) for illustrative purposes. For this station, annual precipitation intensity maxima are available for $n = 76$ years between 1943 and 2021 for the set $\mathcal{D}$ composed of 9 durations: 5-, 10-, 15-, 30-minute and 1-, 2-, 6-, 12- and 24-hour. The IDF curves provided by ECCC for this station are shown in Figure 1. Let Y_d denote the annual maximum of precipitation intensity for the accumulation duration $d \in \mathcal{D}$. The scaling models presented in this section aim to account for the functional dependence between the distributions accross durations.

2.1 No scaling

The first scaling model involves no scaling at all; it consists of modelling independently each duration. It does not take into account the functional dependence, and information across durations is not shared.

According to extreme value theory (Coles, 2001), the distribution for the annual maximum Y_d can be approximated by the Generalized Extreme Value distribution:

$$Y_d \sim \mathcal{G}EV(\mu_d, \sigma_d, \xi_d); \tag{A2}$$

where $\mu_d \in \mathbb{R}$, $\sigma_d > 0$, $\xi_d \in \mathbb{R}$ correspond respectively to the location, scale and shape parameters for duration $d \in \mathcal{D}$ and where the cumulative distribution function (CDF) is as follows:

$$G(y) = \begin{cases} \exp\left[-\left\{1 + \xi_d\left(\dfrac{y-\mu_d}{\sigma_d}\right)\right\}^{-1/\xi_d}\right] & \text{if } \xi_d \neq 0; \\ \exp\left\{-\exp\left(\dfrac{y-\mu_d}{\sigma_d}\right)\right\} & \text{if } \xi_d = 0; \end{cases}$$

defined on $\{y \ : \ 1 + \xi_d(y-\mu_d)/\sigma_d > 0\}$.

In Assumption A2, the location μ_d, the scale $\sigma_d > 0$ and the shape parameters ξ_d depend on the duration d. Assuming that the annual maxima for duration d are independent and identically distributed from the GEV distribution, the parameters (μ_d, σ_d, ξ_d) can be estimated. This estimation procedure is repeated for each duration independently.

Using this no scaling approach, IDF curves at the Pierre-Elliott-Trudeau International airport meteorological station are estimated and shown in Figure 2. The crosses are obtained independently for

each duration and correspond to the quantiles of the associated estimated $\mathcal{GEV}(\mu_d, \sigma_d, \xi_d)$ distribution using maximum likelihood. The following equation applies for return levels:

$$i(T,d) = \begin{cases} \mu_d + \frac{\sigma_d}{\xi_d}\left[\left\{-\log\left(1-\frac{1}{T}\right)\right\}^{-\xi_d} - 1\right] & \text{if } \xi_d \neq 0; \\ \mu_d - \sigma_d \log\left\{\log\left(1-\frac{1}{T}\right)\right\} & \text{if } \xi_d = 0. \end{cases}$$

They differ slightly from the ones represented in Figure 1, which resulted from fitting a Gumbel distribution (assuming $\xi = 0$) for each duration. The solid lines are estimated subsequently using Assumption A1. They result from simple least squares optimization, given the pointwise return levels (crosses). As one may notice, the line fit near the smaller durations is poor, as the return levels seem to curve slightly for durations less than a hour.

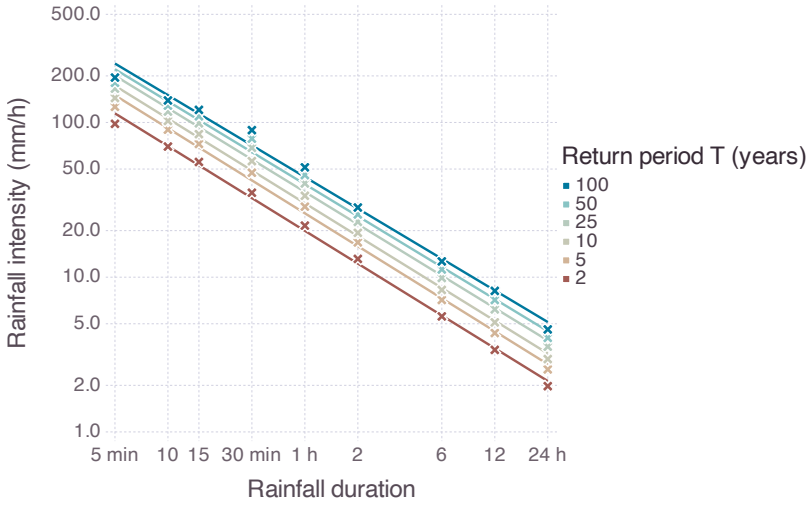


Figure 2: IDF curve for Montréal-Trudeau airport, obtained by fitting a GEV distribution for each duration without any scaling hypothesis.

Due to the uncertainty in GEV parameter estimation when using this model in practice, it is possible for the estimated return levels for different durations to cross each other, which is physically impossible as return levels are increasing in the return period. The scaling models presented in the next sections avoid this behavior.

It is not uncommon in practice to fix the shape parameters at 0 for all durations for precipitation. It is usually done for simplicity because this parameter is difficult to estimate. However, imposing a light-tailed distribution to the heavy-tailed precipitation may lead to quantile and return level underestimation. A simplistic scaling model would be to assume that the precipitations have the same shape parameter across durations:

$$Y_d \sim \mathcal{GEV}(\mu_d, \sigma_d, \xi).$$

Therefore, the information from all durations is pooled to estimate the common unknown shape parameter. This simplistic model decreases the chance of return level crossing along durations by reducing the sampling uncertainty of ξ, but does not eliminate it.

2.2 Simple scaling

Gupta & Waymire (1990) introduced the property of "strict sense simple scaling" for the rainfall process. When applied to precipitation intensity maxima, Burlando & Rosso (1996) showed that it generalizes directly the intensity-duration-frequency relation expressed in Assumption A1 into a relation over the probability distributions of Y_d for $d \in \mathcal{D}$:

$$Y_d \overset{\mathcal{L}}{=} \left(\frac{d}{d_0}\right)^{-\alpha} Y_{d_0};$$

for an arbitrarily chosen reference duration $d_0 > 0$, $0 < \alpha < 1$ and where $\overset{\mathcal{L}}{=}$ denotes equality in distribution. The parameter, α, is referred to as the scaling exponent. Its value does not depend on the chosen reference duration d_0, and it is bounded by physical laws. Assumption A1 can be rewritten as

$$i(T, d) = \left(\frac{d}{d_0}\right)^{-\alpha} i(T, d_0);$$

for any return period $T > 0$ and duration $d > 0$, where $a(T) = d_0^{\alpha} \times i(T, d_0)$ and $B = -\alpha$.

If Assumptions A1 and A2 hold for some α over the set of durations $\mathcal{D}$, then simple computations yield the following relation between the marginal GEV distributions of all durations $d \in \mathcal{D}$ and for any reference duration $d_0 > 0$:

$$Y_d \sim \mathcal{GEV}\left\{\mu_{d_0}\left(\frac{d}{d_0}\right)^{-\alpha}, \sigma_{d_0}\left(\frac{d}{d_0}\right)^{-\alpha}, \xi_{d_0}\right\};$$

where $(\mu_{d_0}, \sigma_{d_0}, \xi_{d_0})$ correspond to the marginal GEV parameters for the duration d_0.

This model, referred to as "Simple Scaling", provides the marginal distribution of all precipitation intensity maxima Y_d for any accumulation duration d and year index j using only the four parameters $(\mu_{d_0}, \sigma_{d_0}, \xi_{d_0}, \alpha)$. It enables the sharing of information across different durations. This model assumes the invariance of the shape parameter across durations.

The Simple Scaling model has been fitted to the precipitation data recorded at the Pierre-Elliott-Trudeau International Airport by maximum likelihood assuming that the data are statistically independent. The chosen reference duration is $d_0 = 1h$ and the parameter estimates are as follows:

$$\begin{aligned}
\hat{\mu}_{d_0} &= 18.1 \quad (17.7, 18.6);\\
\hat{\sigma}_{d_0} &= 5.29 \quad (4.95, 5.63);\\
\hat{\xi}_{d_0} &= 0.049 \quad (-0.009, 0.106);\\
\hat{\alpha} &= 0.694 \quad (0.682, 0.707).
\end{aligned}$$

The values in parenthesis correspond to the Wald 95% confidence intervals. Figure 3 shows the resulting IDF curves. The lines are obtained with the fitted Simple Scaling model, while the crosses consist of the return levels estimated independently for each duration as in Figure 2 for illustration purposes. The IDF curves in Figure 2 and Figure 3 are very similar, as the scaling hypothesis is the same for both methodologies. The Simple Scaling model is, however, more parsimonious, and parameter and return level uncertainties are therefore reduced (not shown in the figures). Yet, the curvature of the crosses along the durations is not captured by the Simple Scaling model.

2.3 General scaling model

Koutsoyiannis et al. (1998) generalized the Simple Scaling model by determining the mathematically feasible shapes for IDF curves to prevent return level crossing across different durations. They stated that all scaling relationships typically found in contemporary literature could be simplified to the following relation

$$i(T, d) = \frac{a(T)}{(d + \delta)^{\alpha}} \text{ for } d \in \mathcal{D} \text{ and } T > 0; \tag{A3}$$

where $0 < \alpha < 1$ and $\delta > 0$. This last equation generalizes Assumption A1 by introducing the duration offset δ. This additional parameter adds concavity in the log-log scale to the return level curve along

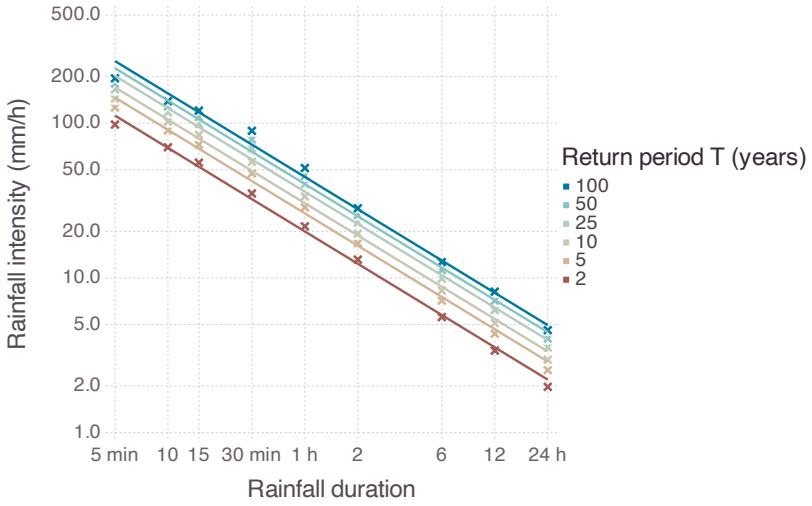


Figure 3: IDF curves at Montréal Pierre-Elliott-Trudeau International Airport estimated with the Simple Scaling model.

the durations and does not depend on the duration. As $\delta \to 0$, the return level curves become linear, simplifying the model to the Simple Scaling, while the curves' concavity increases with δ.

This assumption on return levels leads to the following relation over the probability distributions of Y_d:

$$Y_d \overset{\mathcal{L}}{=} \left(\frac{d+\delta}{d_0+\delta}\right)^{-\alpha} Y_{d_0};$$

for an arbitrarily chosen reference duration $d_0 > 0$ and $d \in \mathcal{D}$. In combination with Assumption A2, the marginal distribution of the annual maximum precipitation intensity for duration $d \in \mathcal{D}$ is given as follows:

$$Y_d \sim \mathcal{GEV}\left\{\mu_{d_0}\left(\frac{d+\delta}{d_0+\delta}\right)^{-\alpha}, \sigma_{d_0}\left(\frac{d+\delta}{d_0+\delta}\right)^{-\alpha}, \xi_{d_0}\right\};$$

where $(\mu_{d_0}, \sigma_{d_0}, \xi_{d_0})$ correspond to the marginal GEV parameters for the duration d_0. This model also assumes the invariance of the shape parameter across durations.

This model, referred to as "General Scaling" and as dGEV distribution by Koutsoyiannis et al. (1998), provides the marginal distributions of all precipitation intensity Y_d for any accumulation duration d using only the five parameters $(\mu_{d_0}, \sigma_{d_0}, \xi_{d_0}, \alpha, \delta)$. As the Simple Scaling model, the General Scaling model enables the sharing of information across different durations, but the latter is more flexible.

The General Scaling model has been fitted to the precipitation data recorded at the Pierre-Elliott-Trudeau International Airport. The chosen reference duration is $d_0 = 1h$ and the parameter estimates along with their 95% confidence intervals are as follows:

$$\begin{aligned}
\hat{\mu}_{d_0} &= 19.8 && (19.1, 20.6);\\
\hat{\sigma}_{d_0} &= 5.59 && (5.20, 5.99);\\
\hat{\xi}_{d_0} &= 0.0405 && (-0.0175, 0.085);\\
\hat{\alpha} &= 0.761 && (0.735, 0.787);\\
\hat{\delta} &= 0.068 && (0.041, 0.095).
\end{aligned}$$

Figure 4 shows the resulting IDF curves. The lines are obtained with the fitted General Scaling model, while the crosses consist of the return levels estimated independently for each duration as in

Figure 2 for illustration purposes. The General Scaling model naturally captures the curvature of the pointwise return levels.

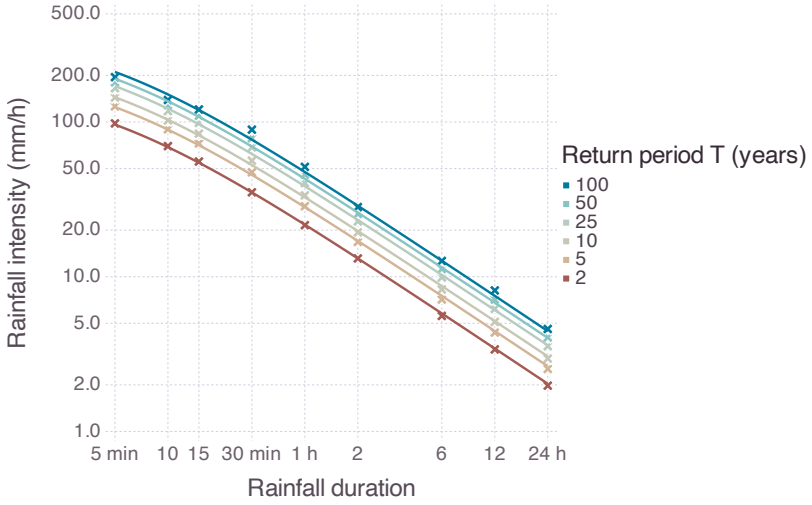


Figure 4: IDF curves at Montréal Pierre-Elliott-Trudeau International Airport estimated with the General Scaling model.

2.4 Additional scaling models

The General Scaling model is not the only one adding more flexibility to the Simple Scaling. Other scaling properties have been proposed, motivated by specific behaviors that may be encountered in real precipitation data. Bougadis & Adamowski (2006) noted that the simple scaling property does not hold for the entire range of precipitation durations but holds separately for short and long durations. They hence proposed a hybrid model comprising two Simple Scaling models, one for short durations and the other for long durations:

$$\begin{cases} Y_d = \left(\frac{d}{d_0}\right)^{-\alpha_1} Y_{d_0} & \text{if } d \leq d_0; \\ Y_d = \left(\frac{d}{d_0}\right)^{-\alpha_2} Y_{d_0} & \text{if } d > d_0; \end{cases}$$

where the reference duration $d_0 \in \mathcal{D}$ consists of the duration at which the break in scaling appears. This duration is used as the reference duration for both sub Simple Scaling models. Each of these submodels has its own scaling exponent: α_1 for durations $d \leq d_0$ and α_2 for $d > d_0$. We refer to this model as "Hybrid Scaling".

Assuming that Assumption A2 holds, another alternative proposed by Courty et al. (2019) to Simple Scaling consists of allowing distinct scaling exponents for the location and shape parameters of the GEV distribution:

$$\begin{cases} \mu_d = \mu_{d_0} \left(\frac{d}{d_0}\right)^{-\alpha_\mu}; \\ \sigma_d = \sigma_{d_0} \left(\frac{d}{d_0}\right)^{-\alpha_\sigma}; \\ \xi_d = \xi_{d_0}; \end{cases}$$

for any reference duration $d_0 > 0$ and for $0 < \alpha_\mu \leq 1$ and $0 < \alpha_\sigma \leq 1$. Unlike the other scaling models, this model is only defined when the marginal distributions family is the GEV. However, it is not a restrictive constraint, as the GEV is the natural distribution for modeling annual maxima according to extreme value theory. Such a model was motivated by an empirical study of the scaling behavior. Courty et al. (2019) observed that, when fitting a duration-independent GEV model, the

scaling exponent they observed for the location parameter was different from the one for the scale parameter. We refer to this model as "Composite Scaling".

3 Review of Anderson-Darling tests

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a random sample of independent and identically distributed one-dimensional variables from an unknown common cumulative distribution function F_X. The problem of testing whether these observations may come from a given distribution or family of distributions is classical and has resulted in the development of many goodness-of-fit tests. The most common one is the Anderson-Darling test (Anderson & Darling, 1952) where the test statistic measures the deviation between the target distribution and the empirical CDF defined as follows:

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{x_j \leq x}. \tag{1}$$

This section summarizes existing results on Anderson-Darling tests.

3.1 Test for a known distribution

When the cumulative distribution function of the target distribution, denoted as F, is fully specified, the null and alternative hypotheses of the Anderson-Darling test can be written as follows:

$$\begin{cases} \mathcal{H}_0 : & F_X = F; \\ \mathcal{H}_1 : & F_X \neq F. \end{cases}$$

Anderson & Darling (1952) proposed the following statistic for testing the previous hypotheses using the random sample $\boldsymbol{x}$:

$$S_n^{\Psi} = n \int_{-\infty}^{+\infty} \{F(x) - F_n(x)\}^2 \, \Psi\{F(x)\} \, dF(x); \tag{2}$$

where $\Psi : (0,1) \to \mathbb{R}_+$ is a weight function. S_n^{Ψ} is a measure of the mean squared error between the target and empirical distributions of X.

For the statistic to exist, the weight function Ψ must meet the following conditions (Anderson & Darling, 1952):

$$\begin{cases} \int_0^t u^2 \Psi(u) du < +\infty; \\ \int_t^1 (1-u)^2 \Psi(u) du < +\infty; \end{cases} \tag{3}$$

for $t \in (0,1)$. When $\Psi(u) = 1$, the tails and the bulk have the same importance and S_n^{Ψ} corresponds to the Cramér-Von Mises statistic. When $\Psi(u) = \frac{1}{u(1-u)}$, more weight is given to the tails of the distribution, and S_n^{Ψ} corresponds to the Anderson-Darling statistic. Other options have been considered in the literature; for instance, Sinclair et al. (1990) proposed $\Psi(u) = \frac{1}{1-u}$ when the upper tail is of specific interest. This could be particularly suitable for our use case as we are studying hydrological extremes.

When $\mathcal{H}_0$ is true, Anderson & Darling (1952) showed, under additional conditions on the weight function Ψ, that S_n^{Ψ} is asymptotically distributed as the integral of a squared Gaussian process:

$$\lim_{n \to \infty} \mathbb{P}(S_n^{\Psi} \leq s) = \mathbb{P}\left\{ \int_0^1 W^2(u) du \leq s \right\};$$

for $s \geq 0$ and where W denotes the Gaussian process of mean 0 and covariance function ρ defined as follows:

$$\rho(u,v) = \sqrt{\Psi(u)\Psi(v)} \, \{\min(u,v) - uv\};$$

where u and v take values in $(0, 1)$. This asymptotic distribution for the test statistic does not depend on the target distribution F. Therefore, the critical values for testing the hypotheses are well-known and tabulated (for instance by Smirnoff (1937) in the Cramér-Von Mises case).

Equation (2) is generally not used for computing the statistic S_n^Ψ. Anderson & Darling (1954) proposed a generic method to obtain an explicit formula for the statistic. Given the ordered sample $x_{(1)} \leq \ldots \leq x_{(n)}$, the formulas for the weight functions $\Psi(u) = 1$, $\Psi(u) = \frac{1}{u(1-u)}$ and $\Psi(u) = \frac{1}{1-u}$ are given respectively by the following equations:

$$S_n^1 = C_n^2 = \frac{1}{12n} + \sum_{j=1}^{n} \left\{ F(x_{(j)}) - \frac{2j-1}{n} \right\}; \tag{4a}$$

$$S_n^{\frac{1}{u(1-u)}} = A_n^2 = -n - \frac{1}{n} \sum_{j=1}^{n} (2j-1) \left[\log F(x_{(j)}) + \log \left\{ 1 - F(x_{(n-j+1)}) \right\} \right]; \tag{4b}$$

$$S_n^{\frac{1}{(1-u)}} = AU_n^2 = \frac{n}{2} - 2\sum_{j=1}^{n} F(x_{(j)}) - \sum_{j=1}^{n} \left(2 - \frac{2j-1}{n} \right) \log \left\{ 1 - F(x_{(j)}) \right\}. \tag{4c}$$

3.2 Test for a family of distributions

Anderson-Darling tests can also be used to test for a parametric family of distributions instead of a specific and fully specified distribution. Let $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ denote the parametric family of target distributions, where $F_{\boldsymbol{\theta}}$ represents the distribution with the parameter vector $\boldsymbol{\theta}$ taking values in the parameter space Θ. The Anderson-Darling test can be written as follows:

$$\begin{cases} \mathcal{H}_0 : & \exists \boldsymbol{\theta} \in \Theta : F_X = F_{\boldsymbol{\theta}}; \\ \mathcal{H}_1 : & \forall \boldsymbol{\theta} \in \Theta, \ F_X \neq F_{\boldsymbol{\theta}}. \end{cases}$$

Suppose that $\hat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimate of the parameter vector $\boldsymbol{\theta}$ based on the random sample $\boldsymbol{x}$ of size n. The Anderson-Darling test statistic can be adapted as follows:

$$S_n^\Psi = n \int_{-\infty}^{+\infty} \left\{ F_{\hat{\boldsymbol{\theta}}_n}(x) - F_n(x) \right\}^2 \Psi \left\{ F_{\hat{\boldsymbol{\theta}}_n}(x) \right\} dF_{\hat{\boldsymbol{\theta}}_n}(x). \tag{5}$$

The conditions on the weight function Ψ to ensure the existence of the statistic remain the same, stated in Equation (3). Furthermore, the explicit formulas given in Equation (4) remain valid, with the only modification being the replacement of the target distribution F with $F_{\hat{\boldsymbol{\theta}}_n}$, the distribution in the family $F_{\boldsymbol{\theta}}$ with the estimated parameters $\hat{\boldsymbol{\theta}}_n$. However, the asymptotic distribution of S_n^Ψ when $\mathcal{H}_0$ is true differs from the previous section where the target distribution is fully specified. Obtaining this asymptotic distribution is more challenging.

When the maximum likelihood estimator satisfies the usual regularity conditions (given, for instance, by Cramér, 1999), Darling (1955) found the asymptotic distribution when $\Psi(u) = 1$ and $\boldsymbol{\theta} = \theta$ is a scalar. Sukhatme (1972) extended the results to a parameter vector $\boldsymbol{\theta}$. Durbin (1973) consolidated the results by focusing on the weak convergence of an empirical process, allowing for a simplification of the necessary hypotheses to ensure convergence. Stephens (1976) studied the asymptotic distribution for several weight functions, particularly in the case of the two-parameter Gaussian family, without providing a formal proof. Laio (2004) extended these results without providing a proof for any weight function Ψ, and applied the test to the GEV family. To the best of our knowledge, a rigorously established asymptotic distribution is currently available exclusively for the scenario in which $\Psi(u) = 1$, and S_n^1 corresponds to the Cramér-Von Mises statistic.

Let's assume that $\mathcal{H}_0$ is true so a $\boldsymbol{\theta}_0 \in \Theta$ exists such that $F_X = F_{\boldsymbol{\theta}_0}$. Using the transformation $u = F_{\boldsymbol{\theta}_0}(x_u)$, let $g_{\boldsymbol{\theta}_0}(u)$ denotes the following function:

$$g_{\boldsymbol{\theta}_0}(u) = \nabla_{\boldsymbol{\theta}} F_{\boldsymbol{\theta}}(x_u)|_{\theta=\theta_0}; \tag{6}$$

and I the Fisher information matrix evaluated at $\boldsymbol{\theta}_0$:

$$I_{\boldsymbol{\theta}_0} = -\mathbb{E}_{(X|\boldsymbol{\theta}_0)}\left[\left.\frac{\partial^2 \log f_{\boldsymbol{\theta}}(X)}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]; \tag{7}$$

where $f_{\boldsymbol{\theta}}$ is the density function related to $F_{\boldsymbol{\theta}}$. Darling (1955) and Durbin (1973) showed that

$$\lim_{n\to\infty} \mathbb{P}(S_n^1 \leq s) = \mathbb{P}\left\{\int_0^1 W^2(u)du \leq s\right\};$$

for $s \geq 0$ and where W corresponds to the Gaussian process of mean 0 and covariance function ρ defined as follows:

$$\rho(u,v) = \min(u,v) - uv - g_{\boldsymbol{\theta}_0}(u)^\top I_{\boldsymbol{\theta}_0}^{-1} g_{\boldsymbol{\theta}_0}(v); \tag{8}$$

where u and v take values in $(0,1)$.

This result extends the one from the fully-specified case by adding the last term $-g_{\boldsymbol{\theta}_0}(u)^\top I_{\boldsymbol{\theta}_0} g_{\boldsymbol{\theta}_0}(v)$ to the covariance function of the Gaussian process W. As a result, the asymptotic distribution of the test statistic depends on the target family $F_{\boldsymbol{\theta}}$. It also depends on the true parameter value $\boldsymbol{\theta}_0$. Consequently, the Anderson-Darling test must be tailored specifically for the considered family of distributions, unlike the fully specified distribution where a general asymptotic distribution is available.

In practice, $\boldsymbol{\theta}_0$ is unknown and can be replaced by its maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$. The Fisher information matrix can also be replaced by its empirical counterpart. In general, the closed-form expression for the asymptotic distribution of S_n^1 is not available. However, its critical quantiles can be approximated using the method proposed by Zolotarev (1961) and presented in Appendix B.

4 Anderson-Darling test for a training and a validation sets

In the typical framework for testing a family of distributions, outlined in Section 3.2 for the Anderson-Darling test, the entire random sample serves two purposes: estimating the parameters of the parametric target family and assessing the goodness-of-fit of the fitted distribution to the data. An alternative approach involves splitting the random sample into a training set and a validation set. The data in the training set are used for parameter estimation, while those in the validation set are used to evaluate the goodness-of-fit of this fitted distribution to the data. This approach leads to a novel family of Anderson-Darling tests. The test procedure, statistic, and its asymptotic distribution under the null hypothesis, which differ from the typical framework, are described in this section. To the best of our knowledge, the theoretical properties of such a test have not been studied in the literature before.

4.1 Test hypotheses

As previously defined, let us recall that $\boldsymbol{x} = (x_1, \ldots, x_n)$ is an independent and identically distributed random sample according to the CDF F_X. Let $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ represent the parametric family of target distributions. The test hypotheses remain the same as those described in Section 3.2 for testing a family of distributions:

$$\begin{cases} \mathcal{H}_0: & \exists \boldsymbol{\theta} \in \Theta : F_X = F_{\boldsymbol{\theta}}; \\ \mathcal{H}_1: & \forall \boldsymbol{\theta} \in \Theta,\ F_X \neq F_{\boldsymbol{\theta}}. \end{cases}$$

4.2 Test statistic

Let $\mathcal{T}$ be the subset of indices of $\boldsymbol{x}$ corresponding to the training set, and let $\mathcal{V}$ be the subset of indices corresponding to the validation set, such that $\mathcal{T} \cup \mathcal{V} = \{1, \ldots, n\}$ and $\mathcal{T} \cap \mathcal{V} = \emptyset$. Let $\boldsymbol{x}^{\mathcal{T}} = (x_i : i \in \mathcal{T})$ denote the $m > 0$ data in the training set, and $\boldsymbol{x}^{\mathcal{V}} = (x_i : i \in \mathcal{V})$ denote the $\ell > 0$ data in the validation set.

The Anderson-Darling test statistic, as described in Section 3.2, can be adapted to the present framework of using a training and a validation set as follows:

$$S_{\ell,m} = \ell \int_{-\infty}^{+\infty} \left\{ F_{\hat{\theta}_m}(x) - F_\ell(x) \right\}^2 dF_{\hat{\theta}_m}(x); \tag{9}$$

where $\hat{\boldsymbol{\theta}}_m$ represents the maximum likelihood estimate of the parameters of the target family obtained using the training set, and F_ℓ represents the empirical CDF of the data in the validation set. In this last equation, the weight function corresponds to the identity function. It could be extended to any weight function Ψ that satisfies the conditions stated in Equation (3).

The explicit formula for the test statistic expressed in Equation (4a) can also be adapted to the present context as follows:

$$S_{\ell,m} = \frac{1}{12\ell} + \sum_{j=1}^{\ell} \left\{ F_{\hat{\boldsymbol{\theta}}_m}(x^{\mathcal{V}}_{(j)}) - \frac{2j-1}{\ell} \right\}.$$

The modifications involve replacing the target distribution F with its estimate $F_{\hat{\theta}_m}$ from the family based on the training set, and using $x^{\mathcal{V}}(j)$ to represent the j-th ordered value of the validation set.

4.3 Asymptotic distribution

In terms of the asymptotic distribution under the null hypothesis of the test statistic, the theoretical results presented in Section 3.2 cannot be directly applied. However, similar asymptotic properties can still be established in this new framework. Indeed, most of the propositions by Durbin (1973) regarding the test for a family of distributions can be adapted when parameter estimation is based on independent observations, namely the training set. While the main arguments remain unchanged, the resulting asymptotic distribution differs in an interesting way, as summarized in the following proposition:

Proposition 4.1. Let us assume that the same regularity conditions of the ones described by Durbin (1973) hold. Let's additionnally assume that the sample sizes of the training and validation sets are asympotically proportional, i.e.

$$\lim_{\substack{\ell\to\infty \\ m\to\infty}} \frac{m}{\ell} = a > 0.$$

If $\mathcal{H}_0$ is true, i.e. a $\boldsymbol{\theta}_0 \in \Theta$ exists such that $F_X = F_{\boldsymbol{\theta}_0}$, then

$$\lim_{\substack{\ell\to\infty \\ m\to\infty}} \mathbb{P}\left(S_{\ell,m} \le s\right) = \mathbb{P}\left\{ \int_0^1 W^2(u)\, du \le s \right\}; \tag{10}$$

where W denotes the Gaussian process with mean 0 and covariance function

$$\rho(u,v) = \min(u,v) - uv + g_{\boldsymbol{\theta}_0}(u)^\top \left(aI_{\boldsymbol{\theta}_0}\right)^{-1} g_{\boldsymbol{\theta}_0}(v); \tag{11}$$

and where $g_{\boldsymbol{\theta}_0}$ and $I_{\boldsymbol{\theta}_0}$ are defined in Equation (6) and Equation (7) respectively.

The proof of this proposition is available in Appendix A.

The test statistic asymptotic distribution is similar to the typical case described in Section 3.2, but the covariance function differs in an interesting way. The last term, involving the Fisher information matrix, is added when using both training and validation sets, whereas it is subtracted when using all the data to estimate $\boldsymbol{\theta}$ as expressed in Equation (8). In the fully specified target distribution in Section 3.1, no such term involving the Fisher information matrix was needed. An intuitive interpretation of the differences in the correlation function is the following. In the classical test for a family of distributions, since the data are used to estimate the distribution, the asymptotic distribution

must compensate for the fact that the estimated distribution fits the data too well. On the contrary, when using both training and validation sets, the asymptotic distribution must compensate for the uncertainty in the estimation of $\boldsymbol{\theta}_0$ and the imperfect specification of the estimated distribution.

In practice, $\boldsymbol{\theta}_0$ is unknown and can be replaced by its maximum likelihood estimate $\hat{\boldsymbol{\theta}}_m$ based on the training set. The Fisher information matrix can also be replaced by its empirical counterpart on the training set. In general, the closed-form expression for the asymptotic distribution of $S_{l,m}$ is not available. However, when it comes to computing critical thresholds and p-values for the test, its quantiles can be approximated using the method proposed by Zolotarev (1961). The details are available in Appendix B.

5 Anderson-Darling test procedure for scaling models using training and validation sets

The proposed Anderson-Darling test for a family of distributions using both a training and a validation set is applied to assess the goodness-of-fit of precipitation IDF scaling models. A slight modification in the definition of the family of distributions to account for the scaling relationship is described, along with the selection process for the training and validation sets.

5.1 Selection of training and validation sets

Given a target scaling model and n annual precipitation intensity maxima for the set of durations $\mathcal{D}$, the null hypothesis now states that "the data are distributed according to the target model".

As illustrated in Section 2 with an example using precipitation data recorded at the Pierre-Elliott-Trudeau International Airport, the various scaling models exhibit the greatest divergence for short durations. This suggests using precipitation corresponding to the smallest duration as the validation set and precipitation from the remaining durations as the training set. Consequently, the empirical distribution of the data in the validation set is compared with the target distribution for the smallest duration, obtained using one of the scaling models fitted on the training set.

The selection of the duration for validation is crucial. It should be a duration at which the discrepancy between the tested models is maximal to ensure maximum test power. In our case, the smallest (5-min) duration is optimal for precipitation IDF curves, and it has been demonstrated to outperform any other duration choice in the example of the Pierre-Elliott-Trudeau International Airport. However, this choice can be easily adapted to different contexts as it does not influence the overall testing procedure.

5.2 Formal definition

Let $d_{(1)}$ be the smallest duration, *e.g.* 5 minutes in the current application. Let $F_{\boldsymbol{\theta}}$ denotes the marginal target CDF for $Y_{d_{(1)}}$ given the parameter vector $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_n$ denotes the maximum likelihood estimation of the target scaling model parameter vector $\boldsymbol{\theta}$ based on the n annual maxima for each of the remaining durations $\mathcal{D} \setminus d_{(1)}$. The test statistic can be written as follows:

$$S_n = n \int_{-\infty}^{+\infty} \left\{ F_{\hat{\boldsymbol{\theta}}_n}(x) - \hat{F}_n(x) \right\}^2 dF_{\hat{\boldsymbol{\theta}}_n}(x).$$

It is important to notice that this is a special case of the goodness-of-fit test presented in Section 4. The set $\boldsymbol{x}^{\mathcal{V}}$ contains one-dimensional data corresponding to the smallest duration and $\boldsymbol{x}^{\mathcal{T}}$ contains multi-dimensional data corresponding to the other durations. In this case, $l = m$. It is not a direct application, as data in the training and validation sets are not identically distributed. Yet, they are supposed to be independent, as this is an inherent hypothesis to the IDF construction procedure.

Hence, Proposition 4.1 applies with $g_{\boldsymbol{\theta}}$ and $I_{\boldsymbol{\theta}}$ being computed over the validation set and the training set respectively. All the details are available in Appendix A.

5.3 Computing the test statistic and its p-value

The explicit formula for the test statistic expressed in Equation (4a) can be adapted to the present context, using the n ordered observations of the validation set $x^{\mathcal{V}}_{(1)} \leq \ldots \leq x^{\mathcal{V}}_{(n)}$ as follows:

$$S_n = \frac{1}{12n} + \sum_{j=1}^{n} \left\{ F_{\hat{\boldsymbol{\theta}}_n}(x^{\mathcal{V}}_{(j)}) - \frac{2j-1}{n} \right\}. \tag{12}$$

The asymptotic distribution of the test statistic depends on the target scaling model as well as its true parameter value $\boldsymbol{\theta}_0$, as expressed in Proposition 4.1. The covariance kernel ρ is estimated by employing $g_{\hat{\boldsymbol{\theta}}_n}$ and $\hat{I}_n$ from, respectively, Equation (13) and Equation (14) available in Appendix A. Efficient computation of the test statistic p-value associated with the asymptotic distribution of Equation (10) can be achieved using the approximation proposed by Zolotarev (1961) as described in Appendix B. Note that the critical region of the test can also be obtained using the same approximation.

For a more precise description of the testing procedure, an algorithm is provided in Appendix C. It's worth noting that this procedure is implemented in the open-source Julia library IDFCurves.jl. On an Apple M1 Pro processor with 16Go of Random Access Memory (RAM), the computation time for the whole testing procedure to complete at Montréal-Trudeau airport is ≈ 0.16 seconds when using Simple Scaling as the null hypothesis, and ≈ 0.25 seconds when using General Scaling as the null hypothesis.

6 Simulation study to assess the performance of the proposed test procedure applied to scaling models

The type I error (the rejection rate when the null hypothesis is true) and the power (the rejection rate of the test when the null hypothesis is false) of the proposed test are assessed through a simulation study. Let's consider the 9 durations as in the real precipitation data, namely 5-, 10-, 15-, 30-minute, and 1-, 2-, 6-, 12-, and 24-hour durations. Without loss of generality, let the reference duration d_0 be 24 hours. For each of these durations, we use the same sample size n. The validation set comprises observations for the 5-minute duration, and the training set corresponds to the remaining observations for the other durations.

6.1 Simple Scaling

6.1.1 Type I error

To assess the type I error when the target family in the null hypothesis is the Simple Scaling model (Section 2.2), 500 samples were generated from the Simple Scaling model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, and $\alpha = 0.7$ for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and different sample sizes ranging from $n = 5$ to $n = 300$. Figure 5 depicts the rejection rate of the null hypothesis as a function of the sample size and the shape parameter at a test significance level of 5%. The 95% confidence bands are obtained through a non-parametric bootstrap procedure. The empirical rejection rates of the null hypothesis when true closely align with the nominal level of 5%, even for small sample sizes. The type I error does not appear to be influenced by the shape parameter.

6.1.2 Test power against the General Scaling model

The power of the test for the Simple Scaling family is assessed against three different alternatives: General Scaling, Hybrid Scaling, and Composite Scaling. In the case of General Scaling (Section 2.3),

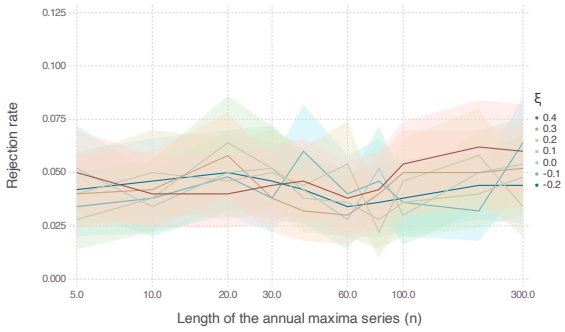
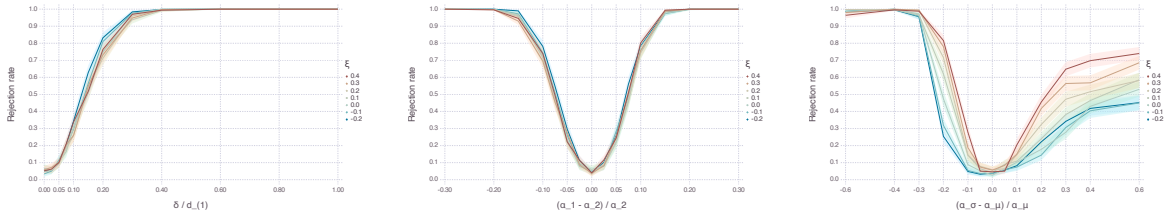


Figure 5: Type I error: rejection rate of the test procedure for the Simple Scaling model under the null hypothesis, as a function of the sample size and the shape parameter.

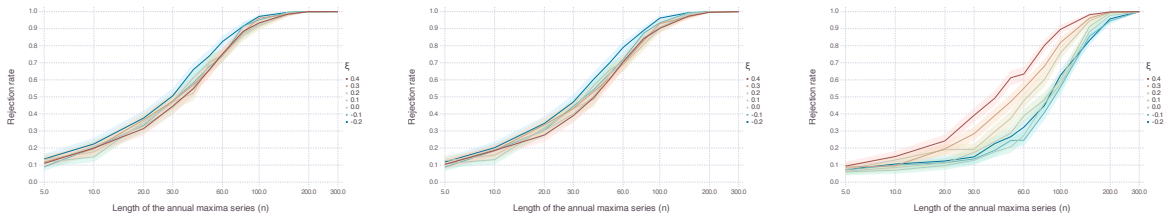
data have been generated from the model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, and $\alpha = 0.7$ for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and the duration offset $0 \leq \delta \leq d_{(1)}$. For a sample size of $n = 60$, Figure 6a illustrates the rejection rate of the null hypothesis stating that the data are distributed according to Simple Scaling, when in reality the true model is the General Scaling model. The test power rapidly increases with the ratio $\delta/d_{(1)}$. As δ approaches 0, the General Scaling model simplifies to the Simple Scaling model, and the rejection rate aligns with the fixed type I error. The test power is not influenced by the shape parameter. Finally, as expected, the test power increases with the sample size n, as shown in Figure 7a for random samples generated with $\delta/d_{(1)} = 0.2$.

In practice, for precipitation IDF curves, $d_{(1)} = 5$ min, and Koutsoyiannis et al. (1998) estimated the duration offset at $\hat{\delta} \approx 8$ min for precipitation recorded in Athens. In this case, the estimated ratio $\delta/d_{(1)}$ is 1.6, and the power of the proposed test is practically one. Even with a moderate sample size of $n = 60$, the test exhibits a good power for a ratio above 0.2, corresponding to a duration offset of $\delta = 1$ min for precipitation IDF curves.



(a) General Scaling model **(b) Hybrid Scaling model** **(c) Composite Scaling model**

Figure 6: Test power: rejection rate of the test procedure for the Simple Scaling model when, in fact, the true model is (a) the General Scaling model, (b) the Hybrid Scaling model, and (c) the Composite Scaling model.



(a) General Scaling model **(b) Hybrid Scaling model** **(c) Composite Scaling model**

Figure 7: Test power as a function of the sample size when the target model is the Simple Scaling model and the true model is (a) the General Scaling model, (b) the Hybrid Scaling model, and (c) the Composite Scaling model.

6.1.3 Test power against the Hybrid Scaling model

In the case of Hybrid Scaling (Section 2.4) where there is a break point in the return levels relation with the duration, data have been generated from the model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, $\alpha_2 = 0.7$ and the break-point at the 1-hour duration for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and the short-durations scaling exponent $\alpha_1 > 0$. For a sample size of $n = 60$, Figure 6b illustrates the rejection rate of the null hypothesis stating that the data are distributed according to Simple Scaling, when in reality the true model is the Hybrid Scaling model. When $\alpha_1 = \alpha_2$, the relative difference between α_1 and α_2 is 0 so the model simplifies to the Simple Scaling model and the rejection rate is equal to the fixed type I error. The test power increases with the relative difference between the two exponent parameters. The test power is not influenced by the shape parameter. Finally, as expected, the test power increases with the sample size n, as shown in Figure 7b for random samples generated with $\frac{\alpha_1 - \alpha_2}{\alpha_1} = -0.1$.

In their study, Huang et al. (2010) utilized the Hybrid Scaling model to construct precipitation IDF curves at various meteorological stations in China. They observed values of $\frac{\alpha_1 - \alpha_2}{\alpha_1}$ around -0.2. The negative nature of this value indicates a flattening of the curve at small durations. In such cases, the power of the proposed test is nearly one, even with a relatively small sample size of $n = 60$.

6.1.4 Test power against the Composite Scaling model

In the case of Composite Scaling (Section 2.4) where the scaling exponent is not the same for modeling μ and σ across durations, data have been generated from the model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, $\alpha_\mu = 0.7$ for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and the σ scaling exponent $\alpha_\sigma > 0$. For a sample size of $n = 60$, Figure 6c illustrates the rejection rate of the null hypothesis stating that the data are distributed according to Simple Scaling, when in reality the true model is the Composite Scaling model. When $\alpha_\mu = \alpha_\sigma$, the relative difference between α_μ and α_σ is 0 so the model simplifies to the Simple Scaling model and the rejection rate is equal to the fixed type I error. The test power increases with the relative difference between the two exponent parameters. The power increases more rapidly when $\alpha_\sigma < \alpha_\mu$ compared to the case $\alpha_\sigma > \alpha_\mu$. For this alternative, the shape parameter influences the test power, especially when $\alpha_\sigma > \alpha_\mu$. The power increases with the shape parameter. Finally, as expected, the test power increases with the sample size n, as shown in Figure 7c for random samples generated with $\frac{\alpha_\sigma - \alpha_\mu}{\alpha_\mu} = 0.3$.

Van de Vyver (2018) applied composite scaling at several rain gauges in Belgium. Model estimations resulted in $\frac{\alpha_\sigma - \alpha_\mu}{\alpha_\mu}$ values around 0.2 and a shape value around 0.08. In this situation, the power of the proposed test is around 0.35 when the sample size is $n = 60$.

6.2 General Scaling

6.2.1 Type I error

To assess the type I error when the target family in the null hypothesis is the General Scaling model, 500 samples were generated from the General Scaling model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, $\alpha = 0.7$ and $\delta = 3$ min for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and different sample sizes ranging from $n = 5$ to $n = 300$. Figure 8 depicts the rejection rate of the null hypothesis as a function of the sample size and the shape parameter at a test significance level of 5%. The 95% confidence bands are obtained through a non-parametric bootstrap procedure. The empirical type I errors align with the nominal level for sample sizes $n \geq 20$. There appears to be a small bias for very small sample sizes (e.g., $n = 5$), where the type I error increases to 7.5%. The type I error does not appear to be influenced by the shape parameter.

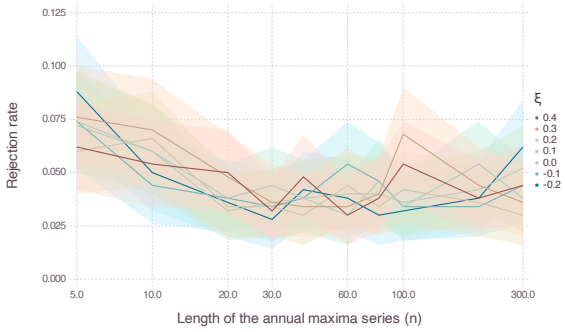
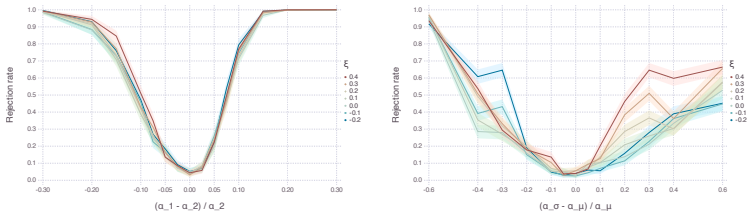


Figure 8: Type I error: rejection rate of the test procedure for the General Scaling model under the null hypothesis, as a function of the sample size and the shape parameter.

6.2.2 Test power against the Hybrid Scaling model

In the case of Hybrid Scaling, data have been generated from the model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, $\alpha_2 = 0.7$, and the breakpoint at the 1-hour duration for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and the short-durations scaling exponent $\alpha_1 > 0$. For a sample size of $n = 60$, Figure 9a illustrates the rejection rate of the null hypothesis stating that the data are distributed according to General Scaling, when in reality the true model is the Hybrid Scaling model. When $\alpha_1 = \alpha_2$, the model simplifies to the Simple Scaling model, and the rejection rate is equal to the fixed type I error, as the Simple Scaling model is a particular case of the General Scaling model when $\delta = 0$.

When $\alpha_1 > \alpha_2$, the return level curve as a function of the duration becomes convex, and the General Scaling model cannot capture this behavior. The more α_1 increases relatively to α_2, the greater the test power becomes, as the convexity increases. When $\alpha_1 < \alpha_2$, the return level curve as a function of the duration becomes concave, and the General Scaling model can sometimes mimic this behavior well. Although the General Scaling model may not be exact, it can provide an excellent approximation of the Hybrid Scaling model in many cases. Finally, as expected, the test power increases with the sample size n, as shown in Figure 7b for random samples generated with $\frac{\alpha_1 - \alpha_2}{\alpha_1} = -0.1$.

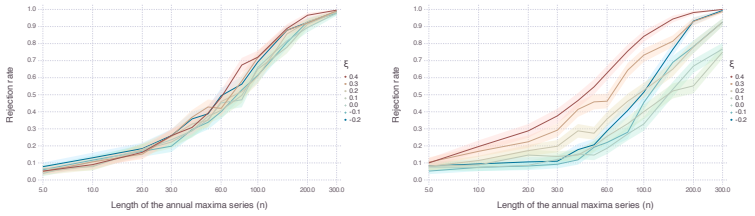


(a) Hybrid Scaling model **(b) Composite Scaling model**

Figure 9: Test power: rejection rate of the test procedure for the General Scaling model when, in fact, the true model is (a) the Hybrid Scaling model and (b) the Composite Scaling model.

6.2.3 Test power against the Composite Scaling model

In the case of Composite Scaling, data have been generated from the model using $\mu_{d_0} = 2$, $\sigma_{d_0} = 0.3$, $\alpha_\mu = 0.7$ for various values of the shape parameter $-0.2 \leq \xi \leq 0.4$ and the σ scaling exponent $\alpha_\sigma > 0$. For a sample size of $n = 60$, Figure 9b illustrates the rejection rate of the null hypothesis stating that the data are distributed according to General Scaling, when in reality the true model is the Composite Scaling model. When $\alpha_\mu = \alpha_\sigma$, the model simplifies to the Simple Scaling model, and the rejection rate is equal to the fixed type I error, as the Simple Scaling model is a particular case of the General Scaling model when $\delta = 0$.



(a) Hybrid Scaling model **(b) Composite Scaling model**

Figure 10: Test power as a function of the sample size when the target model is the General Scaling model and the true model is (a) the Hybrid Scaling model and (b) the Composite Scaling model.

When $\alpha_\sigma < \alpha_\mu$, the return level curve as a function of the duration becomes convex, and the General Scaling model cannot capture this behavior. The more α_σ increases relatively to α_μ, the greater the test power becomes, as the convexity increases. When $\alpha_\sigma > \alpha_\mu$, the return level curve as a function of the duration becomes concave, and the General Scaling model can sometimes mimic this behavior well. The shape parameter ξ has an influence for this alternative, as the test power decreases with ξ when $\alpha_\sigma > \alpha_\mu$.

Finally, as expected, the test power increases with the sample size n, as shown in Figure 10b for random samples generated with $\frac{\alpha_\sigma - \alpha_\mu}{\alpha_\mu} = 0.3$. Overall, in the current simulation framework, the proposed test is less powerful for the General Scaling model than it is for the Simple Scaling model. This could be due to the fact that the alternatives can be well approximated with the General Scaling model.

6.3 Conclusions of the simulation study

The proposed test for assessing whether either the Simple Scaling model or the General Scaling model is appropriate performs well in the simulation study framework. The type I errors of these two tests conform to the fixed nominal level, and the type II errors increase fairly rapidly as the true model diverges from either Simple Scaling or General Scaling. Additionally, the test power is fairly large for practical sample sizes, e.g., $n = 60$, and increases with the sample size. This is due to the fact that information from the data in the training set has been pooled together to predict the distribution of the data in the validation set. In the present framework, the data for the 5-minute duration constitute a satisfying discriminative validation set for testing a precipitation scaling model .

In the given simulation framework, the General Scaling model can approximate the two alternative models (Hybrid Scaling and Composite Scaling) very well, which impacts the test power. Indeed, for several combinations of parameters of the alternative models, the General Scaling model can approximate the data well, even though it is not the true model. Similarly, the Simple Scaling model can also approximate several parameter combinations of the Composite Scaling model effectively. However, these alternative models are retained because they exist in the literature and are physically plausible extensions of the General Scaling model for precipitation.

7 Applications on precipitation IDF data

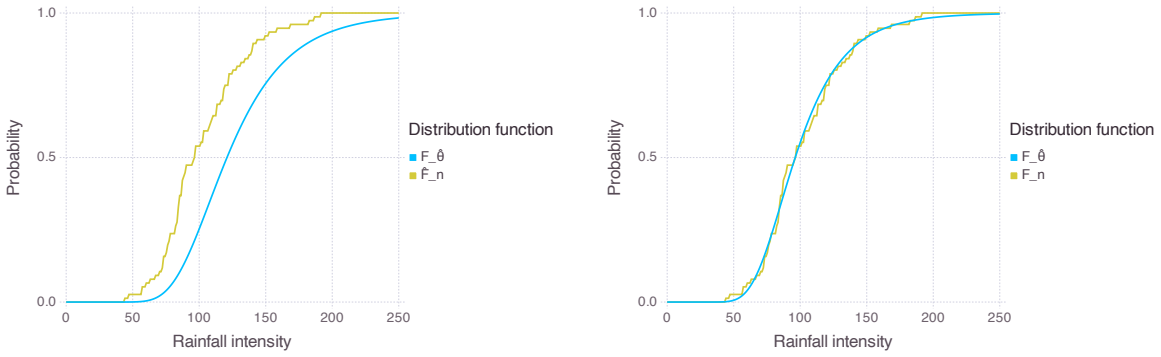
The proposed test is applied in this section to determine whether scaling models are suitable for modeling IDF curves of two Canadian meteorological stations operated by ECCC, namely the stations located at Montréal-Trudeau International Airport in Montréal (QC) and at the Harbour aerodrome in Vancouver (BC). These two stations are located in different climatic zones; humid continental for Montréal and maritime for Vancouver. The data are publicly accessible on the ECCC website: https://climate.weather.gc.ca/prods_servs/engineering_e.html and are also provided in the public repository: https://github.com/jojal5/Publications to reproduce all the results presented

in this article. The proposed goodness-of-fit test procedure is conducted for both the Simple Scaling model and the General Scaling model as the null hypothesis. The annual maxima corresponding to durations larger than 5 minutes form the training set, while those corresponding to the 5-minute duration comprise the validation set.

7.1 Montréal

The proposed test procedure is applied on the precipitation annual maxima showed in Figure 1. The test p-value of 1.4×10^{-6} suggests rejecting the null hypothesis for the Simple Scaling model, whereas a p-value greater than 0.1 suggests not rejecting the null hypothesis for the General Scaling model. Since the Zolotarev approximation (Appendix B) is only valid for estimating small p-values, we know that it is larger than 0.1, although its exact value is imprecise. Figure 4 shows the estimated IDF curves using the General Scaling model.

As a visual indication, Figure 11a displays the empirical cumulative distribution function (CDF) of the 5-minute annual maxima, along with the CDF computed using the Simple Scaling model excluding the 5-minute data. The Simple Scaling CDF is biased compared to the empirical CDF, supporting the rejection of the Simple Scaling Model. In contrast, Figure 11b displays the empirical CDF along with CDF computed using the General Scaling model. The two curves match almost perfectly, supporting the non-rejection of the General Scaling model.



(a) Simple Scaling **(b) General Scaling**

Figure 11: Empirical cumulative distribution function of the 5-minute annual maxima F_n, along with the CDF $F_{\hat{\theta}}$ computed using (a) the Simple Scaling model and (b) the General Scaling model, excluding the 5-minute data.

7.2 Vancouver

Figure 12 displays the IDF curves at the Vancouver Harbour CS station provided by ECCC. The marginal return levels, estimated separately for each duration, seem quite scattered around the regression lines; furthermore, those lines are not parallel. Hence, the linear relationship of return levels across durations does not appear ideal. However, due to the small sample size of only 25, the marginal return levels have large uncertainties. Pooling information across durations using a scaling model could be beneficial for improving return level estimates by reducing estimation uncertainty, assuming the scaling model is reasonable.

Using the Simple Scaling model as the null hypothesis, the proposed testing procedure yields a p-value of 0.048. Therefore, there is not enough evidence to reject the Simple Scaling model for these data.

As a visual indication, Figure 13 displays the empirical cumulative distribution function (CDF) of the 5-minute annual maxima, along with the CDF computed using the Simple Scaling model without

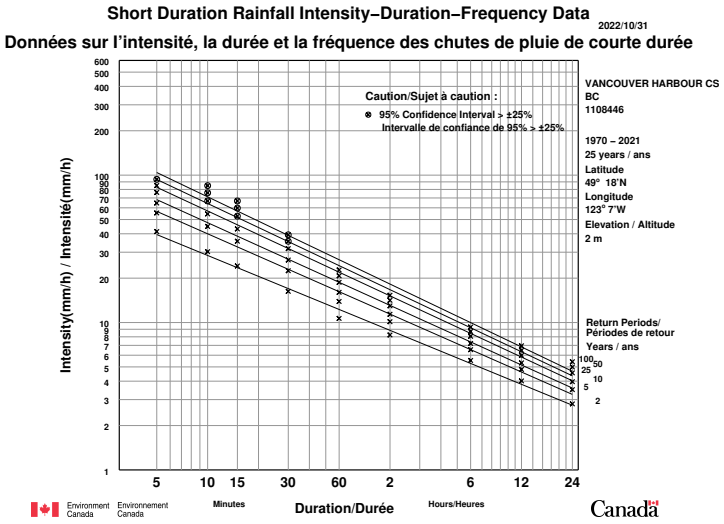


Figure 12: IDF curves at the Vancouver Harbours CS station in BC provided by ECCC.

utilizing the 5-minute data. These two curves appear consistent and support the conclusion of the test.

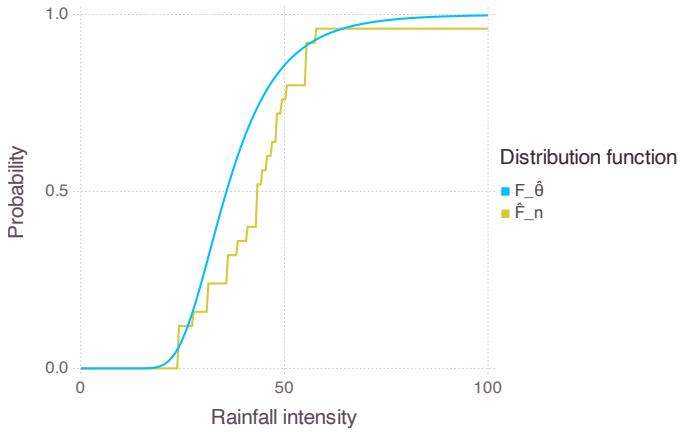


Figure 13: Empirical cumulative distribution function of the 5-minute annual maxima F_n, along with the CDF $F_{\hat{\theta}}$ computed using the Simple Scaling model without including the 5-minute data.

Given that the proposed goodness-of-fit test did not reject the Simple Scaling model, this model can now be fitted to the data of all durations. This approach allows for information sharing across durations and provides estimations with less uncertainty. Figure 14a displays the estimated IDF curves using the Simple Scaling model alongside the estimations of marginal return levels. These curves are more precise than those obtained by the current approach of linearly fitting the curves to the marginal quantiles, which are quite uncertain in this case. With the Simple Scaling model, the quantiles are estimated jointly rather than separately for each duration. As a visual indication, Figure 14b shows the 2-year and 100-year return levels alongside the marginal estimates and their uncertainties. The curve obtained with the Simple Scaling model cuts through the marginal uncertainties.

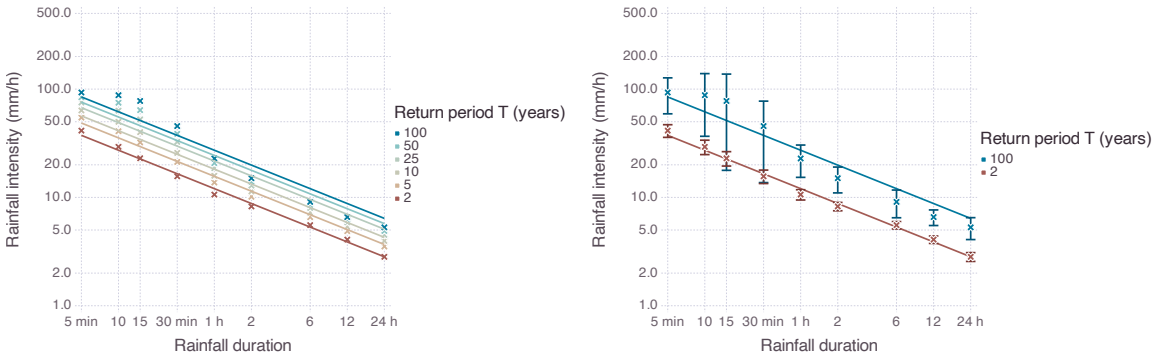


Figure 14: IDF curves estimated with the Simple Scaling model at the Vancouver Harbour CS station for (a) all return periods and (b) for the 2-year and 100-year return levels along with the uncertainties of the marginal return levels.

8 Discussion and conclusion

In this paper, a formal goodness-of-fit test has been proposed to verify if a given scaling model is appropriate to build IDF curves at a location where data is available. The test statistic consists of the Anderson-Darling test statistic, but with the difference that it is computed on a validation set containing the data for one chosen duration. The corresponding model distribution is estimated with the training dataset that contains the other durations. The asymptotic distribution under the null hypothesis has been provided, along with an efficient method for computing its quantiles.

The test has been applied to assess the goodness-of-fit of the Simple Scaling model and the General Scaling model on precipitation IDF data. The simulation study showed that the type I errors for testing both the Simple Scaling and the General Scaling models align with the fixed nominal level, even for small sample sizes. When the target model is the Simple Scaling model, the test power is large, even for small sample sizes. When the target model is the General Scaling model, the test power is moderate when comparing to the Composite Scaling model alternative. This is mostly due to the fact that the General Scaling model is able to approximate this alternative quite well. Other alternative models could be used to evaluate the test power more thoroughly, but it is difficult to develop physically consistent scaling relations between precipitation durations. However, the Hybrid Scaling and Composite Scaling models were retained as possible alternatives because they have been used in the literature and are consistent with precipitation behavior. Except for the Composite Scaling model, the type I and II errors of both tests are not influenced by the shape parameter.

The proposed goodness-of-fit tests were also applied to real precipitation IDF data recorded in Montréal (QC) and Vancouver (BC). For the Montréal data, the General Scaling model could not be rejected, and for the Vancouver data, the Simple Scaling model could not be rejected at the 1% level. Visual indications have been provided to corroborate these results. These conclusions enable the choice of a suitable scaling model for IDF curves, and then enable the sharing of information between precipitation durations to estimate more precise IDF curves.

In the application for precipitation IDF models, the 5-minute duration was used as the validation set as it is the most discriminant duration for distinguishing between precipitation IDF scaling models, notably the Simple Scaling and the General Scaling models. Another duration could be directly used as the validation set. A general subset of durations could also compose the validation set, yet the asymptotic distribution of the test statistic should be adapted accordingly.

The goodness-of-fit test procedure proposed in this paper is general when using a validation set to compute the test statistic and not only for testing scaling models. In the classical Anderson-Darling test for a family of distributions described in Section 3, the whole dataset is used to fit the model, and

the test statistic is computed with the same data. One could also test for a family of distributions after splitting the data into training and validation sets. The test power could be compared with the classical one as a function of the distribution family and the sizes of these sets. As a practical application, the proposed testing procedure could also be developed for testing among spatial interpolation models, whereas scaling models considered in this paper concern temporal interpolation. These models are particularly useful when constructing IDF curves over a large geographical area (see e.g., Jalbert et al., 2022). It would be natural to use a subset of grid cells as the validation set, and the others as the training set.

The general Anderson-Darling test statistic expressed in Equation (5) could have been adapted for different weight functions Ψ. For instance, Sinclair et al. (1990) proposed $\Psi(u) = \frac{1}{1-u}$ to give more weight to the right tail in the Anderson-Darling test statistic for a known target distribution. It could have been interesting to use such a weight function in the application to precipitation IDF models. However, to the best of our knowledge, the asymptotic distribution of the test statistic using $\Psi(u) = \frac{1}{1-u}$ has never been formally proved, yet it has been utilized by Stephens (1976) and Laio (2004). In the training and validation framework, we were unable to derive the asymptotic distribution of the test statistic when $\Psi(u) \neq 1$. It could be interesting for future work to pursue the investigation of the asymptotic distribution for other weight functions. However, in our preliminary analysis, the test power was not improved with a function giving more weight to the right tail.

Finally, an interesting avenue for future work could be to study the effect of statistical dependence in the maxima of a given year across durations. This dependence can arise when several annual maxima of different durations come from the same storm. The marginal distributions would still be given by the scaling model due to the functional dependence, but statistical dependence could impact parameter estimation and change the probability distribution of the test statistic. At this point, it is unknown whether this phenomenon is common in the data or not.

A Asymptotic distribution of the test statistic when using a training and a validation sets

In this appendix, a rigorous proof is given for the result of Proposition 4.1, that states the asymptotic null distribution of the statistic $S_{l,m}$ defined in Equation (9). The given result is slightly more general than the one presented in Proposition 4.1, as it allows the respective distributions of the training set and the validation set to be different.

The proof is mostly adapted from Durbin (1973). In Section A.1, the training / validation framework is defined formally and the necessary notation is introduced. In Section A.2, the precise hypotheses and all the technical details are given in order to prove weak convergence of the sample process $\hat{y}_{l,m}$, with emphasis on how the arguments of Durbin (1973) are modified. A continuity result, given in Section A.3, is necessary to extend convergence to the integral $S_{l,m}$. The final result is wrapped up in Section A.4.

A.1 Theoretical framework

Let $x_1, x_2, \ldots, x_l$ be independent and identically distributed 1-dimensional observations, whose common cumulative distribution function is F_{θ_0}, where the true parameter value $\theta_0 \in \mathbb{R}^p$ is unknown. Let $z_1, z_2, \ldots, z_m$ be independent and identically distributed possibly multi-dimensional observations, whose common cumulative distribution function is H_{θ_0}. Let's suppose that the $(X_i)_{1\leq i\leq l}$ are independent from the $(Z_i)_{1\leq i\leq m}$, but that their distributions are directly bounded as θ_0, the true parameter value, is the same for both. Let $\hat{\theta}_m$ be an estimator of θ_0 based on $z_1, z_2, \ldots, z_m$.

Let's introduce the sample process $\hat{y}_{l,m}(t) = \sqrt{l}\left(\hat{F}_{l,m}(t) - t\right)$, $0 \leq t \leq 1$, where $\hat{F}_{l,m}(t) = \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{\left\{F_{\hat{\theta}_m}(x_i)\leq t\right\}}$. One may notice that $S_{l,m} = \int_0^1 \hat{y}_{l,m}(t)^2 dt$, hence the distribution of $S_{l,m}$ is

entirely specified by that of the process $\hat{y}_{l,m}$. We will focus on the weak convergence of the process $\hat{y}_{l,m}$, in the space D of the right-continuous functions with left-hand limits on $[0,1]$, on which we use the Skorokhod metric. These notions were introduced by Billingsley (1968). All the definitions and results around weak convergence that are necessary for the present work were summarized by Fleming & Harrington (2005).

When $\hat{\theta}_m = \theta_0$, the sample process $\sqrt{l}\,(F_l(t) - t)$ converges weakly to a Brownian bridge, hence the asymptotic result given in Section 3.1. Durbin (1973) treated extensively the case where θ_0 is estimated from the observations $x_1, x_2, \ldots, x_l$, which led to the results given in Section 3.2. We will try and adapt his hypotheses and arguments to the present framework.

A.2 Weak convergence of $\hat{y}_{l,m}$: adaptation of the Durbin proof

Let $x(t,\theta) = \inf\{x : F_\theta(x) \leq t\}$ (the generalized inverse of F_θ). Let's also define:

$$g(t,\theta) = \nabla_\theta F_\theta(x)|_{\theta=\theta, x=x(t,\theta)}\,. \tag{13}$$

Two separated sets of assumptions are made. **(S1)** concerns the estimator $\hat{\theta}_m$, hence the training set, while **(S2)** states necessary regularity conditions on the distribution of the validation set.

(S1) $\sqrt{m}\left(\hat{\theta}_m - \theta_0\right) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} q(z_j,\theta_0) + \epsilon_m$, where:

- **(i)** $\mathbb{E}_{(\boldsymbol{Z}|\theta_0)}\{q(\boldsymbol{Z},\theta_0)\} = 0$;
- **(ii)** $\mathbb{E}_{(\boldsymbol{Z}|\theta_0)}\{q(\boldsymbol{Z},\theta_0)q(\boldsymbol{Z},\theta_0)^t\} = Q(\theta_0)$, and $Q(\theta_0)$ is positive semi-definite;
- **(iii)** $\epsilon_m \xrightarrow[m\to\infty]{\mathbb{P}} 0$.

(S2) There is a compact set $\mathcal{V} \subset \mathbb{R}^p$, containing a neighborhood of θ_0, such that:

- **(i)** $\forall\theta \in \mathcal{V}, F_\theta$ is continuous;
- **(ii)** The function g of Equation (13) exists and is continuous on $[0,1]\times\mathcal{V}$.

Under those assumptions, one may paraphrase the proof of *Theorem 1* in Durbin (1973) and adapt it to the present framework. This leads to:

Lemma A.1 (adapted from Durbin (1973)). Let's assume that **(S1)** and **(S2)** hold. Let's additionnally assume that the sample sizes of the training and validation sets are asympotically proportional, i.e.

$$\lim_{\substack{l\to\infty\\ m\to\infty}} \frac{m}{l} = a > 0.$$

Then $\hat{y}_{l,m}$ converges weakly to the normal process $(W(t))_{0\leq t\leq 1}$ in D, with mean 0 and covariance function

$$\rho(u,v) = \min(u,v) - uv + a^{-1}g(u,\theta_0)^t Q(\theta_0) g(v,\theta_0); \tag{14}$$

where g was defined in Equation (13).

Proof. We will only give a sketch of proof and insist on how the arguments given by Durbin (1973) may be adapted to our case. His proof was divided in five lemmas, that we shall call here " results ".

The first two results are about rewriting $\hat{y}_{l,m}$ in terms of the " real " sample process $y_l(t) = \sqrt{l}\,(F_l(t) - t)\,,\ 0 \leq t \leq 1$, where $F_l(t) = \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{\{F_{\theta_0}(x_i)\leq t\}}$, and the scaled error estimation $w_m = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} q(z_j,\theta_0)$.

Let's introduce the function $\hat{\approx}_m(t) = F_{\theta_0}\left(x(t,\hat{\theta}_m)\right)$. As in Durbin (1973), one may obtain from the fact that $\hat{\theta}_m \xrightarrow[m\to\infty]{\mathbb{P}} \theta_0$ and the *mean value theorem* applied to the function $(x,\theta) \mapsto F_\theta(x)$ that

$\sup_{0\leq t\leq 1}|\hat{\approx}_m(t)-t| \xrightarrow[m\to\infty]{\mathbb{P}} 0$. Using the continuity of the mapping $(y_l,\hat{\approx}_m)\in D\times D\mapsto y_l\circ\hat{\approx}_m - y_l$, one gets the

Result 1: $y_l\left(\hat{\approx}_m(t)\right)=y_l(t)+\epsilon^{(2)}_{l,m}(t)$, where $\epsilon^{(2)}_{l,m}\xrightarrow[l,m\to\infty]{\mathbb{P}}\not\vdash$.

A key thing to notice here is that $\hat{y}_{l,m}(t)=y_l\left(\hat{\approx}_m(t)\right)+\sqrt{l}\left(\hat{\approx}_m(t)-t\right)$. Besides, paraphrasing Durbin (1973), one may obtain that $\sqrt{m}\left(\hat{\approx}_m(t)-t\right)=-w_m^t g(t,\theta_0)+\epsilon^{(3)}_m(t)$, where $\epsilon^{(3)}_m\xrightarrow[m\to\infty]{\mathbb{P}}\not\vdash$. Hence from result 1 and the assumption $\frac{m}{l}\underset{l,m\to\infty}{\longrightarrow} a$:

Result 2: $\hat{y}_{l,m}(t)=z_{l,m}(t)+\epsilon^{(4)}_{l,m}(t)$, where $z_{l,m}(t)=y_l(t)-a^{-\frac{1}{2}}w_m^t g(t,\theta_0)$ and $\epsilon^{(4)}_{l,m}\xrightarrow[l,m\to\infty]{\mathbb{P}}\not\vdash$.

Hence, it boils down to proving the weak convergence of $z_{l,m}$.

Like Durbin (1973), we begin by considering the finite-dimensional distributions. As $y_l(t)=\frac{1}{\sqrt{l}}\sum_{i=1}^{l}\left(\not\vdash_{\{F_{\theta_0}(x_i)\leq t\}}-t\right)$ and $a^{-\frac{1}{2}}w_m^t g(t,\theta_0)=\frac{1}{\sqrt{m}}\sum_{j=1}^{m}a^{-\frac{1}{2}}q(z_j,\theta_0)g(t,\theta_0)$, one may use the multi-dimensional *central limit theorem* twice and use the fact that y_l and w_m are independent to get the

Result 3: $\forall 0<t_1<\ldots t_k<1$, $[z_{l,m}(t_1),\ldots,z_{l,m}(t_k)]\xrightarrow[l,m\to\infty]{\mathcal{L}}[W(t_1),\ldots,W(t_k)]$ where W is the normal process in D with mean 0 and covariance function given in Equation (14).

Strict paraphrase of Durbin (1973) leads to the

Result 4: The sequence $z_{l,m}$ is tight.

Using a theorem from Billingsley (1968) (see also corollary B.1.1 of Fleming & Harrington (2005)), it comes from results 3 and 4:

Result 5: $z_{l,m}\xrightarrow[l,m\to\infty]{\mathcal{L}} W$.

From result 2, $d(\hat{y}_{l,m},z_{l,m})\xrightarrow[l,m\to\infty]{\mathbb{P}}0$. Hence the desired conclusion. □

A.3 Continuity lemma

In order to extend that convergence result to the integral $S_{l,m}$, one needs the

Lemma A.2. The functional $\left\{\begin{array}{ccc} D & \to & \mathbb{R}_+ \\ f & \mapsto & \int_0^1 f^2(t)dt \end{array}\right.$ is continuous with respect to the Skorokhod topology.

Proof. Let be a sequence $(f_k)_{k\leq 0}\in D^{\mathbb{N}}$ that converges to $f\in D$ with respect to the Skorokhod metric. Then:

- **(i)** f has a countable number of discontinuity points, and at every point $t\in[0,1]$ where f is continuous, $f_k(t)\underset{k\to\infty}{\longrightarrow}f(t)$. Hence, $f_k\underset{k\to\infty}{\longrightarrow}f$ almost everywhere.
- **(ii)** As $(f_k)_{k\leq 0}$ is a convergent sequence, it is bounded. Hence, there is a constant $M\geq 0$ such that $\forall k\geq 0, d(f_k,\not\vdash)\leq M$, which leads to $\forall k\geq 0,\|f_k\|_\infty\leq M$.

From **(i)**, **(ii)** and the Lebesgue's *dominated convergence theorem*, we get that $f_k\underset{k\to\infty}{\longrightarrow}f$ in $L^2([0,1])$. Hence $\int_0^1 f_k^2(t)dt\underset{k\to\infty}{\longrightarrow}\int_0^1 f^2(t)dt$, which proves continuity. □

A.4 Final result

From Lemma A.1, Lemma A.2 and the *Continuous Mapping Theorem* (see e.g Fleming & Harrington, 2005), it comes:

Proposition A.1. Let's assume that **(S1)** and **(S2)** hold. Let's additionnally assume that the sample sizes of the training and validation sets are asympotically proportional, i.e.

$$\lim_{\substack{l\to\infty\\ m\to\infty}} \frac{m}{l} = a > 0.$$

Then $S_{l,m}$ (defined in Equation (9)) converges in probability:

$$\lim_{\substack{l\to\infty\\ m\to\infty}} \mathbb{P}\left(S_{l,m} \leq s\right) = \mathbb{P}\left\{\int_0^1 W^2(t)\, du \leq s\right\};$$

where W denotes the Gaussian process in D with mean 0 and covariance function ρ defined in Equation (14).

Finally, Durbin (1973) treated the special case when $\hat{\theta}_m$ is the Maximum Likelihood Estimator. He showed that, when $\hat{\theta}_m$ satisfies the usual regularity conditions stated by Cramér (1999), then assumptions **(S1)** hold, with

$$Q(\theta_0)^{-1} = I_{\theta_0} = -\mathbb{E}_{(Z|\theta_0)}\left[\left.\frac{\partial^2 \log h_\theta(Z)}{\partial\theta^2}\right|_{\theta=\theta_0}\right].$$

In conclusion, the result of Proposition 4.1 holds if $\hat{\theta}_m$ satisfies the usual regularity conditions stated by Cramér (1999) and if the assumptions **(S2)** hold.

B Zolotarev approximation for the proposed tests

The results given in Section 3 and Section 4, concerning the asymptotic distribution of the test statistic, all involve the distribution of $\int_0^1 W^2(t)du$, where $W(t)$ is a Gaussian process with mean 0 and some covariance function ρ. Yet, quantiles from such a distribution cannot be computed explicitly, hence one needs further results in order to compute asymptotic critical thresholds or p-values for our test.

The first result is due to Kac & Siegert (1947). They proved that the gaussian integral $\int_0^1 W^2(t)dt$ corresponds, in terms of distribution, to a discrete weighted sum of $\chi^2(1)$ variables:

Proposition B.1 (Kac & Siegert (1947))**.** Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the eigenvalues of the kernel ρ, ie. the characteristic values associated with the functional equation:

$$\forall x \in [0,1],\ \lambda f(x) = \int_0^1 \rho(x,y) f(y) dy. \tag{15}$$

Then

$$\forall s \geq 0,\ \mathbb{P}\left(\int_0^1 W^2(t)dt \leq s\right) = \mathbb{P}\left(\sum_{k=1}^{+\infty} \lambda_k G_k^2 \leq s\right);$$

where $G_1, G_2, \ldots$ are independent and identically distributed (i.i.d.) according to a $\mathcal{N}(0,1)$.

The cumulative distribution function (CDF) of such a weighted sum of $\chi^2(1)$ variables can still not be computed explicitly. Yet, Zolotarev (1961) provided a useful approximation:

Proposition B.2 (Zolotarev (1961))**.** Let Q be the cumulative distribution function of $\sum_{k=1}^{+\infty} \lambda_k G_k^2$ where $G_1, G_2, \ldots$ are i.i.d. $\mathcal{N}(0,1)$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are such that $\sum_{k=1}^{+\infty} \lambda_k < +\infty$.

Let's account for multiplicity by defining $\gamma_1 > \gamma_2 > \ldots \geq 0$ and positive integers $l_1, l_2, \ldots$ such that $\gamma_1 = \lambda_1 = \ldots = \lambda_{l_1},\ \gamma_2 = \lambda_{l_1} = \ldots = \lambda_{l_1+l_2},\ \ldots\ldots$

Then:

$$1 - Q(x) \underset{x\to+\infty}{=} \frac{\prod_{r=2}^{+\infty}\left(1-\frac{\gamma_r}{\gamma_1}\right)^{-\frac{l_r}{2}}}{\Gamma\left(\frac{l_1}{2}\right)}\left(\frac{x}{2\gamma_1}\right)^{\frac{l_1}{2}-1}\exp\left(-\frac{x}{2\gamma_1}\right)(1+o(1)). \tag{16}$$

Proposition B.2 may be used to approximate high-order quantiles from the distribution of $\int_0^1 W^2(t)dt$, hence explicit theoretical critical values for our test. Yet, this requires knowing the eigenvalues of the kernel ρ (see Equation (15)), which can not be computed analytically. Hence, one needs a further simplification. Schlesinger (1957) provided a method to approximate the q biggest eigenvalues of ρ, with q sufficiently large, by using the Riemann integration rule: $\int_0^1 h(u)du = \lim_{q\to\infty} \sum_{i=1}^{q} \frac{1}{q} h\left(\frac{2i-1}{2q}\right)$. This leads to replace Equation (15) by

$$\forall 1 \leq i \leq q,\ \lambda f\left(\frac{2i-1}{2q}\right) = \sum_{j=1}^{q} K_{i,j} f\left(\frac{2j-1}{2q}\right); \tag{17}$$

where $\forall 1 \leq i,j \leq q, K_{i,j} = \rho\left(\frac{2i-1}{2q}, \frac{2j-1}{2q}\right)$. Hence, this boils down to the problem of computing the eigenvalues of the $q \times q$ matrix K, which is numerically easy. Finally, given $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_q \geq 0$ and q sufficiently large, Proposition B.2 may be used to compute the high-order quantiles of $\sum_{k=1}^{q} \lambda_k G_k^2$, hence those of $\int_0^1 W^2(t)dt$ by replacing $\prod_{r=2}^{+\infty}\left(1-\frac{\gamma_r}{\gamma_1}\right)^{-\frac{l_r}{2}}$ by its finite counterpart.

C The testing procedure as an algorithm

The testing procedure can be written as a formal sequence of computations listed in the Algorithm 1. The inputs of the algorithm are the target scaling model and precipitation IDF data $(y_{dj})_{d\in\mathcal{D}, 1\leq j\leq n}$. It returns the p-value of the test.

Algorithm 1: The testing procedure.

Input: A target model $\mathcal{M}$, data $(y_{dj})_{d\in\mathcal{D}, 1\leq j\leq n}$.
Output: The p-value of the test p.
Parameters: The validation duration $d_{(1)}$, the number q of eigenvalues to compute.

`/* Splitting the data into validation and training sets: */`
1 $\boldsymbol{y}^{\mathcal{V}} \leftarrow (y_{d_{(1)}j} : 1 \leq j \leq n)$;
2 $\boldsymbol{y}^{\mathcal{T}} \leftarrow (y_{dj} : d \in \mathcal{D} \setminus \{d_{(1)}\}, 1 \leq j \leq n)$;
`/* Estimation using the training set: */`
3 $\ell_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y}^{\mathcal{T}}) \leftarrow$ the log-likelihood function of model $\mathcal{M}$ over the training data ;
4 $\hat{\boldsymbol{\theta}}_n \leftarrow \arg\min \ell_{\mathcal{M}}(-|\boldsymbol{y}^{\mathcal{T}})$;
5 $\hat{I}_n \leftarrow -\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ell_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y}^{\mathcal{T}})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}\right\} / n$;
`/* Computation of the test statistic: */`
6 $s_n \leftarrow$ the test statistic from Equation (12) using $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{y}^{\mathcal{V}}$;
`/* Computation of the p-value */`
7 $g(t,\boldsymbol{\theta}) \leftarrow$ the function g of Equation (13) ;
8 $\rho(u,v) \leftarrow$ the covariance function of Equation (14) using $g(-,\hat{\boldsymbol{\theta}}_n)$ and $\hat{I}_n$;
9 $\lambda_1, \ldots, \lambda_q \leftarrow$ the eigenvalues of $\rho(-,-)$ computed using Equation (17) ;
10 $Q(x) \leftarrow$ the approximated CDF for the theoretical distribution of the test statistic, given by Equation (16) using $\lambda_1, \ldots, \lambda_q$;
11 $p \leftarrow 1 - Q(s_n)$;
12 **return** p ;

The training set is used to compute $\hat{\boldsymbol{\theta}}_n$ and $\hat{I}_n$ (lines 3, 4 and 5 of the Algorithm 1). The test statistic is computed using the validation set and $\hat{\boldsymbol{\theta}}_n$ (line 6). The covariance kernel ρ is estimated by employing $g_{\hat{\boldsymbol{\theta}}_n}$ and $\hat{I}_n$ (lines 7 and 8). Computing the p-value of the test requires a numerical approach, as detailed in Appendix B. It involves computing the q biggest eigenvalues of ρ (line 9) and subsequently approximating the distribution of the test statistic (line 10). The approximated CDF is employed to determine the p-value of the test (line 11). When a significance threshold is provided by the user, the p-value leads to decide whether or not the null hypothesis should be rejected.

References

Anderson, T. W., & Darling, D. A. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. The Annals of Mathematical Statistics, 23(2), 193–212.

Anderson, T. W., & Darling, D. A. 1954. A test of goodness of fit. Journal of the American Statistical Association, 49(268), 765–769.

Billingsley, P. 1968. Convergence of probability measures. Wiley Series in Probability and Mathematical Statistics. Wiley.

Blanchet, J., Ceresetti, D., Molinie, G., & Creutin, J.-D. 2016. A regional GEV scale-invariant framework for intensity-duration-frequency analysis. Journal of Hydrology, 540, 82–95.

Bougadis, John, & Adamowski, Kaz. 2006. Scaling model of a rainfall intensity-duration-frequency relationship. Hydrological Processes, 20(17), 3747–3757.

Burlando, Paolo, & Rosso, Renzo. 1996. Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. Journal of Hydrology, 187(1), 45–64.

Coles, S. 2001. An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. Springer.

Courty, Laurent G, Wilby, Robert L, Hillier, John K, & Slater, Louise J. 2019. Intensity-duration-frequency curves at the global scale. Environmental Research Letters, 14(8), 084045.

Cramér, Harald. 1999. Theory of estimation. Princeton University Press. Pages 473–524.

Darling, D. A. 1955. The Cramér-Smirnov test in the parametric case. The Annals of Mathematical Statistics, 26(1), 1–20.

Durbin, J. 1973. Weak convergence of the sample distribution function when parameters are estimated. The Annals of Statistics, 1(2), 279–290.

Fleming, Thomas R., & Harrington, David P. 2005. Appendix B : an introduction to weak convergence. John Wiley and Sons, Ltd. Pages 331–341.

Gupta, Vijay K., & Waymire, Ed. 1990. Multiscaling properties of spatial rainfall and river flow distributions. Journal of Geophysical Research: Atmospheres, 95(D3), 1999–2009.

Haruna, Abubakar, Blanchet, Juliette, & Favre, Anne-Catherine. 2023. Modeling intensity-duration-frequency curves for the whole range of non-zero precipitation : a comparison of models. Water Resources Research, 59(6), e2022WR033362.

Huang, Qin, Chen, Yuanfang, Xu, Sui, Liu, Yong, & Li, Xinkai. 2010. Scaling models of a rainfall intensity-duration-frequency relationship. Pages 3415–3419 of: 2010 Sixth International Conference on Natural Computation, vol. 7.

Innocenti, Silvia, Mailhot, Alain, & Frigon, Anne. 2017. Simple scaling of extreme precipitation in North America. Hydrology and Earth System Sciences, 21(11), 5823–5846.

Jalbert, J., Genest, C., & Perreault, L. 2022. Interpolation of precipitation extremes on a large domain toward IDF curve construction at unmonitored locations. Journal of Agricultural, Biological and Environmental Statistics, 27, 1–26.

Kac, M., & Siegert, A. J. F. 1947. An explicit representation of a stationary Gaussian process. The Annals of Mathematical Statistics, 18(3), 438–442.

Koutsoyiannis, Demetris, Kozonis, Demosthenes, & Manetas, Alexandros. 1998. A mathematical framework for studying rainfall intensity-duration-frequency relationships. Journal of Hydrology, 206(1), 118–135.

Laio, Francesco. 2004. Cramér–Von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters. Water Resources Research, 40(9), W09308.

Lima, Carlos H.R., Kwon, Hyun-Han, & Kim, Yong-Tak. 2018. A local-regional scaling-invariant Bayesian GEV model for estimating rainfall IDF curves in a future climate. Journal of Hydrology, 566, 73–88.

Menabde, Merab, Seed, Alan, & Pegram, Geoff. 1999. A simple scaling model for extreme rainfall. Water Resources Research, 35(1), 335–339.

Petley, David. 2023. The failed dams in Wadi Derna in Libya. https://www.preventionweb.net/news/failed-dams-wadi-derna-libya. Published by the UNDRR.

Rodriguez-Sola, Raul, Casas-Castillo, M. Carmen, Navarro, Xavier, & Redano, angel. 2017. A study of the scaling properties of rainfall in Spain and its appropriateness to generate intensity-duration-frequency curves from daily records. International Journal of Climatology, 37(2), 770–780.

Schlesinger, Stewart. 1957. Approximating eigenvalues and eigenfunctions of symmetric kernels. Journal of the Society for Industrial and Applied Mathematics, 5(1), 1–14.

Sinclair, C. D., Spurr, B. D., & Ahmad, M. I. 1990. Modified Anderson-Darling test. Communications in Statistics - Theory and Methods, **19**(10), 3677–3686.

Smirnoff, N. V. 1937. Sur la distribution de ω^2 (criterium de R. von Mises). Rec. Math. Moscou, n. Ser., 2, 973–993.

Stephens, M. A. 1976. Asymptotic results for goodness-of-fit statistics with unknown parameters. The Annals of Statistics, 4(2), 357–369.

Sukhatme, Shashikala. 1972. Fredholm determinant of a positive definite kernel of a special type and its application. The Annals of Mathematical Statistics, 43(6), 1914–1926.

Van de Vyver, Hans. 2018. A multiscaling-based intensity–duration–frequency model for extreme precipitation. Hydrological Processes, 32(11), 1635–1647.

Yeo, Myeong-Ho, Nguyen, Van-Thanh-Van, & Kpodonu, Theodore A. 2021. Characterizing extreme rainfalls and constructing confidence intervals for IDF curves using scaling-GEV distribution model. International Journal of Climatology, 41(1), 456–468.

Zolotarev, V. M. 1961. Concerning a certain probability problem. Theory of Probability & Its Applications, 6(2), 201–204.