A non-parametric procedure to assess the accuracy of the normality assumption for annual rainfall totals, based on the marginal statistics of daily rainfall: An application to NOAA-NCDC rainfall database

Dario Ruggiu[1], Francesco Viola[1], and Andreas Langousis[2]

1. Department of Civil, Environmental and Architectural Engineering, University of Cagliari, Italy (dario.ruggiu@unica.it)

2. Department of Civil Engineering, University of Patras, Patras, Greece

Università degli Studi di Cagliari

UNIVERSITY OF PATRAS ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

1. Introduction

The statistical modelling of annual rainfall totals (ART) is fundamental in many fields of geosciences, in hydraulic design, and to support water management decisions. We focus on the normality assumption for ART samples, which theoretically arises from central limit theorem (CLT). In particular, the goal of our work is oriented towards creating a non parametric procedure to assess the accuracy of the normality assumption of ART, based on the marginal statistics of daily rainfall. The analysis has been conducted using the NOAA-NCDC daily rainfall database with global coverage, from which 3007 timeseries have been properly selected to ensure statistical significance of the results. Our work can be divided into three consecutive steps: a) use of goodness-of-fit statistics to classify NOOA-NCDC ART samples into two complementary groups: Gaussian (G) and non-gaussian (NG) distributed, b) implementation of a logistic regression analysis to identify the most influential daily and intra-annual marginal statistics of rainfall in determining convergence to the normal shape, and c) evaluation of a set of constraints using a random-search algorithm, which ensures reliable classification to N and NG groups.

2. Dataset and case study

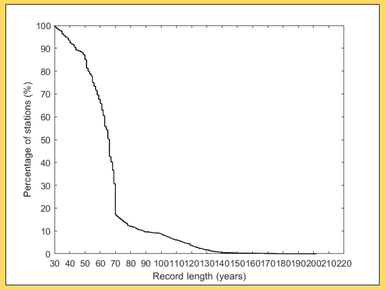

Figure 1. Percentage of the considered NOAA-NCDC stations exceeding different record length requirements.

From NOAA-NCDC Global Historical Climatology Network (GHCN) rainfall database (Menne et al., 2012), 3007 daily rainfall timeseries have been selected and analyzed, with: global coverage, percentages of missing data below 5%, yearly completeness above 98%, and more than 30 years of recordings (Figure 1 and 2).

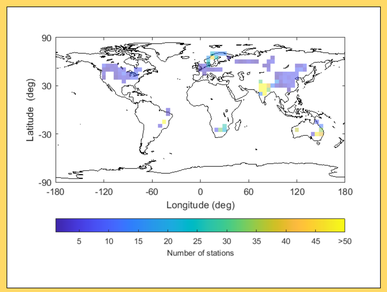

Figure 2. Spatial density of the 3007 NOAA-NCDC rainfall stations considered in the analysis.

References

Menne, M. J., I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston, 2012: An overview of the global historical climatology network-daily database. Journal of Atmospheric and Oceanic Technology, 29, 897-910.

3a) Goodness-of-fit tests

Three non-parametric goodness-of-fit test statistics for assessing the validity of the normality assumption (Kolmogorov-Smirnov, KS; Anderson Darling, AD; Cramer Von Mises, CVM), have been implemented to the 3007 daily rainfall timeseries aggregated at annual scale (ART timeseries).

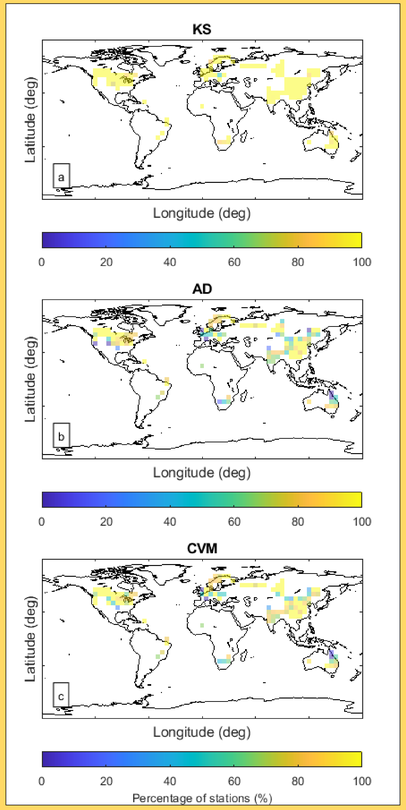

Figure 3. Global maps with spatial resolution 5 × 5 deg, illustrating the local fractions of stations that belong to group G (i.e. approximately Gaussian distributed annual rainfall totals), according to the three normality tests considered (Kolmogorov Smirnov, KS (a); Anderson-Darling, AD (b); Cramer Von-Mises, CVM (c)), at the 5% significance level .

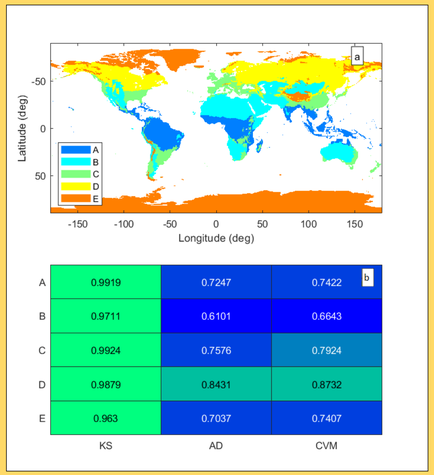

Figure 4. (a) Global map illustrating the Köppen – Geiger climate classification, featuring five distinct climate types: equatorial (A), arid (B), warm temperate (C), continental (D) and polar (E). (b) Fraction of stations with approximately Gaussian distributed ART (group G), on the basis of the three normality tests considered (Kolmogorov Smirnov, KS; Anderson-Darling, AD; Cramer Von-Mises, CVM), at the 5% significance level.

- In regions exhibiting continental climate (D), most ART samples can be effectively described by a normal distribution (AD 84.3% and CVM 87.3%, Figure 4b);
- In regions with arid climate (B) the gaussian distribution is not always the proper model to describe ART (AD 61.0% and CVM 66.4%, Figure 4b).
- The Kolmogorov-Smirnov (KS) test overestimates the percentage of gaussian distributed (G) ART samples (99% of the 3007 considered stations), relatively to Cramer-Von Mises (CVM) and Anderson Darling (AD) tests.
- CVM and AD tests indicate that 77% and 74% of the 3007 selected ART samples, respectively, approximate the gaussian distribution.
- Some spatial patterns of nomally distributed ART samples have been identified; see high density areas in Figure 3.

3b) Logistic regression

To detect the most influential local climatic properties in determining the approximate convergence of ART samples to the normal shape, we use logistic regression analysis between different sets of daily and intra-annual marginal rainfall statistics and the results of the Anderson-Darling test (α=0.05).

		f_{dd}	PCI	sk_{dw}	σ_{dw}
p-value	Set I	1.38E-18	--	3.74E-11	--
	Set II	8.69E-12	--	7.24E-10	1.32E-02
	Set III	6.004906e-06	5.502648e-03	7.057651e-11	--
VIF	Set I	1.000032	--	1.000032	--
	Set II	1.310212	--	1.024746	1.333897
	Set III	1.984055	1.985027	1.001464	--

Table 1. Results of the logistic regression analysis in the form p-values, and VIFs (Variance Inflation Coefficients), for the classification of the considered NOAA-NCDC rainfall stations into G and NG subsets, based on the Anderson-Darling test statistic at 5% significance level, and for three selected sets of predictor variables (Set I: f_{dd} and sk_{dw}; Set II: f_{dd} , sk_{dw} , and σ_{dw} ; Set III f_{dd}, PCI, and sk_{dw}).

Set I, namely the fraction of dry days f_{dd} and the skewness coefficient of rainfall in wet days sk_{dw}, , is the best performing one as indicated by the low p-values and a VIF value close to one; see Table 1.

3c) Random-search algorithm

Given a set of climatic predictors (Set I), the constraints S are tuned by optimizing the following objective function, using a random search algorithm, for different levels of significance α:

$$\max_{S}(P[A \cap T] + P[A^C \cap T^C])$$

- S = set of constraints;
- $P[A \cap T]$= joint probability of A and T events;
- $P[A^C \cap T^C]$ = joint probability of A^C and T^C events.

- A = approximate gaussianity;
- T= fulfillment of constraints S;

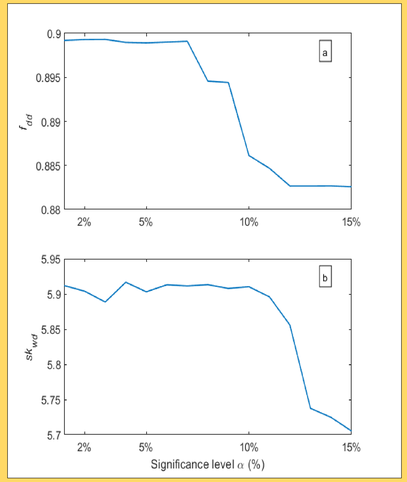

Figure 5. Dependence of the optimal thresholds of the fraction of dry days f_{dd} (a), and skewness coefficient sk_{dw} of rainfall in wet days (b), on the level of significance α of the Anderson-Darling (AD) test, for the 3007 NOAA-NCDC daily rainfall timeseries analyzed.

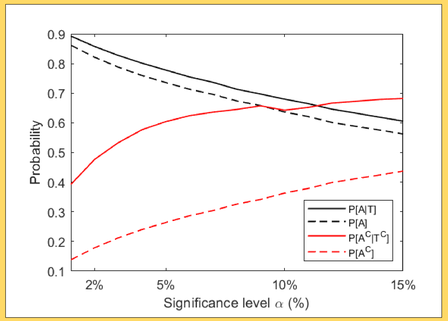

Figure 6. Comparison of the conditional probabilities $P[A|T]$ and $P[A^C|T]$, with the marginal probabilities $P[A]$ and $P[A^C]$= 1- $P[A]$, as a function of the level of significance α used for the Anderson-Darling (AD) test.

- When the level of significance increases, the optimal thresholds for the fraction of dry days f_{dd} , and the daily skewness coefficient tend to drop (Figure 5);
- Both conditional probabilities exhibit higher values relative to their respective marginals, indicating the significant information value of the non-parametric procedure (Figure 6).

4) Conclusion and future research

- Goodness-of-fit tests can be used to assess the approximate gaussian behaviour of ART samples in different climates;
- We developed a non-parametric procedure to assess the accuracy of the normality assumption for ART based on the marginal statistics of daily rainfall, particularly suited for hydrologic applications in data poor regions;
- Future research could be oriented towards assessing the extremal behavior of rainrates at different temporal resolutions, based on the marginal statistics of daily, monthly, and annual rainfall accumulations.