ISSN 2522-9400 **European Modern Studies Journal** *Vol 8 No 5*
available at www.journal-ems.com
DOI: 10.59573/emsj.8(5).2024.7

Employing an Interdisciplinary Approach and MATLAB Practicality in Studying the Water Flows in HPP Systems

Erarda Vuka [1], Dode Prenga [2]*
[1]Department of Information Technology, Mediterranean University of Albania
[2]Department of Physics, Faculty Natural Sciences, University of Tirana, Albania

Abstract. MATLAB is known for offering an intriguing computation environment for a large spectrum of applications and various levels of the users programming. In this work we have represented a technical analysis concerning the improvement of the forecasting capacities for water inflows in HPP structures and for the rivers that feed their dams in Albania. This work involves mathematical approaches and data-oriented analysis for a real system, where natural processes and anthropogenic activities are interlinked in a complex structure. In this regard, the codes written in MATLAB/Octave environment have been considered as technical tools for calculation without commenting comparison with other programming languages, but the deepness and usefulness of the results obtained straightforwardly throughout the work justify selecting it for facilizing a complex analysis. For this reason, we have highlighted and commented on the findings from the engineering aspect, revealing also the effectiveness of using very simple and practical codes that can be easily implemented and used by beginners.

Keywords: MATLAB, lakes' side in flows, simulation, stationarity distribution, q-Gaussian, multifractal

Introduction

The study of real and anthropogenic systems poses several technical challenges for the researchers. Their statistical description and the correctness of measuring on them require professional harmonization of modelling, regression, calculation techniques and simulation and deep interdisciplinary augments, which are not very likely to be fulfilled easily. Fortunately, the MATLAB environment can be used successfully for overcoming those difficulties, not only through its high profile and qualitative Toolboxes, but even by its simplicity and elegance of the programming milieu. In this work we want to underline both elements of the analysis for a nonlinear and heterogenous-data system, throughout a balancing of technical procedures of the research concerning the improvement of the analysis for a current system, and the practicability of the calculating environment used therein. We highlighted this second indirectly, by evidencing the deepness of the findings herewith. The system consists of the water flows in HPP lakes and the rivers where those settlements are built, which are highly nonlinear and heterogenous data series, and the calculations performance regarding them are of key importance. Concerning on strictness of the outcomes for the current analysis we appraised also the flexibility of the codding in this language that helped for achieving good results with little programming efforts. Because the codes written for this purpose were typically simple and considering that herein we mostly contemplated directly the functions of the language, we have skipped their representation in the edited form by admitting that methodically, mentioning the respective pseudocodes would be sufficient. Also, the details regarding syntax and other programming technicalities are considered irrelevant for describing hereto, leaving more space for the analysis of the system itself for clarifying their nonlinear and heterogeneous nature, which in turn, confirm the effectiveness of using MATLAB. As said, the system analyzed herein consists of a particular case of the time data series of the water inflows in the dam's cascade in the Drin River, Albania. This system is complex due to the

*Corresponding Author

pairing of natural heterogeneity of the precipitations and the dependency of the water flows and discharges from economic issues, urban area safety and risk hazard management. In those circumstances the improvement of prediction and forecasting for the water's quantities flows become an intriguing activity from a scientific and engineering point of view. Particularly, this typical feature invokes complexity and interdisciplinarity approach for their study, and advanced programming skills. Several methods and methodologies employed for improving the accuracy of measurement for their characteristic quantities can be clarified under the general analysis provided in the literature of the NaTech time series studies and their risk management, like (Wang et al, 2023; Qiang, et al, 2024; Plot et al, 2022) etc., but empirical methods are always a basic start for understanding the problem. In this case, the efficacity of MATLAB programming which is typically a destinated for scientific research and engineering application can be advanced by a rigor treatment of the features of the system under scrutiny. Following the idea elaborated in our recent works concerning similar analysis in (Vuka et al, 2018; Prenga & Vuka, 2014) etc., we have employed herein also for comparison and for guaranteeing the trustworthiness of our findings several algorithms written in this language provided by adequate literature (Ilhen, 2012; Liu, Zh 2022, Shimazaki, 2024) etc. The codes are run in a common desktop PC running Windows 10, without any supplementary specifics. For managing the processing time and memory space we used several auxiliary functions which are passed easily to each one grace of the flexibility of the language used. As result, we have realized the study of the system of the water inflows whose analysis needed for some calculating capacities with no effort and in a manageable time. In this regard, this work can contribute as an example for students and researchers on their early stage of academic advance how to do in deep analysis with little struggle by using MATLAB or its Linux based homologue, the Octave.

Data and Methods

In this work we have used a set of data for daily side inflows on three lakes of the Drin cascade covering the period [1998-2013], that are typically heterogeneous, covering a short period regarding prediction's capabilities of the models. The climate changes and the composite impact of human activities on the HPP's watersheds has become a challenging pursuit for the responsible bodies, mirrored on flooding of the Shkodra lowlands, water reserve misbalances etc. (Prenga & Sula, 2019). Considering those specifics and characteristics, the analysis is interesting and important from researchers' and management's point of view, and the accuracy of the outcomes is determinant for their usefulness. To begin with analysis, we notice that several methods and tools used for analyzing and predicting climate data series face challenges and a balance of advantages and disadvantages. A very basic but indispensable step for addressing this balance is the preliminary discussion of the general assumptions which are typically made when employing routine calculations. So, when measuring the characteristic quantity x on time series [0, T], given confidence level α_n, one need to calculate in principle

$x_{\alpha_n}^{T} = \bar{x}^{T} \pm n\sigma(\overline{x}) + \varepsilon_t$ where $\overline{x} \equiv \int_{support} x\rho(x)dx$, $\sigma(\overline{x}) \equiv \int_{support} (x-\bar{x})^2\rho(x)dx$ (1)

providing that the distribution of the stochastic variable (x) is stationarity! However, this last requirement is skipped in some imprudent assessment. The question is to estimate in what extend we can perform the measurement (1) if the distribution of x values is not stationary! So, we have observed in (Kovaçi & Prend, 2016) that the distribution of the raw daily average water inflows in one of the Dams of Drin cascades (Albania) was unstable for a period of several years, therefore, formula (1) is mathematically unapplicable! Physically speaking, the raw average calculated $< x = \frac{1}{T}\sum_{i=1}^{T} Daily.Inflows_i$ has not the significance of a statistical representative feature. Concerning on a based statement regarding averaging procedure and result, we might search for a more stationary daughter series for proceeding with a partial evaluation, or we must estimate the inaccuracy inflicted by the non-stationarity of the physical state of the system which is recognized by the time data series. To perform such a calculation the researcher must discuss and investigate thoroughly the stationarity issue. In this case, the fitting and regression opportunities of MATLAB are very helpful for a technician who is not professional in programming. Cleary this can be realized by every language, but in our case, the effort and codding abilities needed, are very basic. Another atypical measurement regarding the ordinary differential equations (ODE) used on time series modelling is analyzed in (Kushta et al, 2023), and (Prenga, 2024). In this case, by employing physics concepts, the regression procedures for models

$$Y = AX,\ \frac{dY}{dt} = AY, Y(t) = F(y_{t-T} \dots y_{t-1})) \tag{2}$$

where some matric elements must be estimated, are considered also as measurement activities. Again, preliminary precautions must be undertaken in order to grant meaningful implantation of the ODE dynamics. For our computation steps of the category (1) or (2), we must consider also the quality of the estimation of the characteristic's distribution, stationarity issue etc., regardless of protocols or assumptions, because predictions based on them contains significant uncertainty (Sula et al, 2019, 2016; Prenga, 2014). According to those studies, very intensive flooding has happened several times more frequently than predicted by standard protocols. For those reasons and others that we will show in the following, every calculation procedure is preceded by estimation of the non-stationary level and of heterogeneity measure. Basically, this process consists of a descriptive analysis aiming on the evidence of conditions for physical measurement of the key quantities, mostly the average side inflows, and when some predefined criterions are met, we procced with the measurement or a discussion about measuring process. Next, we investigated similarities in the dynamics of the time series of each lake and discussed for improving the prediction and forecasting for water inflows and their management. Therefore, we have considered a rigor histogram optimization based on corresponding function of the MATLAB, and next we employed also alternative calculation optimal bandwidth like (Shimazaki, 2024). The improvement on calculating the averages is considered also for examining the presence of extreme events. In this case, we analyzed the nature of the high-level inflows compared to the long-term mathematical average. Those events could be associated with extreme events and in this case, we have checked for the nature of the distribution. If the distribution does not follow the appropriate one, we have an indicator that those events are not extreme regarding the full data set, that suggests re-evaluating the long-term average. Next, we investigated the presence of a trend for averaged data regarding a meaningful period e.g., based on year, season etc., by simple regression. It would signal the presence of a long-term regime under the basic process of data generation. Also, by exploring the clustering and correlation distance between series, we could judge the nature of the data in terms of efficacity of using the same protocol for prediction forecasting and controlling of the water flows. In all those elements, the heterogeneity of the time data series is important. Initially we observed that standard tests based on unit root test have disqualified the opportunity of linearization. Instead, we proposed to measure the heterogeneity level, and following this measure, to consider an inaccurate measurement but not unexpectable one. This step is performed by employing multifractal analysis. We have disqualified a given daughter time series for use in modeling process if one observes discontinuity for the corresponding multifractal spectrum. For completing those analytic stopes before measuring a quantity, we have used several inner functions of the MATLAB regarding fitting, multivariate regressions, factorial analysis, the own made codes for Genetic Algorithm, Particle Swarm Optimization, Adjusted Monte Carlo Simulation, the Empirical Mode Decomposition and its MATLAB subvariants etc. Finally, if non-stationarity and the heterogeneity measure are not very significant, we have procced with measurement in the sense of this discussion, that is, evaluation of parameters in the models and discussion of their physical significance for our time series. We believe that by this procedure, we have avoided mechanical measurement and therefore we have improved the analysis for our system under scrutiny.

Remarks on General Behavior of the Extreme Event's Occurrences on Inflow's Series

The importance of enhancing calculating steps aiming on helping management bodies of the HPPs to maintain stable and controlled regime on the discharges from the Dams can be understood by a simple discussion regarding extraordinary precipitation. Firstly, we observe that during intensive prescription like the one that occurred in 2011, the final discharges are very large, and the consequences form the flooding of Shkodra lowland were very hard. Several other flooding indicates low predictiveness level. Oslo, according to (Prenga et al, 2019; Sula et al, 2018), the hourly behavior of the water level in the end gate of the cascade, the Vau i Dejes' Dam follow a log-periodic precursor, indicating harmful pairing of the operational activities, (Prenga et al, 2019). Notice that log- periodic precursor is not natural in the water level series of an HPP, because it signifies harsh magnitudes and the risk for very high ones. If those are present, it signals failures of the management of the waters, and consequently, intensive discharges. All those facts suggest that the models used for forecasting are not updated with recent changes in the climate behavior, watersheds' features etc. Notice that the goodness on prediction ais very important for those systems, not only for productive management of the waters during operational activities, but also for the safety of the lowland next to the dams. So, by assuming for hypothesis that flows of above 10,000 m3/sec could happen (it is predicted as one at 10,000 years for Fierza basin), it would fill all dams in few days (corresponding to around 5 milliard m3 for Fierza HPP, se Figure 1). Consequently, the discharges would be forced at this level; and the process would follow for the subsequent Dams. It is considered an uncontrollable scenario, but if the prediction and forecasting were highly accurate, the half-empty Dams of Koman and VDeje would provide several days barrier for flooding). Note also that the scenarios for the Danms' managements are very strict, but within our best knowledge, the calculation used for stipulating the rules therein are based on a lognormal distribution of the average inflows, that cannot be the real case due to several changes in the water system and cycles. Also, even though it is not clarified on the Regulation Protocol which we are not commenting or discussion herein, it is likely that calculations refer to theoretical distribution of the extreme value which is the Gumble distribution (Gumbel, 1935, 1940; Plots et al, 2000) etc., given by

$$f(x) = \frac{1}{\beta} exp\ ([-\frac{x-\mu}{\beta} + exp\left(-\frac{x-\mu}{\beta}\right)) \qquad (3)$$

with cdf $F(x) = e\,\mathrm{xp}\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right)$. Literally speaking, this is classical for rivers flows, referring to (Gumbel et al, 1940). However, based on theoretical arguments and the analysis for current system provided in (Prenga et al, 2024; Prenga et al 2019), time series of the side inflows covering [1998-2013] do not provide sufficient information regarding the long-term distribution, so the averages, and their variance calculated on those series are local and characterize this period only. Specifically, the distribution of the daily inflows are typically non-stationary distributions. (Prenga et al, 2019; Kovaçi et al, 2016)) etc. For all those arguments and facts, a careful analysis of the distribution for the inflows on the Dams consists of a key procedural step for improving knowledge of the averages, their variances and for enhancing the prediction of the events. In this regard, we must know also the stationarity or merely, the non-stationarity measure for the daily water inflows on each of the HPP's Dams.

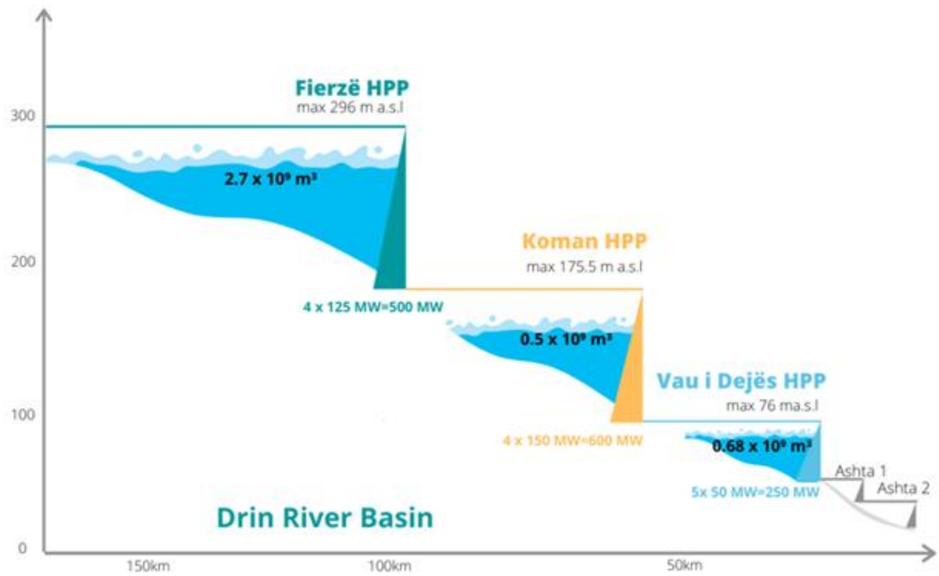

Figure 1. Drin HPP cascade. Next to the Vau i Dejes is Shkodra and Lezha lowland

Based on the suggestions of the relevant literature (Tsallis et al, 2000; Umarov et al, 2008), etc., and our previews works, we have utilized as q-Gaussian (qG) of the form

$$\rho(x) = \frac{\sqrt{\beta}}{\sqrt{\pi}} \frac{\sqrt{q-1*}\Gamma\left(\frac{1}{1-q}\right)}{\Gamma\left(\frac{3-q}{2(1-q)}\right)} (1-(1-q)\beta(x-\mu)^2)^{\frac{1}{1-q}} \quad \text{for } 1 < q < 3 \tag{4}$$

$$\rho(x) = \frac{\sqrt{\beta}(3-q)\sqrt{1-q}\,\Gamma\left(\frac{3-q}{2(1-q)}\right)}{2\sqrt{\pi}\,\Gamma\left(\frac{1}{1-q}\right)} * (1-(1-q)\beta(x-\mu)^2)^{\frac{1}{1-q}} \text{for } -\infty < q < 1 \tag{5}$$

Notice that a q-lognormal form is recommended in literature, but it is difficult to fit adequately, and also, the q-Gaussian theoretically is more general, as attractor of the non-stationary distribution (Tsallis, 2000; Umarov et al, 2008). Notice that a stable q-Gaussians corresponds to $1 < q < 1.667$, and the variance exist for $q < 2$. In this case it is an attractor of the unknown distributions according to (Umarov et al, 2008), that legitimates using it as a general form of the water inflows distributions herein.

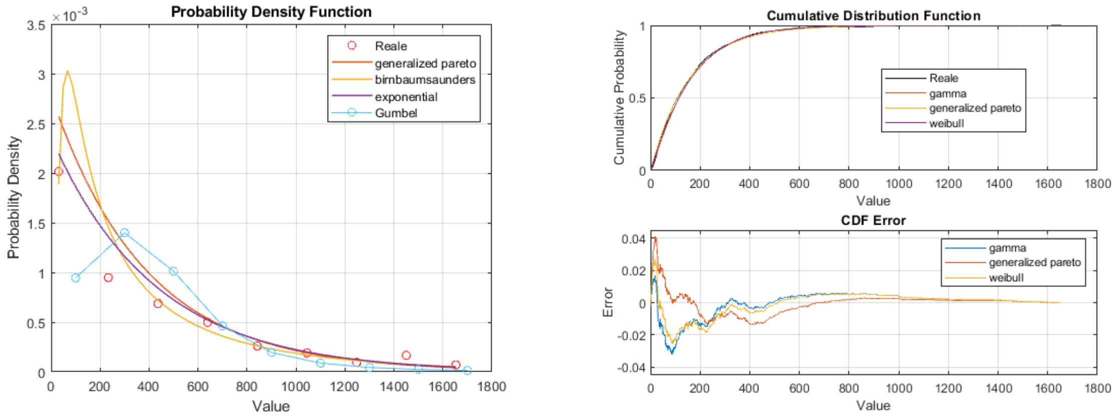

Figure 2. PDF for monthly maximum inflows. Illustration for Koman (left, pdf) and Fierza Lake (right, cdf)

Now the analysis is realized step by step. Firstly, we verified the Gumpel distribution for excessive inflows, but the fit was not significant. Notice that initially we have optimized the histogram as discussed above and performed the fit procedure for several basic distributions, and one qualify five of them according to their goodness of fit. So, initially we observed that tLocationScale which is included on the basic distribution in MATLAB has fitted better (we tried 17 distribution types) and recognizing that this practically the q-Gaussian, we have used the corresponding form (4) and (5) for further analysis. Notice that a q-lognormal form is recommended in literature, but it is difficult to fit adequately. Next, on measure the non-stationarity level, and according to it, we discussed the meaning of the characteristic parameters. So, for calculating the average inflows and their variances and considering them as references, we have employed the following shortened pseudocode

- *Read (data); operate a check for meaningful distribution using standard kernel (we used mostly epnanchkov)*
- *Identify the optimal histogram (we used mostly Freedman-Djaconic); calculate pdf and cdf correspondingly by using inner function*
- *Operate fitting routine for main distributions included in MATLAB routines*
- *Operate fitting routine for cfd and compare the goodness of fit; disqualify fitted distribution when pdf and cdf fit do not match*
- *Fit the q-Gaussian with optimal histogram by operating nonlinear least squares (we use mostly LM and Trusted Region option), PSO an GA; disqualify the fit if parameters do not match; identify the distance from stationary distribution*
    - *Remove fie last bins and rerun the procedure of identification of the distribution.*
    - *Look which of four best fitted distribution reproduce better the bin removed*
- *Conclude the report based on the distance from the stationarity*

In Figure 2 are shown for illustration the graphs of the best distribution fitted to the histogram based on the pdf and cdf approach for the daily of inflows on Koman and Fierza lake. The heterogeneity of the data and their small number impose difficulty on fitting routines, resulting in non-unique distribution obtained. This investigative exploration suggests precautions in the identification of the underlying distributions. Specifically, for the cases where the non-stationary indicator $d = q - 1 > \frac{2}{3}$, we conclude that the mean and its variance are typically local parameters which characterize current series only and do not represent the long-term mean. This finding is important however, because it out a limit between the correct statements and u supported ones. Herein we are not going into details for the causes that implicate the non-stationarity for the daily water inflows, but recognizing this factum, we might explore for alternative solution regarding a desired descriptive analysis. Particularly, we cannot use any of standard distribution like lognormal which is used in several forecasting protocols, because the non-stationarity of the distribution implicates that any parameters estimated by the fit would depend on current series, hence the long-term parameters are unknown. This why we have considered the importance of correctly identifying the q-parameter by the fitting process. Technically speaking, for achieving trustworthy results knowing the sensitivity of the cost function to the q-parameter in formula (1) and (2) above, we have refined the fitting procedure by using basic fitting and evaluating functions and nonlinear least square or partial LS regressions and employing PSO or GA algorithm. Also, we have practiced fitting based on pdf functions by arguing that in this case, local defects such as the empty bin presence for example are more influent on fitting process, which impede fake fitting. Norice that when we chose the cdf fit, we have considered also a special optimization technique proposed in (Stoenoiu et al, 2023), but since we didn't get improvement due to the empty bin presence on our histograms, we are not commenting on it herein. In fitting subroutines, we have used a logic criterion for accepting it: if the corresponding parameters obtained by two different calculation techniques match, the obtained q-Gaussian is qualified as representative. By employing this procedure, we can measure with good certainty the non-stationarity level, which is very important for the analysis herein. If q-Gaussian is qualified as the best fitted distribution, we might use the differences between the q-mean and q-variance obtained after the fit, with arithmetic mean and variance, as a measurable quantity.

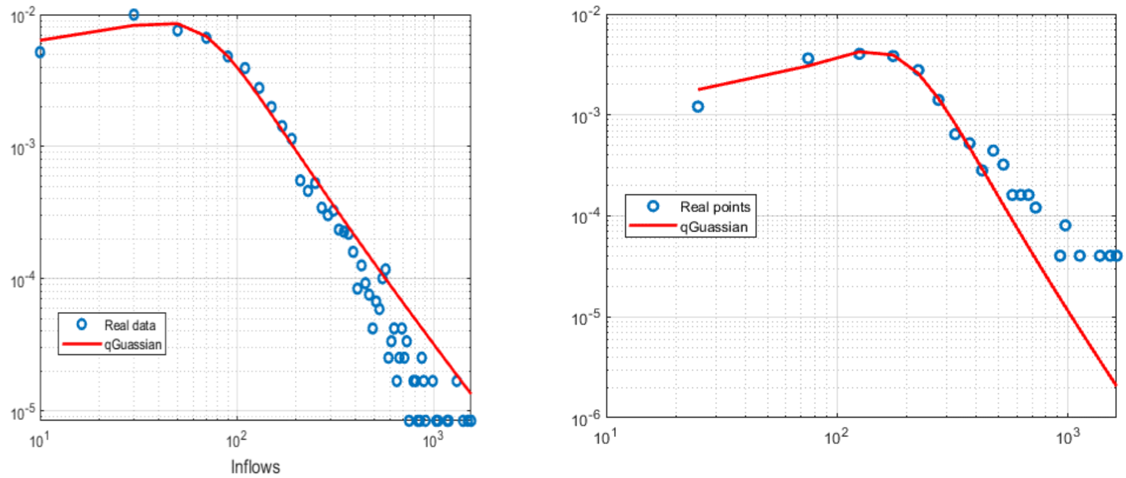

Figure 3 Distribution of daily inflows For Fierza and Koman

Better results are found when using an ad-hoc genetic algorithm and another Particle Swarm Optimization algorithm for fitting (5) to the optimal histograms, whereas nonlinear LS was unstable and didn't converge always. Also, this last was highly sensitive to the bin number for all the cases when preliminary investigation resulted in non-stationary distribution. For inflows of Fierza and Koman, the fit was very good as seen on loglog scale, whereas for Vau Dejes, the fit deteriorates for high values of the inflows. The key parameters (q) are $q(qG_{Koman}) = 2.0408$, $q(qG_{Fierza}) = 1.6285$and $q(qG_{VauDejes}) = 2.5187$. It followed that the distributions are highly non-stationary; therefore, the mean does not represent the long-term average. The distribution of Fierza inflows has definite variance (e.g. $\sigma^2 > 0$), and stationary distribution. The non-stationarity feature of the distributions measured directly by q-gaussians fit, disqualify the capacity of generating correct data by simulation, or by employing Monte Carlo techniques, because of the instability of the target distribution. To explore for measurement conditions and measurable variables, we have considered daughter series of the daily inflows for every month separately. Physically speaking, the cutting of the natural continuity of the processes by mounting monthly series is debatable and mechanical-like option, but on the other side, the seasonality nature of the time data series is supportive for this partitioning of the series. By the way it is not meaningless to talk about the historical average of precipitation/temperatures/inflows etc., during certain month's reference. After portioning the series in monthly fighter series, we observed that the q-Gaussians remain the best fitted PDF to the corresponding histograms but regarding to the non-stationarity measure by $d = q - 1$, there are remarkable differences between lakes inflows series (Table 2). So, all monthly-based data for the end-chain of HPP cascade- the VDeje dam - the corresponding distributions are highly non-stationary ($d > \frac{2}{3}$). Consequently, the mean and variance are again typically local, featuring our data set only. In this respect, we cannot use them for assessing probabilities of certain occurrences and for forecasting. This statement is valid regardless of the pdf function considered for the fit, including the ones used theoretically. Next, the distributions of the January series are stationary for the Koman and Fierza lakes' inflows (Table 2). Also, the distributions of the mixed inflows for the period considered are stationary in 3 cases. We concluded that, for those daughter series we can evaluate the probability of a certain event with good confidence. Respectively we can measure the average and its variance for February, March, June, July. For those subcases we can consider prediction too. From management point of view, we can be based on predictions and scenarios of those months to compensate missing informative prediction for other months. We conclude that additionally, for all series of months that exhibit stationary behavior, we can forecast inflows by using Neural Network tools.

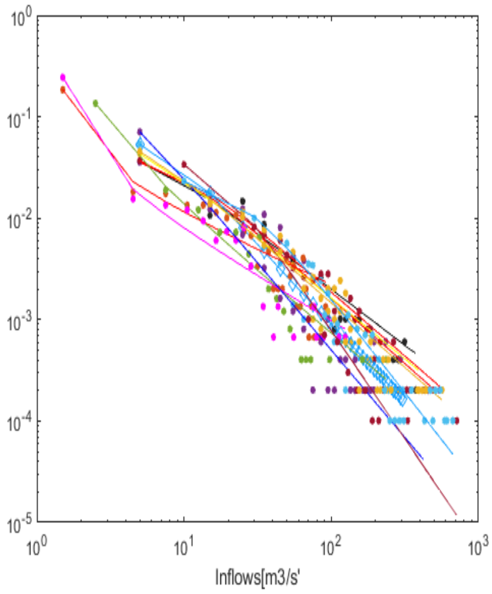

Table 1. Non-stationarity index for distributions of the side inflows

	Fierze	VD	Koman	All together
Jan	**1.18**	2.82	**1.63**	**1.62**
Feb	2.22	2.63	**1.19**	**1.29**
Mar	**1.21**	2.65	2.24	2.02
Apr	2.05	2.77	**1.20**	1.95
May	2.66	2.31	**1.27**	2.30
Jun	2.00	3(ND)	2.30	2.65
Jul	**1.19**	2.69	2.07	2.30
Aug	2.22	3(ND)	2.73	1.78
Sept	2.31	2.23	2.07	2.30
Oct	2.78	1.91	2.26	2.28
Nov	2.46	2.60	2.20	**1.41**
Dec	2.78	2.12	1.68	2.96

Figure 4. Nonstationary distribution -power law like-for VD inflows

The shortened pseudo codes look like following

- *Read the data, check for outlier*
- *Perform histogram optimization (Scott, Freedman Djaconic, MaxEntropy Algorithm etc)*
- *If (several empty bins) call the function based on entropy optimization, select the binning with lowest empty bins*
- *Check fitting of all the set of distributions-fitdist -Proceed with CDF and PDF fit and verify uniqueness Call GA, PSO and NLLS functions, evaluate and compare distribution parameters.*
- *If (not unique) exit,*
- *Measure non-stationarity* $S = q - 1$*; compare* $\Delta\mu = \mu_q$-μ, $\Delta\sigma = \sigma_q$-σ.
- *Display the parameters*
- *If* ($q < \frac{5}{3}$),
    - *Execute Monte Carlo subroutine: Makedist q-Gaussian based on parameters obtained, use as target distribution*
    - *Run Neural Networks for predicting scenarios*
- *end*

Evidenced Dissimilarities on the Trends of the Side Inflows in HPP's of Drin

By direct analysis of the stationarity for the time data series provided in standard statistical literature like (Walter et al, 2016) for example, we realized that daily inflows raw series and their differences have unit roots, so those series are not usable for linear modeling. This finding suggests the presence of a trend underlining their dynamics. Because those series are characterized by seasonality and short-term quasi cycles, we were interested for checking the presences of a potential long-term trend. The idea is very rude but meaningful because long term cyclic behavior is also typical for time data series of environmental and climate quantities. Alosd, we do not examine for short-term cyclic behavior because it is difficult to separate the trend from complex stochastic signal which compose the recorded data. Regarding the evidence of any 'cyclic behavior' of the seasonality type, the autocorrelation (ACF) can help, it is not the case for long term cycles, because in this case we must extrapolate beyond measured data. Alternatively, we proposed an empirical approach based on regression. In this case, the finding could be only suggestive, and more sophisticated tools and more data are needed to confirm the presence of the trend. So, we observe initially that through the years after 1998-2013 an increasing tendency appear on the annual averages of daily side inflows on the lakes, which suggest the trial empirical form as follow

$$I_F(t) = I_F^{Trend}(t) + x(t);\ I_K(t) = I_K^{Trend}(t) + y(t); I_K(t) = I_K^{Trend}(t) + z(t) \qquad (7)$$

where x,y,z are zero-mean processes and superscript ‘Trend’ clarify the first term. In this case the process that generate our data series has the form $X \sim PDF(q, \mu_q, \sigma_q, \beta)$, where $\mu_q$is the long-term average, q and β are parameters of the underlying distribution, and σ_q is the variance.

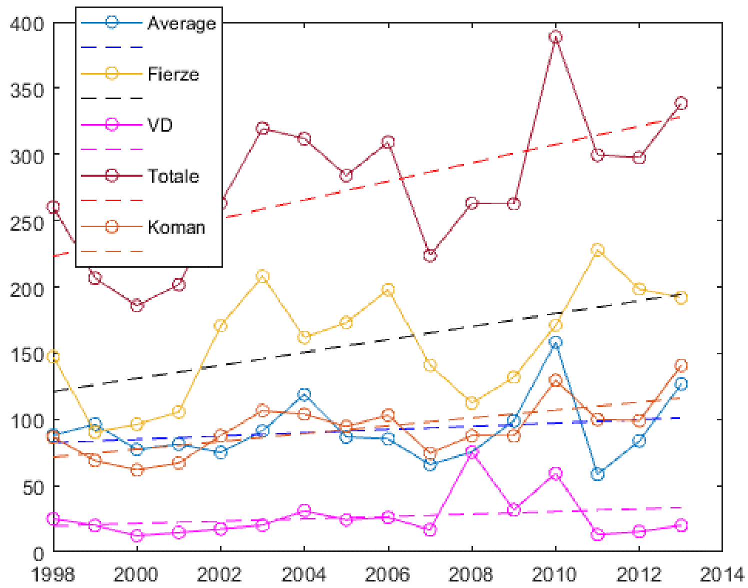

Figure 5. The observed trend on averaged data based on 4-month seasonality

A more realistic approach could be based on variable if the variance like in (Gubler et al, 2023) $\sigma_{q,m} = \sigma_q + \sum_{j=1}^{2} a_{1j}\cos\left(\frac{2\pi j}{n} * m\right) + \sum_{j=1}^{2} a_{2j}\sin\left(\frac{2\pi j}{n} * m\right)$, where m counts periods (let start with year) and n=1, 2,12 reproduce months. Based on similar arguments we can mimic empirically this idea and proposed for the first term of Fourier decomposition the following

$$I^{trend}(t) = I_{mean} + u_t, u_t = u_0 + a_1 \, cos\left(\frac{2\pi}{T}t + \varphi_o\right) + a_2 sin\left(\frac{2\pi}{T}t + \varphi_o\right)) \qquad (6)$$

referring to the yearly coordinate. If we assume that the variance is normally distributed $\sigma \sim N(0,1)$, this naïve approach make sense, taking also $I_{longTerm}^{average} \sim I_{mean}^{shortTerm} + u_0$. The same idea is used for tinner data, if a periodicity is logical, and in Figure 3 we have represented those averages based on 4-month reference, are grouped as [Dec-March, April-July, August-November]. Notice that to reduce high frequency fluctuations we have smoothed the original series by employing moving average setting empirically the lags=5 days. Also, for evidencing the nature of possible trends the series, we procced initially with empirical mode decomposition on original series of the daily inflows considering that in principle, last IMF mode report an empirical underlying periodicity comparable to the length of the series. The medium term-trend obtained by using the EMD diagnostics could be also a numerical coincidence, therefore, we have performed e sequential evaluation for series that finish n, 2n, ….; days before the original one. Here we chose n=30 days. It resulted that the period for the supposed underlying trends varies from one subseries to the other. This finding suggests that the near to periodic trend observed by EMD diagnosis could be result of the fluctuations of daily inflows. However, we obtained that longer series have longer length of the period. In this step we cannot accept or reject the presence of the medium-term periodicity for daily series, because of its dependency from the length of the series, but this observation is used as supportive for the initial assumption that if a trend really exists, it could be a long-term one. Based on this observation we expect that a longer averaging reference like the 4-month would preserve the long-term tendency whereas diminishing the effect to the short-term undulations. This trick is expected to reduce the uncertainty of regression (6). By repeating this intermediate analysis, we obtained also that the trend of the series depended on their length, but the differences have decreased. It mirrors the reduction of the number of points included for regression, but also, it suggests that a larger longer reference would be meaningful. This is not a solid argument thought, so we assumed the validity of (6) empirically in the sense that if a trend exists on the inflows time series, the calculation of (6) is acceptable. Finally, we have used the yearly averaged series. The result is shown in Table 3.

Table 3. Coefficients of the Empirical cyclic regression

	A(sin)	FreqCyclic	A(cos)	InitPhase	Threshold	Period	
Reference: 1 year							
Koman	125.43	0.0101	0	1.57	121.84	624.4	
Fierza	450.68	0.0101	188.97	1.61	100.04	624.2	
V.Deje	118.64	0.0098	8.6268	1.57	100.05	640.2	
TotaleDrin	619.39	0.0101	359.88	1.66	102.03	622.74	
Average	293.67	0.0101	0.7582	1.55	123.26	620.6	
4 months							
Fierze	81.7	0	142.7	0.9	100	1194.1	Dec-March
	75.8	0	132	1	100	1640	Apr-Jul
	408	0	787.9	2.5	993.6	5465.8	Aug-Nov
Vdeje	27.32	0.0134	0	3.14	100	470.4	Dec-March
	19.56	0.0108	3.608	2.32	100	582.6	Apr-Jul
	14.51	0.0108	0.2966	3.1149	100	579.2	Aug-Nov
Koman	24.19	0.0129	0	3.1416	126.0882	485.3	Dec-March
	14.12	0.0112	7.7645	3.0487	116.6989	561.5	Apr-Jul
	21.91	0.0106	0	3.1416	100	592.6	Aug-Nov

The pseudocode looks like

- *Read the data*
- *Perform unit root check, detrend the data and look for seasonality, return if yes*
- *Utilize the raw fit (6), evaluate parameters*
- *Run EMD we used mostly vmd function (forerun a q-Gaussian routine to estimate q-variance that is used in the vmd function);*
    - *if short term trend is verified, return, exit*
- *Conclude with reporting the period and basic inflows calculated (Table 3)*
- *End*

By this empirical approach seems to optimize the period ~620-640 years for the reference based on the yearly averages. This parameter changes significantly for references based on 4-month averages in the case of Fierza. Also, the supposed periodicity differs among groups of moths for this case. It suggests that empirical model (6) is not descriptive for this case but can be considered for the other two lakes. Noticing that Fierza lake is supplied partially by discharges from the Ohrid Lake in N.Macedonia, those differences can be paired with irregularities on those discharges during heavy precipitation last years. For two other lakes the periodicity behind the trend for two references considered is found at the interval 470 years - 640 years, suggesting that for this case it can be accepted as a possible scenario. The initial phase is nearly the same and produces a time shift of around 169 years. Also, this description identifies a near to similar threshold of the average flows of Drin at around 100-120 m3/sec. Those estimations are helpful in the sense of the range of terms of interests, once the cyclic behavior is supported by additional arguments out of the scope of this work.

A Comparative View of Correlation Distance and Clustering

When discussing side inflows of waters on artificial lakes, we must consider several factors that act independently or are interlinked to some extent, but practically hinder their relationship. While managing bodies may effectively adopt unique procedures for administering waters of communicating systems only if relationships are monotone, more precaution in reals scenarios. To shed some light on this idea, we checked for clustered behavior on several pairs of the series. For illustrating the idea, we have considered series of side inflows for month-based data and analyzed the correlation distance between series of daily inflows $[x_i,\ x_j]_{month}$ for two HPP's, using directly the corresponding function of the MATLAB. Initially we used the Euclidian correlation distance between two series given by $d_{x_i,x_j} = \sqrt{2(1-\rho_{x_i,x_j})}$ where ρ_{ij} is the correlation coefficient $\rho_{x_i,x_j} = \frac{cov(x_i,x_j)}{\sqrt{var(x_i)var(x_j)}}$. We observe that correlation distance varies by the months, figure 4. It suggests that the management should be based on local features of each series. Notice that the side inflows do not involve HPP's discharges, therefore we might compare also the pairing behavior for two basins which are not in contact (in Figure 4 are displayed the result for the pair Vaus i Dejes-Fierza). The highest correlation is obtained in the pair {Koman, Vau i Dejes}, that is expected because they are feed by watersheds belonging to similar terrains and neighbor area. Regarding this feature, the differences between each of those basins with Fierza is distinguishable as seen on the low correlation distance. Those differences are highlighted during the wet seasons and diminish during dry season, Figure 4. To advance on the pairing and similarity in the behavior, we have considered also the dynamic time warping check and clustering features, using the purposed function of the MATLAB. For this investigation we have compared series of the precipitation in Shkodra region during the same time, and inflows series on each lake, Figure 5, left frame. It resulted that the similarities are comparable for second semester of the year, whereas highest pairing is observed locally in periods December- February. The fact that the basin of VDeje has shown better paring is logical because the participation data were gathered from Shkodra station which is closer to this HPP than the others. The important finding to be underlined is the similarity observed for the second semester that suggests better predictability of the side inflows, by the intermediacy of the precipitation forecasting. Notice that side inflows depend on precipitation, characteristics of the terrain of the corresponding watersheds and their largeness etc., so it is very difficult to be predicted.

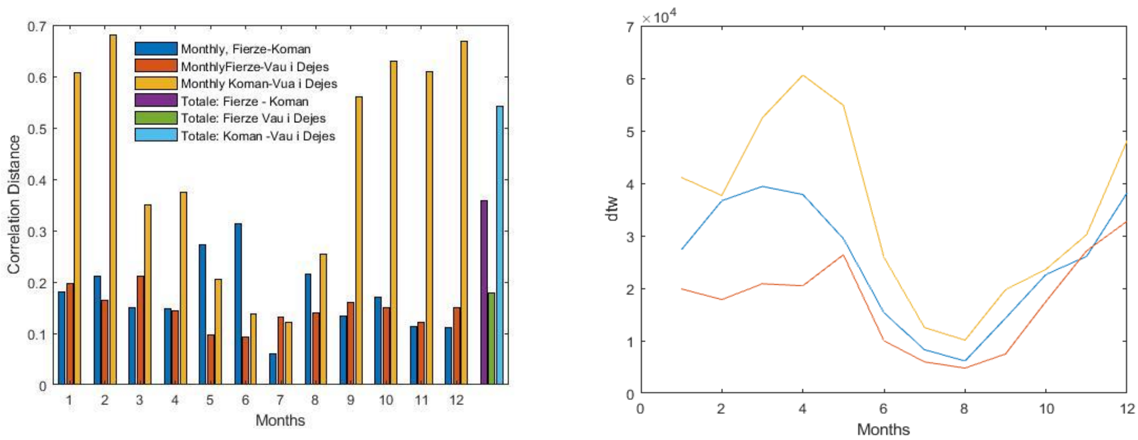

Figure 6. Correlation distance and dynamic time warping

By extending the analysis using clustering analysis, we discussed the centroids of clusters by the help of k-means or k-medoids. We used Manhattan difference for series of the same type, and Euclidian distance for different type series. When searching for clustering on a pair of two flows that logically do not have relationship cause-consequence, we have in mind a common cause that is based on terrain features [modified by human activities] that might cause similarity between them. In the pair [inflows, rain] we want to explore matching features. Notice that a monotone correlation is expected, but the inflow series are continuous whereas raining series have many terms null.

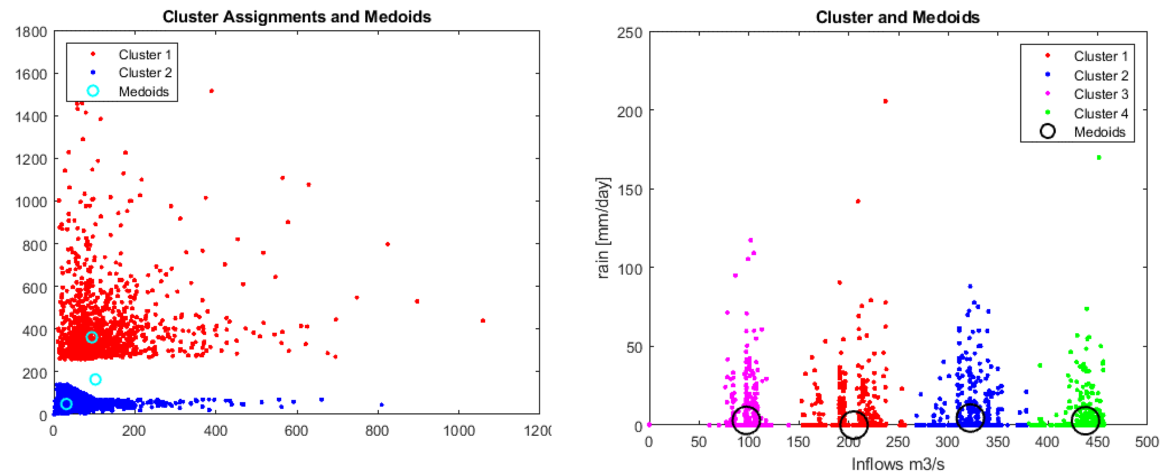

Figure 7. identification of clusters on pairing inflows/rain

In Figure (6) we observe that for the pairing [inflows at VD lake inflows, raining quantities in Shkodra region] there are 4 clusters, around centroids {205.3853 0}, {322.1477 4.5000}, {97.3459 3.0240} and {437.6960 2.8000}. The relationship is only indicative (we do not have the data for VD basin, but for a neighbor region), the presence of multiple centroids indicate that several factors are involved and act hiddenly.

Using Multifractal Heterogeneity to Improve Linear Modelling

When employing a model $Y = AX,$ where Y is an independent variable or even a derivative of X, assessment of the matrix parameters $[a_{ij}]$ can be seen as a kind of physical measurement. For modeling with them, the removing of the nonlinearity is realized by differencing. However, if series are highly heterogenous, the un-stationarity and nonlinearities cannot be avoided by first differences. Let's consider using Neural Network for forecasting series. Those algorithms are based on multiple linear relationships that can become insignificant or too mechanical when series are highly nonlinear. In this case we might suggest an approximative modelling procedure, by identifying initially the intervals where the time series behave as more relaxed. For this purpose, one might use multifractal properties, (McLeod et al, 2012); based on general literature, (Morales et al, 2012; Pavlos et al, 2014). For analyzing nonlinear time series X(t) with stationary increments obeying a scaling form $\langle |X_{n+\tau} - X_n|^q \rangle_n \sim \tau^{qH_q}$, where H_q parameter is Holder exponent, τ is the time lag and $q > 0$ is a parameter, and <…> designates averaging for all (n), see (Kantelhardt et al, 2002, Pavlos, 2012), etc., the multifractal spectrum given by

$$f(\alpha) = q\big(\alpha - H(q)\big) + 1 \tag{7}$$

can be very useful. Its smoothness can be considered as a visual display of continuity of the scaling behavior, and the un-smoothness indicates heterogeneity. Multifractal spectrum given by $f(\alpha) = q\big(\alpha - H(q)\big) + 1$ can be considered as a visual display of continuity of the scaling behaviour, and the un-smoothness, indicates heterogeneity.

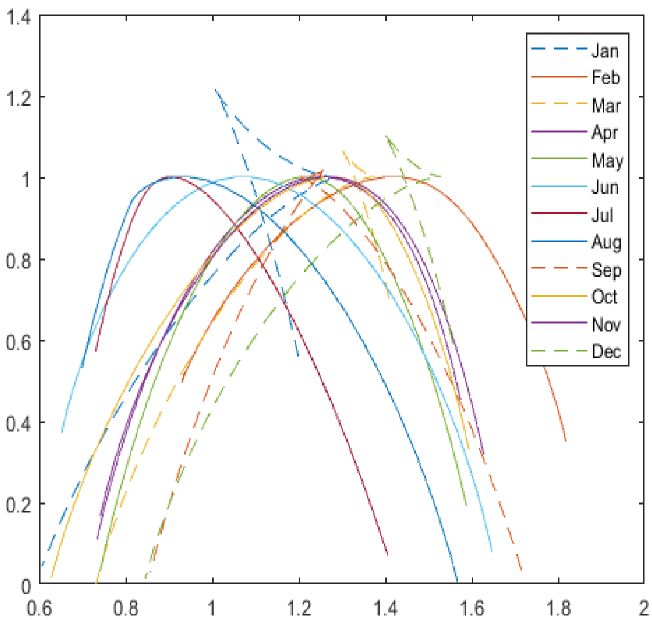

Figure 8. Multifractal spectrum for daily inflows in Fierza

Considering similar analysis used elsewhere (Prenga et al, 2024; Kushta et al, 2024), we assume that a low level of the nonstationary of the corresponding distribution and the smoothness of the multifractal structure can be acceptable condition for employing NN forecasting. We observe that smoothness of the MF spectrum and other MF indicators for different series differs significantly. For illustration we have shown in Figure 8 MF spectrum for monthly-based series of daily side inflows in Fierza Lake. Considering also the data from stationery of the corresponding distributions in Table 2, we proceeded with NN forecasting by checking the reproduction of the series for 15 days. The results are shown in Figure 9. For this calculation we have excluded the series with high heterogeneity identified herein throughout the discontinuity of the multifractal fractal spectrum, e.g, January, March, September and December. Notice that this procedure seeks to predict scenarios for the month of the next year and not the data for the incoming days. Also, this prediction could not be confused with forecasting probability of a given event using distribution obtained in section 2, but both calculations can improve each other. Once the conditions for a based application of the NN are fulfilled, we explored several starting times in this series to obtain the best reproduction of the recorded values, Figure 9.

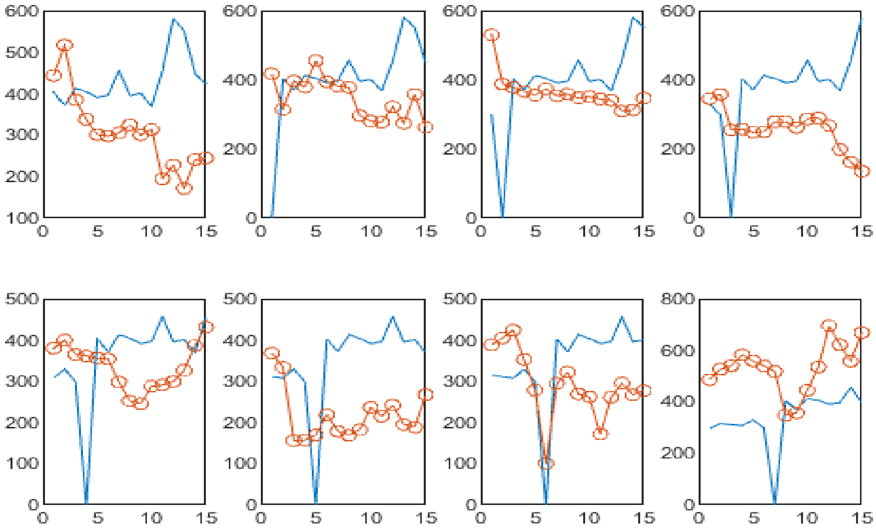

Figure 9. Reproduction of the inflows for January series attempting intervals of 15 days, the calculations are illustrative based on simplest network with only two layers and using trainlm algorithm (the fasted)

Conclusions

Based on a in interdisciplinary approach and using very simple routines in MATLAB we were able to perform a substantial analysis of the statistical features of water inflows for the Drin cascade, overcoming the problems imposed from the heterogeneity of the data and limited length of the time series. Starting from a methodological improvement intention, we achieved on concrete and useful for the system considered as a case study. Specifically, climate changes suggest a careful review of basic rules of the management of loads and discharges stipulated in the Original Regulation, based on fresh data and historical records. Herein we have discussed from a scholarship point of view the features of side inflows data series for three HPP lakes of the Drin River. Using a limited data set that we had at the time of this work, we have recognized those series as highly nonlinear, heterogenous and nonstationary, therefore, forecasting, prediction and simulation needs for specifical tools and methodologies. It resulted that series of daily inflows exhibit stationary distribution for few months only, limiting the legibility of modelling on these references. Methodically this work reveals the importance and easiness of using MATLAB to achieve.

References

Azure, I., Wiredu, J. K., Musah, A., & Akolgo, E. (2023). AI-Enhanced Performance Evaluation of Python, MATLAB, and Scilab for Solving Nonlinear Systems of Equations: A Comparative Study Using the Broyden Method. *American Journal of Computational Mathematics*, *13*(4), 644-677. https://doi.org/10.4236/ajcm.2023.134036

Denaj, A., Prenga, D., & Tahiri, V. (2023, September). General features of the time data series of Covid-19 in Albania. In *AIP Conference Proceedings* (Vol. 2872, No. 1). AIP Publishing.

Geraskin, P., & Fantazzini, D. (2011). Everything you always wanted to know about log-periodic power laws for bubble modelling but were afraid to ask. *The European Journal of Finance 19*(5), 366-391.

Gubler, S., Fukutome, S., & Scherrer, S. C. (2023). On the statistical distribution of temperature and the classification of extreme events considering season and climate change—an application in Switzerland. *Theoretical and Applied Climatology*, *153*(3), 1273-1291. https://doi.org/10.1007/s00704-023-04530-0

Gumbel, E.J. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de l'Institut Henri Poincaré*, *5*(2), 115–158.

Gumbel, E.J. (1941). The return period of flood flows. *The Annals of Mathematical Statistics, 12,* 163–190.

Ihlen, E.A.F. (2012). Introduction to multifractal detrended fluctuation analysis in Matlab. *Front. Physiol.*, *3*, 141. https://doi.org/10.3389/fphys.2012.00141

Innerberger, M., & Praetorius, D. (2023). MooAFEM: An object oriented Matlab code for higher-order adaptive FEM for (nonlinear) elliptic PDEs. *Applied Mathematics and Computation*, *442*, 127731. https://doi.org/10.1016/j.amc.2022.127731.

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., & Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, *316*(1-4), 87-114.

Kotz, S., & Nadarajah, S. (2000). *Extreme Value Distributions Theory and Applications*. Department of Engineering Management and Systems Engineering the George Washington University, Washington DC, USA School of Mathematical Sciences, The University of Nottingham, UK.

Liu, Zh, Peng, D., Zuo, M., Xia, J. & Qin, Y. (2022). Improved Hilbert-Huang transform with soft sifting stopping criterion and its application to fault diagnosis of wheelset bearings. *ISA Transactions*, *125*, 426-444.

McLeod, I., Yu, H., & Mahdi, E. (2012). Time Series Analysis with R. In T. S. Rao, S. S. Rao, & C.R. Rao (Eds.), *Handbook of Statistics* (Vol. 30, pp. 661-712). Elsevier. https://doi.org/10.1016/B978-0-444-53858-1.00023-5.

Morales, R., Di Matteo, T., Gramatica, R., & Aste, T. (2012). Dynamical generalized Hurst expo-nent as a tool to monitor unstable periods in financial time series. *Physica A: statistical mechanics and its applications, 391*(11), 3180-3189. https://doi.org/10.1016/j.physa.2012.01.004.50.

Pasdar, A., Hossein, A., & Kazemi, R. (2020). A Comparative Study on the Efficiency of Compiled Languages and MATLAB/Simulink for Simulation of Highly Nonlinear Automotive Systems. *Journal of Applied and Computational Mechanics, 8.* https://doi.org/10.22055/jacm.2020.33053.2137.

Pavlos, G. (2012). Complexity in Theory and Practice: Toward the Unification of Non-equilibrium Physical Processes. *Chaotic Modeling and Simulation (CMSIM), 1*, 123-145.

Pavlos, G. et al. Universality of Tsallis Non-Extensive Statistics and Fractal Dynamics for Complex Systems. arXiv:1203.5556 [nlin.CD]

Pavlos, G. P., Karakatsanis, L. P., Xenakis, M. N., Pavlos, E. G., Iliopoulos, A. C., & Sarafopoulos, D. V. (2014). Universality of non-extensive Tsallis statistics and timeseries analysis: Theory and applications. *Physica A: Statistical Mechanics and Its Applications*, *395*, 58–95. http://doi.org/10.[8] 6/j.physa.2013.08.026.

Plot, E., Balouin, Y., Ferreira, S., Mathilde, B. A. (2022). monitoring-based management approach for Natech-related risks: reflection from a case study. *ESReDA Seminar on Technological disruptions triggered by natural events: identification, characterization, and management,* Politecnico di Torino, Sep 2022, Torino, Italy.

Prenga, D. (2014). Using q-distributions on the study of side inflows for Koman basin in the Drin River, Albania. *International Journal of Engineering and Technical Research (IJETR), 2*(10), 77-81.

Prenga, D., & Ifti, M. (2016, March). Complexity methods used in the study of some real systems with weak characteristic properties. In *AIP Conference Proceedings* (Vol. 1722, No. 1, p. 080006).

Prenga, D., & Vuka, E. (2014). Complex systems method approach to the ECG analysis. *Int. J. Eng. Tech. Res*, *2*(11), 189-194.

Qiang, C. Junhong, L., & Khu, S. (2024). Environment Risk Management System in EU Natech Risk Management and Its Enlightenment to China[J]. *Research of Environmental Sciences, 37*(5), 1162-1168. http://doi.org/10.13198/j.issn.1001-6929.2024.01.22

Shimazaki, H. (2024). Histogram Binwidth optimization MATLAB Central File Exchange.

Stoenoiu, C. E., & Jäntschi, L. (2023, August). Least Squares for Generalized Gauss–Laplace Distribution of the Error in Certain Nonlinear Regressions with Perpendicular Offsets. In *International Conference on Mathematical Modeling in Physical Sciences* (pp. 153-162). Cham: Springer Nature Switzerland.

Sula, S., & Prenga, D. (2016). *Improving the Analysis of Hydrologic Time Data Series in our basins and economic impact of hydro-industries*. ISTI Faculty of Economy, University of Tirana.

Sula, S., & Prenga, D. (2019). A Case Study of Hydrometric Variables in the Lakes of Drin Cascade, Albania. *Bulletin of Natural Sciences, 28*, 205-217.

Tsallis, C. (2009). Nonadditive entropy and nonextensive statistical mechanics-an overview after 20 years. *Braz. J. Phys., 39,* 337–356

Umarov, S., Tsallis, C., & Steinberg, S. (2008). On aq-central limit theorem consistent with nonextensive statistical mechanics. *Milan Journal of Mathematics*, *76*(1), 307-328.

Vuka, E., & Petritaj, O. (2018). A Review on Traditionally ETL Process for Better Approach in Business Intelligence. *RTA-CSIT*, 17-23.

Walter, C., & Le Courtois, O. (2016). Lévy Processes and Extreme Value Theory. François Longin. *Extreme Events in Finance: A Handbook of Extreme Value Theory and its Applications, 1.* John Wiley and Sons, 2016, 9781118650196. https://dx.doi.org/10.1002/9781118650318.ch8

Wang, J., & Weng, W. (2023). Methodology for Natech coupling risk assessment using correlative multi-criteria decision-making method. *Journal of Loss Prevention in the Process Industries*, *83*, 105060. https://doi.org/10.1016/j.jlp.2023.105060.

Witkovský, V. (2023). Characteristic Function of the Tsallis q-Gaussian and Its Applications in Measurement and Metrology. *Metrology*, *3*(2), 222-236. https://doi.org/10.3390/metrology3020012

Xu, C., Ke, J., Peng, Z., Fang, W., & Duan, Y. (2022). Asymmetric Fractal Characteristics and Market Efficiency Analysis of Style Stock Indices. *Entropy, 24,* 969. https://doi.org/10.3390/e24070969

Zuway, M., Amani, F., & Shatshat, M. (2023). Effectiveness of MATLAB and Neural Networks for Solving Nonlinear Equations by Repetitive Methods. *Indian Journal of Computer Graphics and Multimedia*, *3*(2), 1-7. https://doi.org/10.54105/ijcgm.H9683.083223