EGU General Assembly 2024

Session HS3.8
Advances in stochastic analysis, modelling, simulation and prediction for hydrological and water-related processes

EGU24-8580

Detection of the spatial clustering mechanisms of streamflow extremes in the USA and relevance to flood insurance data

[1*]Konstantinos Papoulakos, [1]Theano Iliopoulou, [1]Panayiotis Dimitriadis, [2]Dimosthenis Tsaknias and [1]Demetris Koutsoyiannis

[1]Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zografou, Greece

[2] Independent researcher, Greece

* Corresponding author. E-mail address: papoulakoskon@gmail.com

Abstract

During the last decades, scientific research in the field of flood risk management has provided new insights and strong computational tools towards the deeper understanding of the fundamental stochastic behaviour that characterizes such natural hazards. Flood hazards are controlled by hydrometeorological processes and their inherent uncertainties.

Historically, a high percentage of flood disasters worldwide are investigated regarding the aggregated number of the affected people, economic losses, and generated flood insurance claims.

In this respect, the recently published National Flood Insurance Program data by the Federal Emergency Management Agency may yield novel perspectives into flood impacts.

The objective of this study is to conduct a spatial analysis on the daily flow series within the US-CAMELS dataset.

Specifically, we seek to identify spatial clustering mechanisms of over-threshold streamflow extremes, considering them as proxies for collective risk, in order to examine their underlying stochastic structure.

Furthermore, we explore their relevance to the actual insurance data and develop some additional stochastic modelling approaches.

US-CAMELS dataset

This analysis is applied on the US-CAMELS dataset, which comprises of **671 daily streamflow time series** from catchments in the contiguous United States (CONUS) that are **minimally impacted by human activities** (Newman et al., 2014).

From this dataset, **360 streamflow time series** with the maximum temporal overlap (namely, 35 years from 1980 to 2014) and less than 10% of missing values **were selected**. Figure 1 shows the study area and stream gauge locations for the full dataset including the finally selected 360 stream gauge locations.

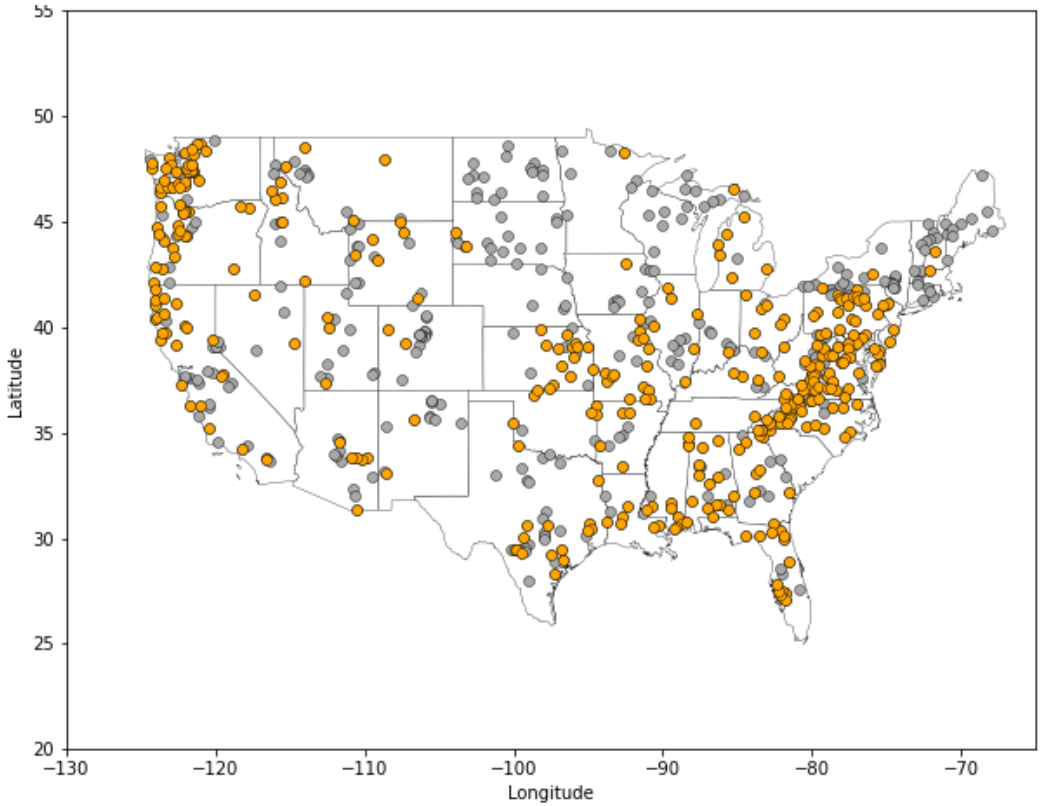


Fig. 1 The 671 US-CAMELS stream gauge locations. The selected 360 US-CAMELS stream gauge locations are colored orange.

FEMA's NFIP claims records dataset

EGU General Assembly 2024

Federal Emergency Management Agency (FEMA) published in **2019** the National Flood Insurance Program (NFIP) data, including more than **two million claims records** dating back to 1970 and more than **47 million policy records** for transactions (FEMA, 2019).

It is evident that this is a **giant contribution** for supporting scientists and policy-makers on their research on **how** the National Flood Insurance Program (NFIP) works, **where** flood damage occurs, and **what** the costs are.

Methodology: Extreme value analysis (EVA) distributions

EVA is widely used and applied as a tool to analyze and study statistics on sample values that **deviate extremely** from the mean of the full sample, in order to develop a deeper understanding of the sample and **precise modeling strategies**. It generates significant applications across many scientific fields such as **hydrology**, **insurance** and **finance** and can be also used to predict the occurrence of rare events, such as **extreme flooding**, large **insurance losses**, crashing of the **stock market** and many others (Reis and Thomas, 2007).

Methodology: Threshold selection

Threshold selection is a challenge in insurance and especially in flood insurance practices (Robinson and Botzen, 2020). The threshold should be chosen such that **all losses above the threshold are "extreme losses"** in the sense of the underlying extreme value analysis.

On one hand, we want to choose a **high threshold** in order to investigate the behavior of the (really) extreme events. On the other hand, for the estimation of the parameters in the distribution of the extreme losses, we need **many observations** above the threshold to create a solid statistical foundation for our conclusions, based on a long sequence of values.

Methodology: Threshold selection

In order to characterize the **dynamics of extreme streamflow values**, this study performed a POT analysis using **four** different percentage **thresholds**, i.e., 90%, 95%, 98%, and 99%.

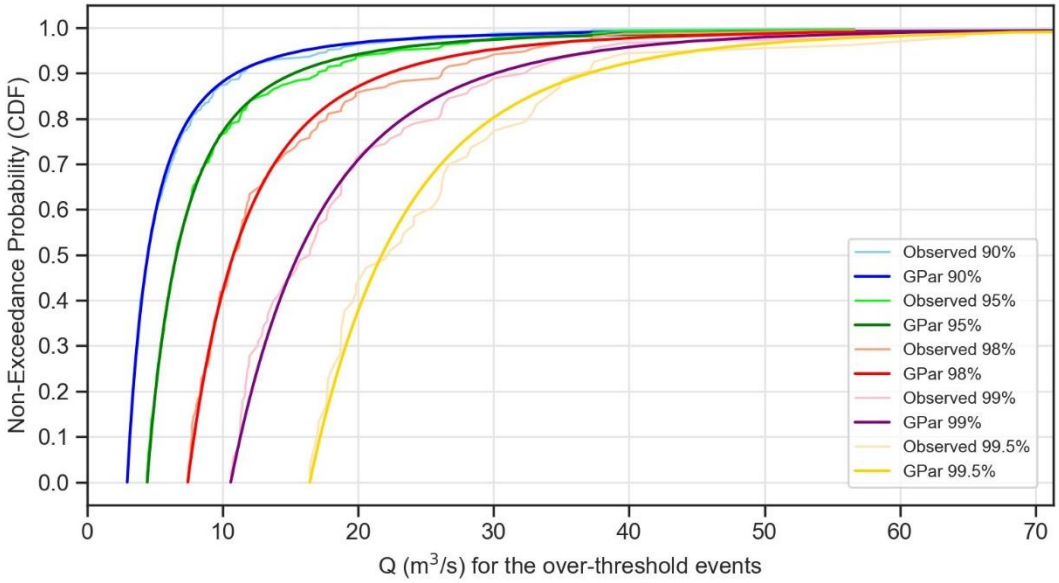


Fig. 2 Diagram that shows the impact of threshold selection on Non-Exceedance Probability (CDF) of streamflow of the over-threshold events regarding the observed streamflow records as well as the ones that were developed by the process of fitting these observed data with the generalized Pareto distribution. Gauge ID: 01552500.

Methodology: Collective risk model in insurance

The distribution of **total claim amounts**, considering the insurance company's portfolio as a collective that produces a random number N of claims in a certain time period, can be described by the **collective risk model** (Kaas et al., 2008).

Collective risk S_x is defined as

$$S_x = X_1 + X_2 + \cdots + X_N, \tag{1}$$

where X_i is the i^{th} claim amount during a certain time period, e.g. a year. Apparently $S_x = 0$ if $N = 0$.

Methodology: Collective risk model in flood insurance

EGU General Assembly 2024

Similarly, regarding **flood insurance** practices and in case of an extreme flood event, the collective risk *S* is the **total claim amount**, considering again the portfolio of (re)insured properties as a collective that produces a random number *N* of claims in a certain time period of **one year** in our case.

Methodology: Collective risk model in flood insurance

Denoting the records y_t of a time series, a **proxy of temporal collective risk *S*** is defined by Serinaldi and Kilsby (2016) as

$$S = \sum_{j=1}^{N} Y_j \tag{2}$$

where Y_j is the jth claim amount proxy (over-threshold flow fluctuation severity). Again, the total claim amounts $S = 0$ if $N = 0$. The definition of collective risk regarding flood insurance practices is a proxy of the actual collective risk, as it involves **hydrological series** and not actual claim amounts. In this study, regarding the aforementioned proxy of temporal collective risk, we use the term **Proxy Aggregated Losses *S***.

Methodology: Sequence of independent variables

In order to characterize the dependence and the clustering mechanisms, it is important to quantify how the time series differs from a sequence of independent variables.

A widely used method to create a sequence of independent variables is to **shuffle** (randomize) the series in order to get a new series which has the **same marginal distribution** but **no correlation**; the quantification of the distance between the independent and the observed variables is performed by **comparing specific characteristics**, i.e. the annual Proxy Aggregated Losses, the duration of the peak-over-threshold events and the occurrence frequency of return periods in the original time series and in the shuffled one.

Hence, in order to assess the **clustering of extremes** of the 360 observed time series, 100 new shuffled time series were **reproduced** for each one of the 360 original time series.

Methodology: The Hurst – Kolmogorov dynamics

The exhibited **persistence** in many **natural processes**, including streamflow and rainfall dynamics, is known as the Hurst phenomenon or **Hurst-Kolmogorov (HK) dynamics** and is quantified by the Hurst coefficient *H*.

In order to calculate the Hurst coefficient *H* and detect the potential **long-term dependence** (or else persistence, clustering) of a process, the most accurate method is by formulating the ***Climacogram*** (Koutsoyiannis, 2010), which has been shown to outperform estimators based on the autocovariance and power-spectrum (Dimitriadis and Koutsoyiannis, 2015).

Methodology: Generalized-HK (GHK) process

In some cases, such as in this study, fitting of straight line in the Climacogram derived from the observed data cannot capture the full variance behavior of the process at the whole range of scales. Thus, the generalized-HK (GHK) model is applied.

The **generalized-HK (GHK) model** is applied, which exhibits also an HK behavior in large scales but has **more flexibility** in **smaller scales** (Dimitriadis and Koutsoyiannis, 2018; note that a more advanced scheme has been introduced that can preserve any number of moments; Koutsoyiannis and Dimitriadis, 2021).

The ***Climacogram*** of the **GHK model** is the following, where the Hurst coefficient H is bounded between zero and one inclusive, q is positive, while λ and q have dimensions [x2] and [T], respectively:

$$\gamma(k) = \frac{\lambda}{(1 + k/q)^{2-2H}} \quad (3)$$

Results:
Impact of clustering mechanisms on the correlation between the Average Loss and the number of over-threshold events

EGU General Assembly 2024

Clustering mechanisms

Impact of clustering mechanisms on the correlation between the Average Loss and the number of over-threshold events

A common assumption in the computation of collective risk is the **independence** between *Average Yi* and Number of over-threshold events *N*. Here, losses *Yi* denote the flows exceeding the selected threshold and *N* is the number of such exceedances (number of events) over 365-day time windows. The relationship between *Average Yi a*nd *N* is emerged by the **Spearman**, **Pearson** and **Kendall** correlation coefficient between *N* and the average value of the **over-threshold** flow intensities.

$$\frac{1}{N}\sum Y_i = \frac{S}{N} = \textit{Average Yi} \quad (1)$$

Clustering mechanisms

Impact of clustering mechanisms on the correlation between the Average Loss and the number of over-threshold events

Insurance companies' concern about this correlation factor is noteworthy, as they try to investigate the **dependence** between the annual number of **extreme** events and the provoked *Average Yi*, which is a proxy of the **average claim amounts** per over-threshold event on a specific region. Introducing this parameter, **Pearson**, **Spearman** and **Kendall** correlation coefficient for the 4 selected thresholds are investigated.

Fig. 1 Cumulative histogram curves of the Pearson, Spearman and Kendall correlation coefficient between Average Yi and number of over-threshold events N for the 360 selected gauge locations and for all thresholds.

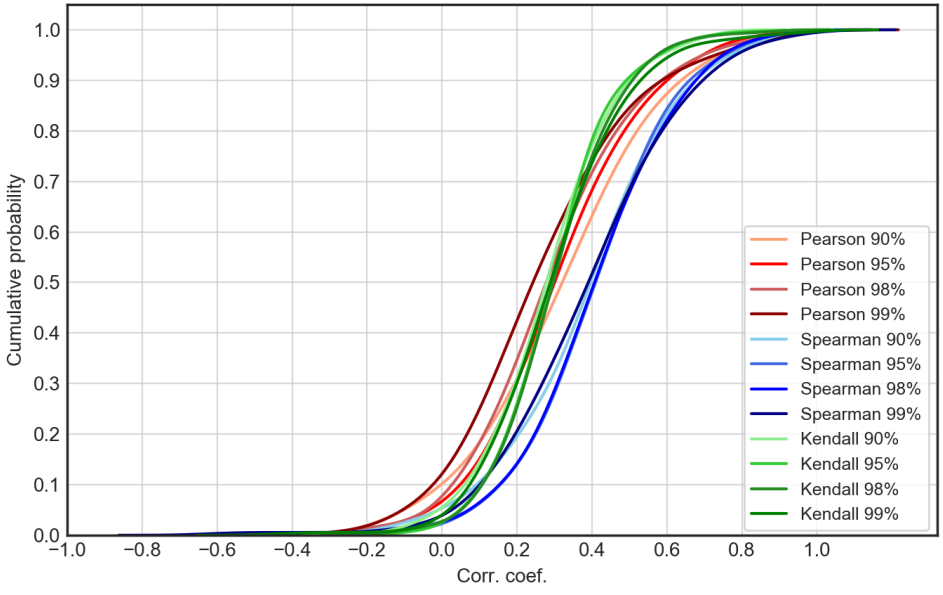
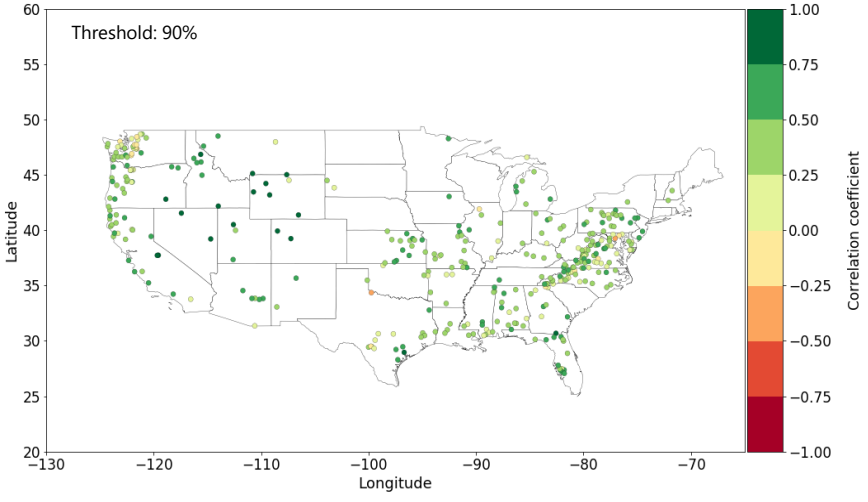
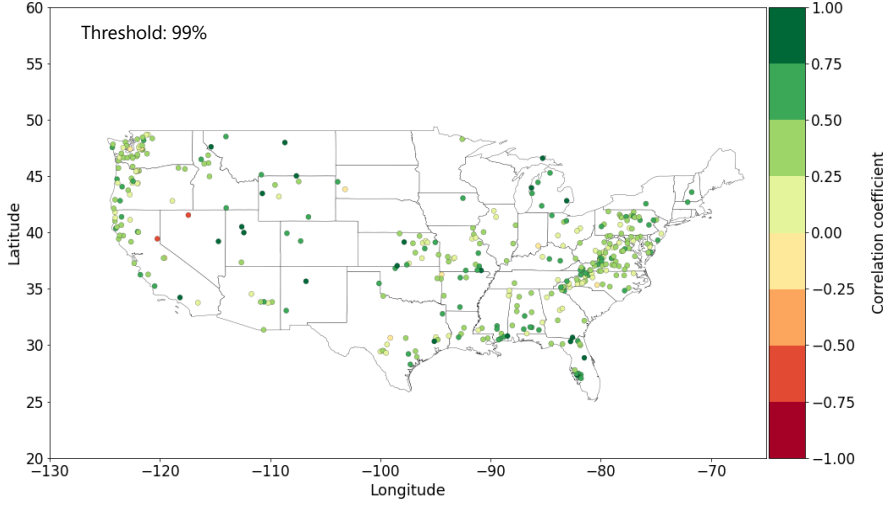


Figure 1 shows **cumulative histogram curves** of the Pearson, Spearman and Kendall correlation coefficient between *Average Yi* and number of over-threshold events *N* for the 360 selected gauge locations and for all thresholds. This study evaluates the **Spearman correlation coefficient**, as it is considered as the most suitable tool for the analysis of **extremes**. Instinctively, someone would expect that years that are more active in terms of Number of Events *N* tend to exhibit extreme events also in terms of *Average Yi* magnitude.

Fig. 2 Spearman correlation coefficient between Average Yi and Number of over-threshold events N for all gauge locations, left: threshold 90%, right: threshold 99%.





Indeed, our study shows that **this assumption holds true** in most cases, yet there are **exceptions** shown in the maps in Figure 2 suggesting that it may not be **universally applicable**. The following Figure (2) present the Spearman correlation coefficient between *Average Yi* and Number of over-threshold events *N* for all the gauge locations for the selected thresholds.

Impact of clustering mechanisms on the correlation between the Average Loss and the number of over-threshold events

EGU General Assembly 2024

This depiction (Fig. 2) offers a **spatial categorization** of areas with **high** Spearman **correlation coefficient** between the *Average Yi* and the Number of over-threshold events *N*, in contrast with the ones where the correlation coefficient is noticeably **lower**. In other words, this spatial categorization indicates the regions that are **subjected to** numerous **claim** amounts in case of a year that an extreme number of over-threshold events occur. Moreover, it is shown that threshold selection influence slightly the Spearman, Pearson and Kendall correlation coefficient.

In addition, for each one of the **360 gauges** and for all selected thresholds, the Spearman correlation coefficient between the *Average Yi* and the Number of over-threshold events *N* was calculated for the observed as well as the shuffled (independent) time series in order to evaluate the **clustering mechanisms** on this correlation parameter. The following **box plots** show that clustering mechanisms that are prevailing over the **observed data** introduce significant correlation between the *N* and *Average Yi* in many gauge locations.

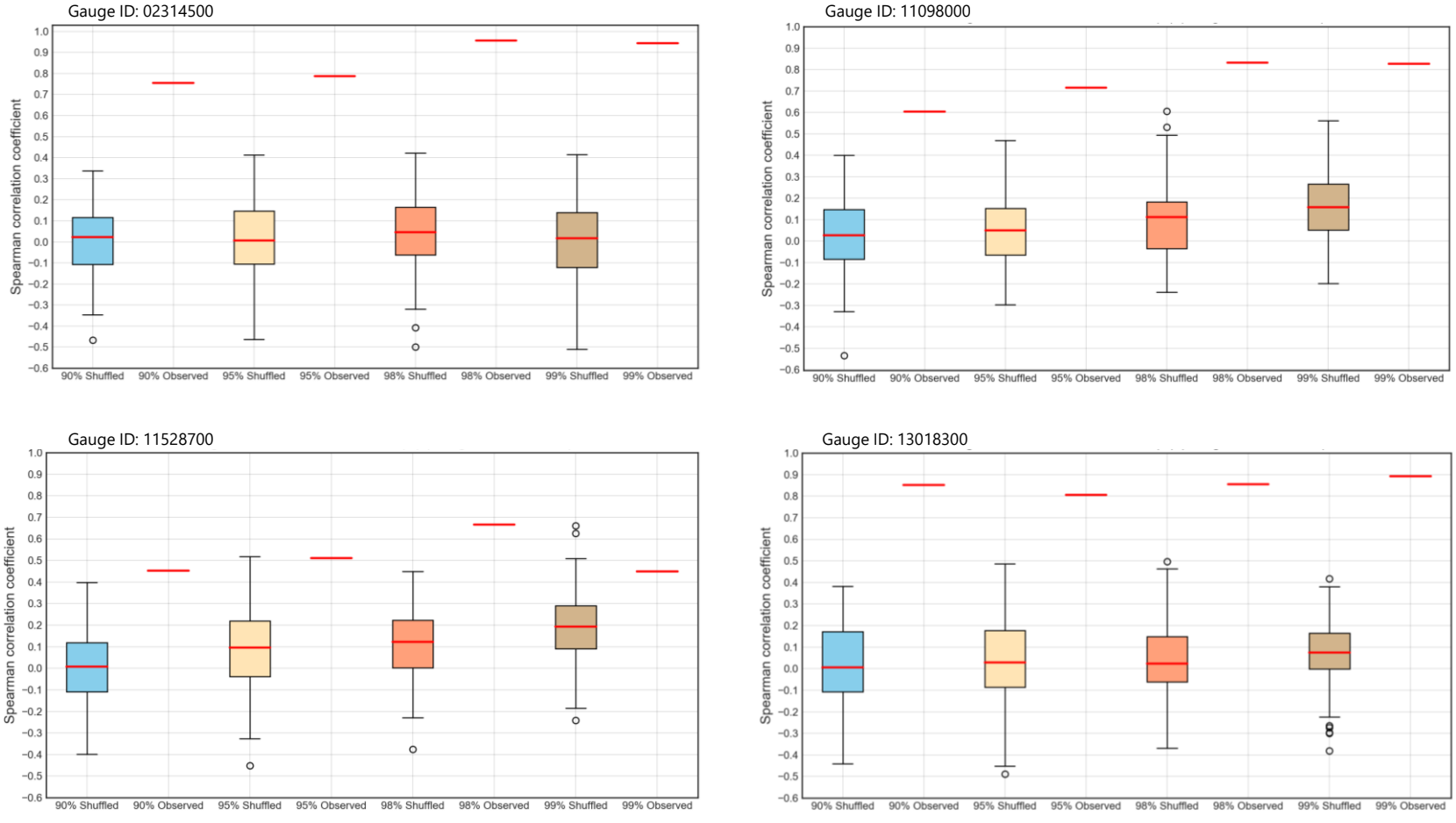
Impact of clustering mechanisms on the relation between Average Loss and Number of over-threshold events *N*

The conclusions of this investigation are quite impressive once again, as the **divergence** of the correlation coefficient between the **observed** and the **shuffled** ones in many gauge locations is profound. In more detail, the shuffled series tend to underestimate the correlation coefficient in comparison with the observed ones, which apparently introduce the **impacts** of **clustering mechanisms**. Ignorance of this behavior could lead insurance policy-makers on inaccurate conclusions which could potentially provoke **financial impacts**.

The **results** that are shown on the following Figure (3) present the above-mentioned **conclusions**. The gauge locations of these figures are:

- Suwannee River AT US 441 AT Fargo, GA (ID: 02314500)
- Arroyo Seco NR Pasadena, CA (ID: 11098000)
- SF Trinity R BL Hyampom, CA (ID: 11528700)
- Cache Creek Near Jackson, WY (ID: 13018300)

Fig. 3 Box plot of Spearman correlation coefficient between *Average Yi* and Number of over-threshold events *N* for the shuffled as well as the observed time series for all thresholds.



Reproducing observed clustering using HK dynamics and Monte Carlo Simulations

(Manolis et al.,2024)

Generalized-HK (GHK) model

EGU General Assembly 2024

Based on the **mean *Climacogram*** of the **GHK process** regarding the 360 empirical streamflow time series of the US-CAMELS dataset, a **persistent behavior** was indicated with parameters H = 0.63 and q = 6.94 days (Manolis et al., 2024).

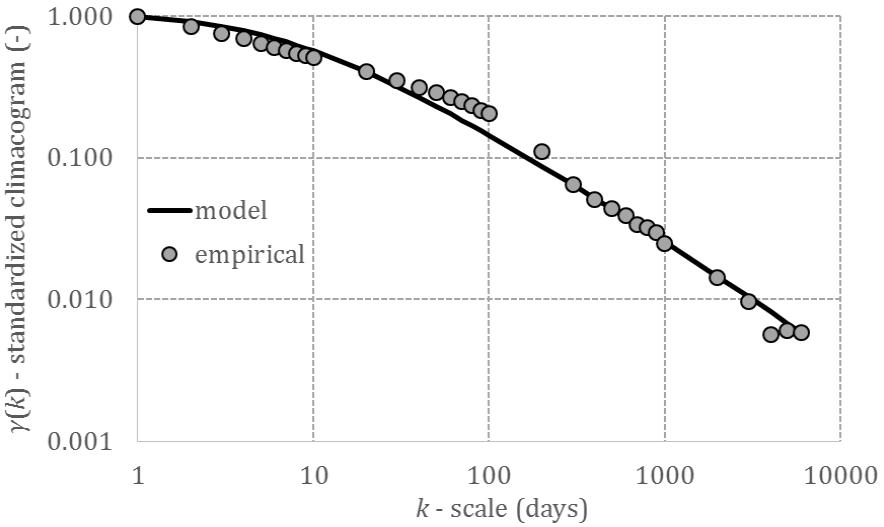


Fig. 4 The mean Climacogram of the 360 selected gauge locations of the US-CAMELS dataset.

Reproducing observed clustering using HK dynamics and Monte Carlo Simulations

(Manolis et al.,2024)

Generalized-HK (GHK) model

EGU General Assembly 2024

The effect of this **dependence structure** is tracked on the behaviors of POT flows at the annual scale and the estimation of the **Proxy Aggregated Losses**. The behavior of daily streamflow in our dataset is found to be **consistent with HK dynamics** (Dimitriadis et al., 2021) characterized by moderate *H* parameters (in the range 0.6-0.7), through Monte Carlo simulations.

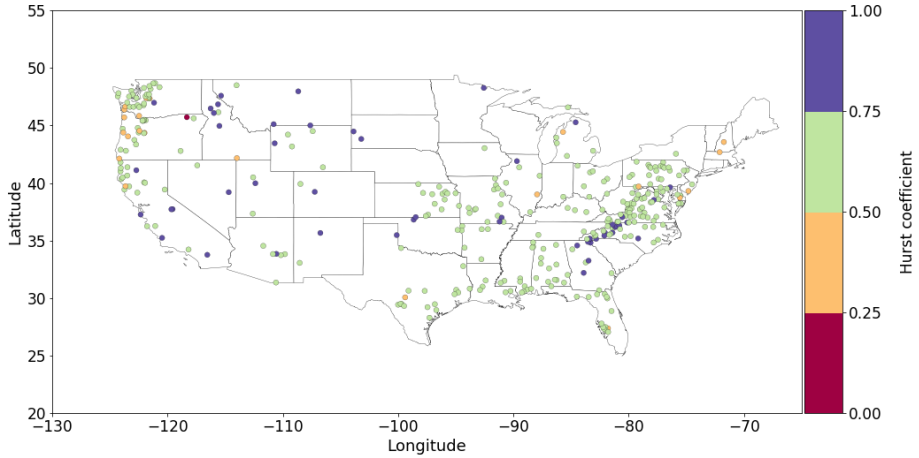


Fig. 5 Hurst coefficient *H* of each one of the 360 selected gauge locations of the US-CAMELS dataset.

EGU General Assembly 2024

A brief case study on spatial dependence mechanisms of US-CAMELS dataset

A brief case study on spatial dependence mechanisms of US-CAMELS dataset

EGU General Assembly 2024

Extreme river flows often appear to have **spatial patterns**, as a result, among others, of the complicated **hydrological processes**, weather systems' interplay and catchment structure.

On a global scale, firms that operate in the flood insurance and reinsurance sector utilize the theory and the applications of the **spatial dependence** in order to assess the probability distributions of **annual losses** to which their portfolio is exposed (i.e., the so-called loss-exceedance curve) and to reassure market regulators.

Moreover, governments, policy-makers and local authorities are interested in spatial dependence mechanisms, as they evaluate **economic scenarios** for investments and payments regarding **flood defenses** structures as well as compensations because of extreme **flood events**.

In all these cases, univariate extreme value analysis conducted at a series of sites **independently** cannot generate the year-to-year **variation** in flood losses that insurers experience (Pielke et al., 2008), and a **multivariate** analysis considering spatial dependence is required (Quinn et al., 2019).

A brief case study on spatial dependence mechanisms of US-CAMELS dataset

EGU General Assembly 2024

In the following figures (6-9), regarding **spatial correlation mechanisms** in **Hydrological Unit 3** of the US-CAMELS dataset, the **Spearman** correlation coefficients between annual Proxy Aggregated Losses *S* (which in our study is in fact a streamflow-based **proxy** for flood claims amounts) of the gauge locations that belong to this Hydrological Unit are presented for all the selected **thresholds**. In lower thresholds, the correlation coefficient is **higher** across the unit. In contrast, increasing the threshold has a strong impact on the results, as the range of the correlation coefficient across the unit seems to **vary**.

In more detail, when the correlation coefficient between two gauge locations ranges between:

- -0.25 and 0.25, it means that the correlation is practically zero and, as a result, flood events in these gauge locations can be considered as uncorrelated.
- 0.25 and 1, it means that when a flood event occurs in one of these gauge locations, it is highly probable that another event will occur on the other gauge locations, too.
- -0.25 and -1, it means that when a flood event occurs in one of these gauge locations, it is highly improbable that another potential event will occur on the other gauge locations.

A brief case study on spatial dependence mechanisms of US-CAMELS dataset

Insurance companies aim at having in their portfolios risks which have **negative** spatial or temporal correlation or, at least, zero correlation, in order to combine and aggregate risks which represent extreme flood events that are unlikely to happen **at the same space or time**.

Accordingly, the following figures are extremely significant from an **insurance** viewpoint, as they reveal which combination of insured properties on the mentioned gauge locations could compose a **profitable portfolio** in terms of **zero or negative dependence**.

Fig. 6 Spearman correlation between annual collective risk of gauge locations in Hydrological Unit 3; Threshold 90%.

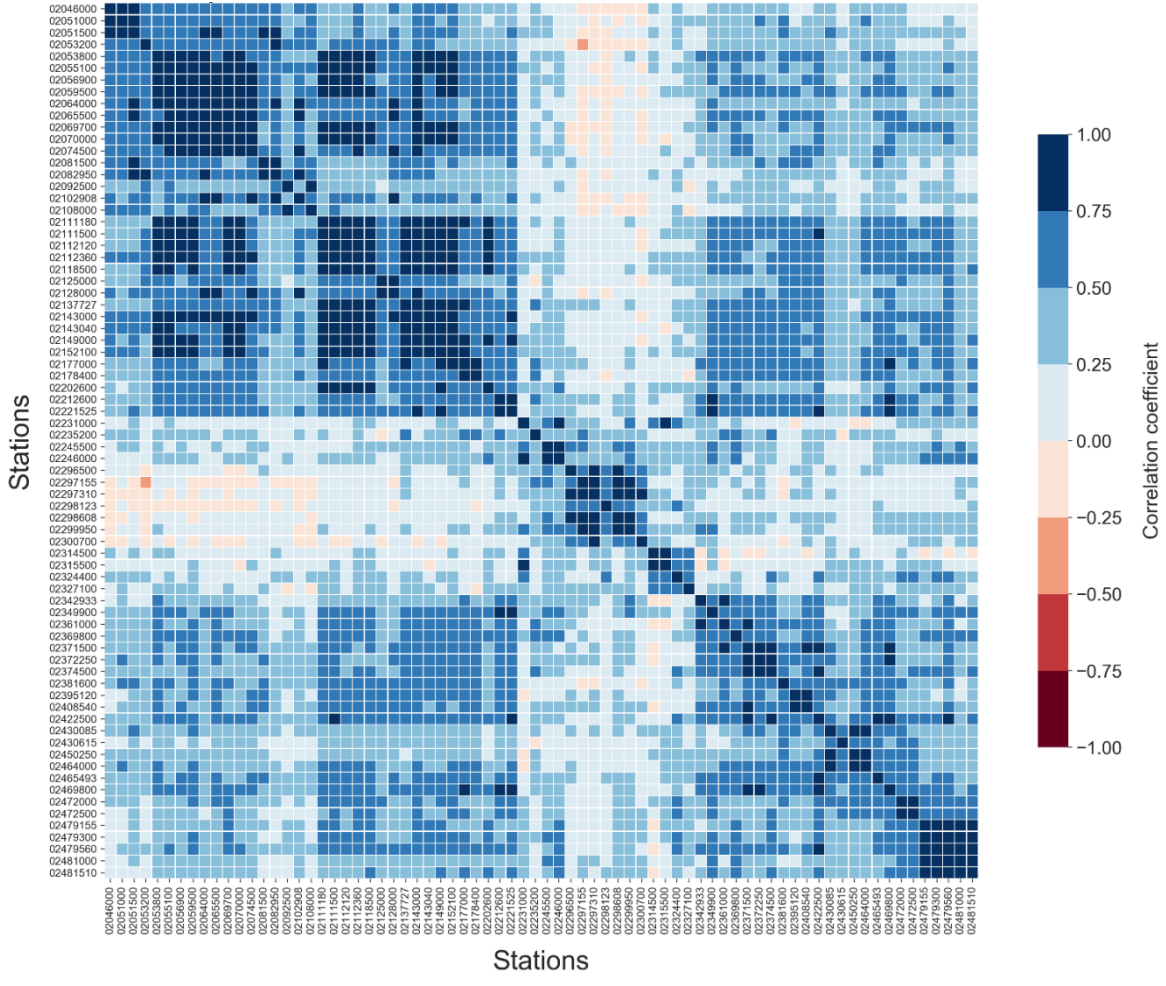


Fig. 7 Spearman correlation between annual collective risk of gauge locations in Hydrological Unit 3; Threshold 95%.

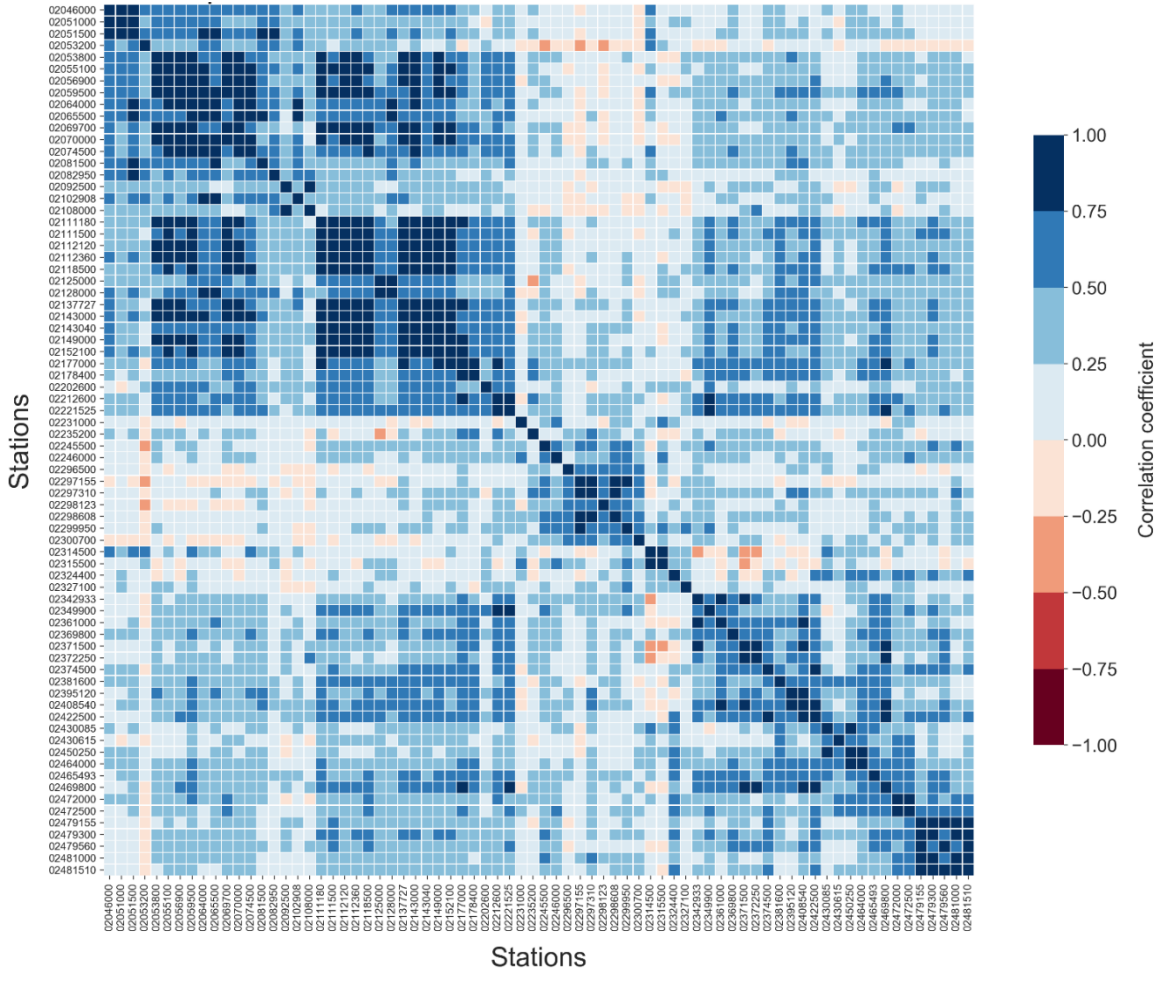


Fig. 8 Spearman correlation between annual collective risk of gauge locations in Hydrological Unit 3; Threshold 98%.

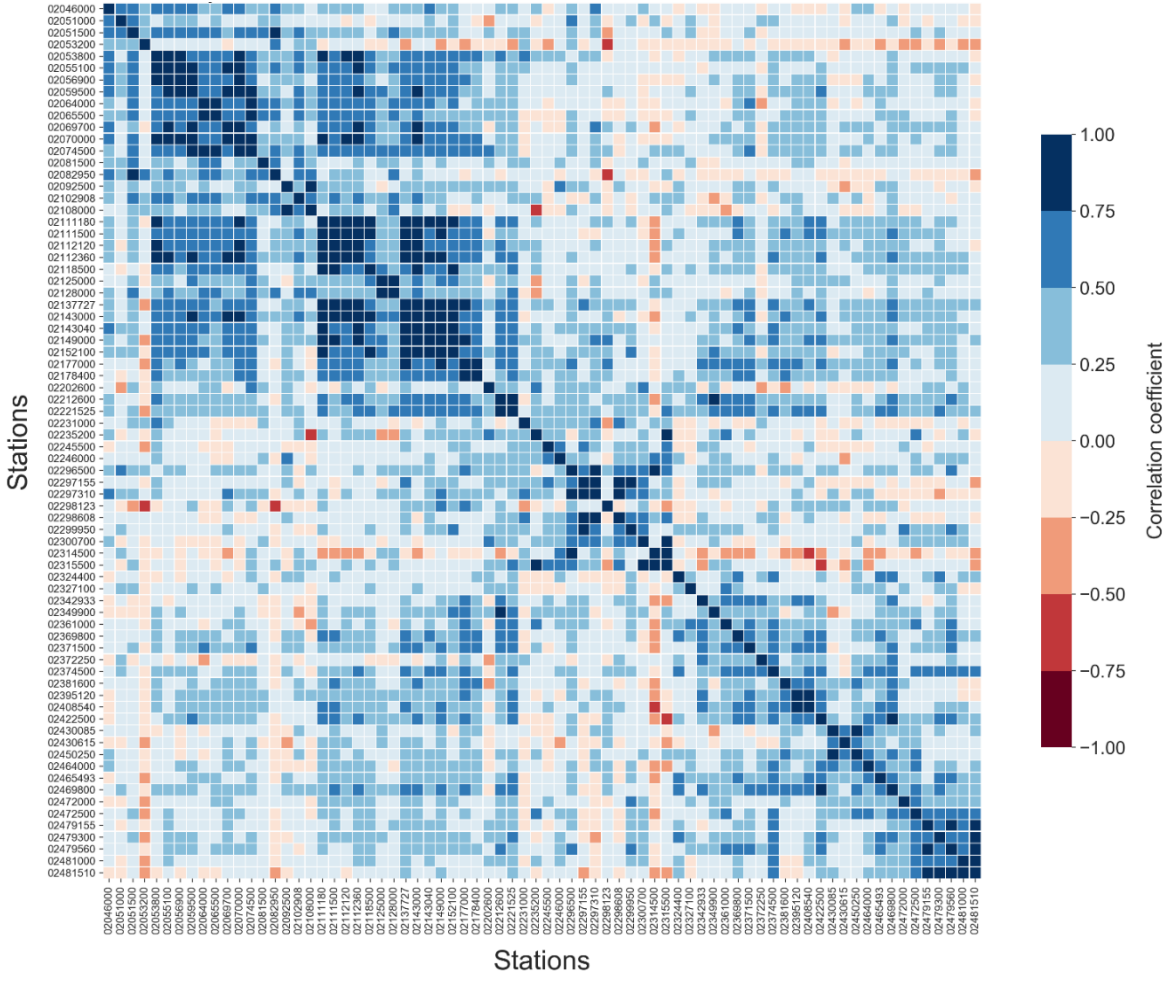
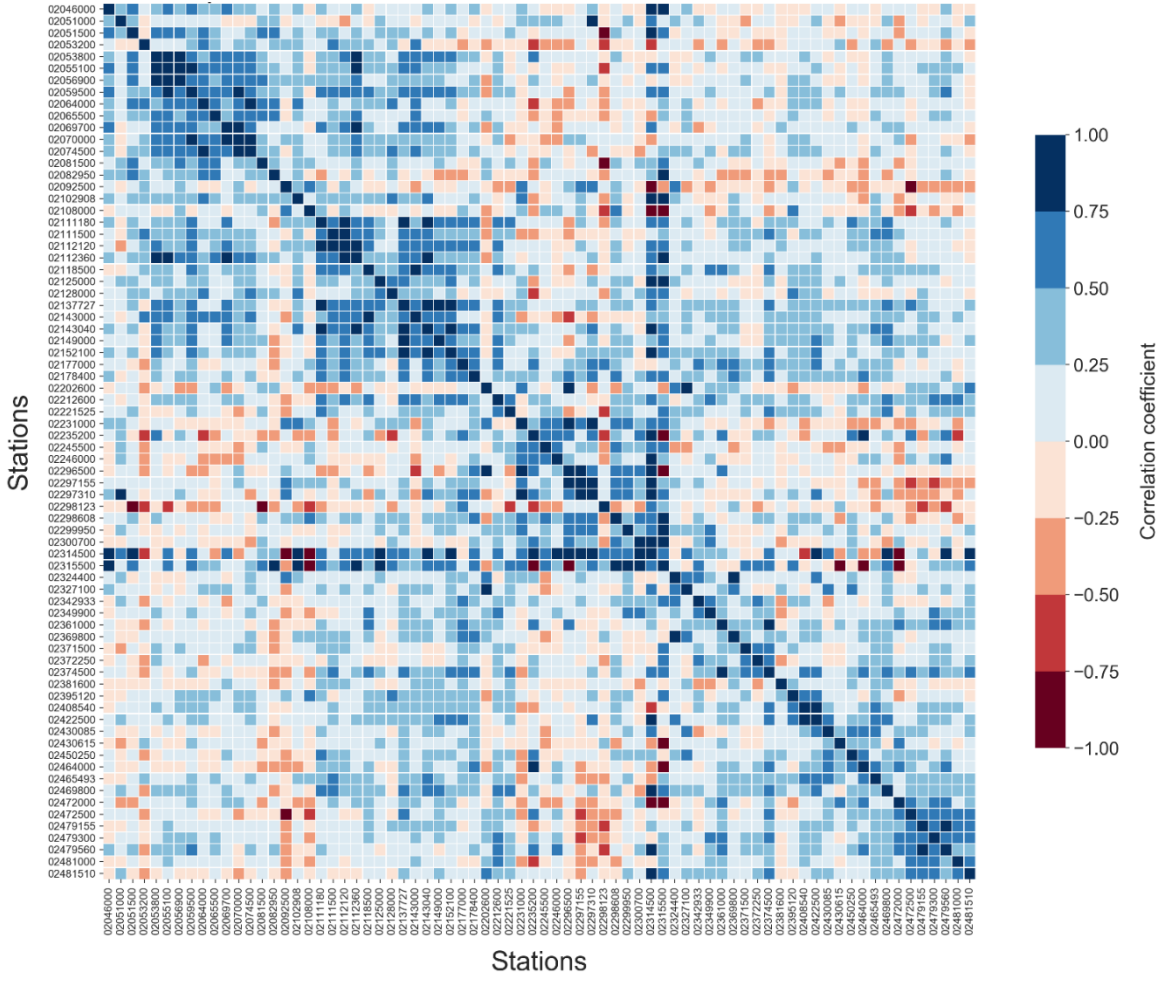


Fig. 9 Spearman correlation between annual collective risk of gauge locations in Hydrological Unit 3; Threshold 99%.



Conclusions

Conclusions on

the correlation between *Average Yi* and *N*

EGU General Assembly 2024

A common assumption in the computation of **Proxy Aggregated Losses *S*** *(*collective risk) is the **independence** between *Average Yi* and Number of over-threshold events *N*. Insurance companies' concern about this **correlation** factor is noteworthy, since they try to investigate the **dependence** between the annual number of **extreme** events and the provoked *Average Yi*, which is a proxy of the average claim amounts per over-threshold event on a **specific region**.

Initially, we showed that the **assumption** that years which are more active in terms of Number of events *N* tend to exhibit extreme events also in terms of *Average Yi* magnitude generally holds **true** yet it cannot be **universally applied**.

We categorized **spatially** the areas with **high** Spearman correlation coefficient between the *Average Yi* and the Number of over-threshold events *N*, and the others (areas) where the correlation coefficient is noticeably **lower**.

Conclusions on

the correlation between *Average Yi* and *N*

In other words, this **spatial** categorization indicates the regions that are subjected to numerous **claim amounts** in case of a year that an extreme number of over-threshold events occur. Subsequently, we highlighted through **a box plot analysis** the existence of **clustering mechanisms** that are prevailing over the **observed** data, as they introduce significant correlation between the *N* and *Average Yi* in many gauge locations, in contrast to the **shuffled** ones, in which the correlation is zero.

Conclusions on

the spatial dependence on US-CAMELS dataset

EGU General Assembly 2024

Regarding the **spatial** dependence mechanisms of US-CAMELS dataset, these mechanisms were investigated as part of a **case study**. The analysis showed that in lower thresholds, the correlation coefficient is higher across the unit. In contrast, increasing the threshold has a strong impact on the results, as in that case the range of the correlation coefficient across the unit seems to **vary greatly**, especially when the threshold is set to 99%.

Moreover, the extracted **heat maps** revealed which combination of insured properties on the mentioned gauge locations could compose a profitable portfolio in terms of zero or negative spatial dependence, as insurance companies' aim is to have in their **portfolio risks** which have **negative spatial** or **temporal correlation** or, at least, zero correlation, in order to combine and **aggregate risks** which represent potential extreme flood events that are **unlikely** to happen at the same space or time.

References

Dimitriadis, P., and Koutsoyiannis, D. (2015). Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes. Stochastic Environmental Research & Risk Assessment, 29 (6), 1649–1669.

Dimitriadis, P., and Koutsoyiannis, D. (2018). Stochastic synthesis approximating any process dependence and distribution. Stochastic Environmental Research & Risk Assessment, 32 (6), 1493-1515.

Dimitriadis, P., Koutsoyiannis, D., Iliopoulou, T., and Papanicolaou, P. (2021). A Global-Scale Investigation of Stochastic Similarities in Marginal Distribution and Dependence Structure of Key Hydrological-Cycle Processes. Hydrology, 8(2), 59.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, pp. 226–231, ISBN 1-57735-004-9.

FEMA (2019). FEMA publishes NFIP claims and policy data. [Online] Available at: https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v2. Last Data Refresh: 12-12-2023 [Accessed: December 2023]

Kaas, R., Goovaerts, M., Dhaene, J., and Denuit, M. (2008). Modern Actuarial Risk Theory Using R. Springer.

Koutsoyiannis, D. (2010). A random walk on water. Hydrology and Earth System Sciences, 14, 585-601.

Koutsoyiannis, D., and Dimitriadis, P. (2021). Towards generic simulation for demanding stochastic processes. Sci, 3, 34, doi:10.3390/sci3030034.

Manolis, G. T., Papoulakos, K., Iliopoulou, T., Dimitriadis, P., Tsaknias, D., and Koutsoyiannis, D. (2024). Exploring Hurst-Kolmogorov Dynamics: Unraveling the (temporal) link between Flood Insurance Claims and Streamflow Extremes in the contiguous USA. EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-12363, https://doi.org/10.5194/egusphere-egu24-12363.

Newman, A., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., and Blodgett, D. (2014). A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR.

Pielke, R. A., Gratz, J., Landsea, C. W., Collins, D., Saunders, M. A., and Musulin, R. (2008). Normalized hurricane damage in the United States: 1900-2005. Natural Hazards Review, 9 (1), 29- 42.

Quinn, N., Bates, P.D., Neal, J., Smith, A., Wing, O., Sampson, C., Smith, J., and Heffernan, J. (2019). The Spatial Dependence of Flood Hazard and Risk in the United States. Water Resources Research, 55 (3), 1890-1911.

Reiss, R.D., and Thomas, M. (2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser.

Robinson, P.J., and Botzen W.J.W. (2020). Flood insurance demand and probability weighting: The influences of regret, worry, locus of control and the threshold of concern heuristic. Water Resources and Economics, 30, 100144.

Serinaldi, F. and Kilsby, C.G. (2016). Understanding Persistence to Avoid Underestimation of Collective Flood Risk. Water, 8 (4), 152