Politecnico di Torino

Master’s Degree in Environmental and Land Engineering

Master’s Degree Thesis

Along-Stream Flood Estimation in Ungauged Locations in the Po River Basin

Supervisors:	**Candidate:**
Prof. Alberto Viglione	Saeed Shayanseresht
Prof. Daniele Ganora	

July 2025

Acknowledgment

I want to express sincere gratitude to Prof. Alberto Viglione and Prof. Daniele Ganora from the Polytechnic University of Turin for the great help and continuous interest and attention throughout the entire duration of the academic year, great support and guidance, and for insights that really helped with the successful conclusion of my master's thesis with their great knowledge and fantastic guidance that encouraged me a lot. I have been lucky to work with them.

Then I want to thank my family, who always stood by me during the many years at the university through the various ups and downs along the way. Thank you for being part of this journey; I will forever be grateful for your encouragement and support.

Cheers then go out to my classmates, who made the hours spent in lessons and exams more bearable throughout a period so peculiar and unprecedented for all of us in the past years.

Abstract

This thesis examines the performance of the Along-Stream approach in the prediction of flood statistics in ungauged basins and compares it with a regional approach officially in use in the Po river district in Italy. This method, introduced by Ganora (2013) and based on information transfer from close (short distance) donor sites, is here used for a larger dataset of 175 basins with three additional distance definitions. Distance definitions based on basin area, Euclidean distance between the basins, basin mean elevation and the combination of basin area and mean elevation are investigated. The study attempts to assess how different distance-threshold levels change the uncertainty of the estimates propagated and derived compared to the regional model. It is shown through the analysis that smaller distance limits, for example those based on the basin area ratio, result in much lower errors of propagated estimates when compared to the regional model, which, in turn, indicates that the predictions may be more accurate if confined to smaller distance limits. Increasing the distance threshold continues to increase the uncertainty of propagated estimates, thus indicating a trade-off between expanding the applicability domains while retaining prediction accuracy. The analysis locates D_{lim}= 1.6 (corresponding the basin area ratio of 4.95) for the area-based definition, D_{lim}= 1.3 (corresponding to the distance of 3.67 km) for Euclidean distance, and D_{lim}= 1.72 (corresponding to the basin area and elevation of 5.58) on the combined definition of the elevation-area ratio as the optimal thresholds. It is concluded from the analysis that the mean elevation may be unrepresentative of the basins and is not included in the optimum distance determination. At these limits, a wider percentage of basins (up to 30%, 25%, and 10% for area ratio, area-elevation multiplication and Euclidean distance, respectively) have lower RMSE for propagated estimates than the "regional" model. Thus, these thresholds provide a good balance between the applicability of the models and prediction uncertainty, hence making the Along-Stream approach a proper method for discharge prediction in ungauged basins along with regional models.

Contents

1- Introduction.. 1

2- Data and Methods 11

2-1- Study area 11

2-1-1- Po River Hydrological Behavior 13

2-1-2- Po River discharge variability 14

2-2- Methods and Hypotheses 16

2-3- Organization of Nested Basins and Definitions of Distances 20

2-4- Model Fitting and Reliability 23

2-4-1- Uncertainty of the propagated estimate 23

2-4-2- Parameter Estimation 24

2-4-3- Validity of the Approach 26

3- Results and discussions 28

3-1- Exploratory data analysis 28

3-2- Along-Stream approach 30

3-2-1- Distance based on Basin Area 30

3-2-2- Distance based on Euclidean distance 33

3-2-3- Distance based on Basin Mean Elevation 35

3-2-4- Distance based on Basin Area and Mean Elevation 37

3-3- Comparison of the Results 39

4- Conclusion 45

5- References 47

6- Appendices 51

6-1- Appendix I: List of the basins in the whole dataset 51

6-2- Appendix II: R Environment and Scripts for the index flood with D = log(A(max)/A(min))........ 57

List of Figures

Fig. 1. Po River basin map (from Wikipedia)

Fig. 2. Rainfall (Mean annual) in the Po River (from Po River Basin Authority., 2006).

Fig. 3. Hydrological fluxes (Mean annual) for the Po River basin (from Montanari., 2012)

Fig. 4. River discharges for the different locations (from Montanari., 2012)

Fig. 5. Annual maxima (left) and minima (right) of the Po River at Pontelagoscuro daily discharge series (1920–2009) and linear regression line (from Montanari., 2012)

Fig. 6. Sketch of the along-stream propagation of information (from Ganora et al., 2013)

Fig. 7. Basins in the whole dataset

Fig. 8. Schematic representation of the basins and the connections

Fig. 9. Process of iterative procedure

Fig. 10. Scatter plots of the local index flood versus the basin area and mean elevation

Fig. 11. (a) Box plots of the index flood for the local and regional estimate, (b) scatter plot of the regional model and local values

Fig. 12. Plot of the propagated estimate P_t and the local values S_t, (a) standard (b) log-log

Fig. 13. The maximum likelihood estimator of α on the loglikelihood plot for $D \leq D_{lim}$, $D = \log\,(A_{max}/A_{min})$

Fig. 14. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log\,(A_{max}/A_{min})$

Fig. 15. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log\,(Distance_{pair})$

Fig. 16. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log\,(H_{max}/H_{min})$

Fig. 17. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log\,(A_{max}/A_{min}\,.\,H_{max}/H_{min})$

Fig. 18. Graph of percentage of basins multiplied by the percentage of basins with propagation standard deviation lower than the regional model for different D_{lim} and different distance definitions

Fig. 19. Distribution of the distance definitions

Fig. 20. Root Mean Square Error of the propagated, operational and mean of regional for different distance thresholds

Fig. 21. Root Mean Square Error of the propagated, operational and mean of regional for different basin percentage

Fig. 22. Absolute Normalized Mean Error of the operational and mean of regional for optimum distance thresholds

Fig. 23. R-Studio visual interface

List of Tables

Table. 1. Summary of the literature reviewed in the text

Table. 2. Summary of the recent research and literature reviewed in the text

Table. 3. Observation period, mean value μ and standard deviation σ of the observed time series, along with the catchment area A at the considered location according to the Po River Basin Authority (2006)

Table. 4. Rules for the choice between regional model and operational approach

Table. 5. Results obtained from the first definition of distance, $D = \log\left(A_{max}/A_{min}\right)$

Table. 6. Results obtained from the first definition of distance, $D = \log\left(Distance_{pair}\right)$

Table. 7. Results obtained from the first definition of distance, $D = \log\left(H_{max}/H_{min}\right)$

Table. 8. Results obtained from the first definition of distance, $D = \log\left(A_{max}/A_{min}\,.\;H_{max}/H_{min}\right)$

1- Introduction

Hydrology is the science that studies the water cycle, the movement of water, in all its forms, through the atmosphere, land, and oceans. This dynamic cycle strongly governs all life on Earth. It defines the ecosystems and their biodiversity, and gives sustainability far beyond merely living, making the necessity of its total understanding of the hydrological mechanisms in academics as well as relevant to the global issues facing humankind. Surveillance of the charging phenomena of the water cycle would be necessary for ameliorating the climate change which has already disrupted the earlier pattern and has led to an increase in precipitation, evaporation, and runoff. An understanding of these mechanisms gives an idea to foresee extreme weather events like floods and droughts, which are becoming more and more common and severe. Hydrological studies also hold tremendous importance for the efficient management and sustainable use of water resources. With an increasing population and demand for clean drinking water, efficient distribution of this finite resource becomes paramount.

The water cycle is very difficult to understand indeed, and for many, the primary reason would be the vast number of physical phenomena that are involved in it, as well as the scales at which they happen. In fact, these scales are large both from a spatial and a temporal point of view. Knowledge and understanding about most of the elements of the water cycle are far from being complete or even well coupled. The majority of hydrological problems are referred to different areas of expertise. Uhlenbrook. (2006) comments on catchment hydrology as an all-comprehensive one, dealing with every terrestrial water cycle and its interactions over all the basins. In fact, basins can be termed as the elemental landscape units that incorporate hydrological cycles into geochemistry, ecology, morphology, and other processes (Sivapalan et al., 2003). All these processes, however, are strictly related to the fluxes through the boundaries of the basin, especially to and from the atmosphere and groundwater. However, one of the most significant elements, with respect to the water flux through the basin outlet is streamflow or runoff due to the importance that this variable has for many applications in practice. On common usage, streamflow, nearly always, is an easily interpretable index that captures in some way all the processes of the catchment.

One of the approaches to understanding the complex systems associated with a basin is through macro-characteristics, such as the magnitude, frequency, or duration of a certain kind of event. These macro-characteristics can be treated statistically, trying to describe the hydrological phenomena without the physical processes which is called statistical hydrology.

Quantitative and reliable characterization of surface water flows is becoming increasingly important both because they are being overexploited and because of the impacts of changing land use and urbanization, for proper management of this resource. Different users with conflicting requirements- the farmer, the industrialist, and the energy plant- make it less easy to find solutions when it comes to water exploitation. There is then the increasing concern for environmental issues brought about by water quantity and quality which demands practical management tools. Apart from the problem of conserving water, it is equally important to protect communities and properties against extreme events of water, especially floods and droughts. These are very pressing demands for quantitative, wide-ranging datasets over numerous variables in hydrology.

The straightforward way of analyzing the catchment behavior is through streamflow time series and, thereby, other variables: for example, precipitation, soil characteristics, vegetation, etc. For this, however, the discharge time series needs to be known at that site of interest. When it is not set up to directly monitor river flow or when collected data are not adequate to analyze, then the basin is called ungauged, and indirect approaches must be applied to study its hydrological characteristics. Indirect procedures are based on the concept of information transfer from gauged to ungauged basins. This refers to the hypothesis of compensating the deficiency of time series data from gauged sites through data of other sites. This is represented in the field of catchment hydrology, and the very clear proof of this is the Prediction in Ungauged Basins (PUB) initiative (Sivapalan et al., 2003).

In Regional approach the identification of the recurrence frequency of floods in ungauged catchments is done primarily using suitable statistical models that have been developed for flood statistics and different basin characteristics based on a given set of gauged stations. Such models transfer the information induced from the gauged sites and employ it to the target basin requiring only morphoclimatic catchment characteristics. Such procedure named a regional model as it identifies a subset, here called the region, of homogeneous basins to later use as a pooling set for the estimation at the ungauged site. Such basins in a region must therefore donate statistical proprieties to ungauged ones in the same region.

All flood frequency methods may have sampling variability when applied to a data collection of an isolated site for estimating return periods greater than the period of record at a site. (Hosking & Wallis., 1993, Cunnane., 1988). Hence, it is likely that regional flood frequency analysis (RFFA) would be more appropriate because an estimate at a single-gauged site can be improved by pooling data from other sites established as having similar frequency distributions. Some regional methods, furthermore, provide a measure for estimating flood frequencies even at ungauged sites within a region to which observations exist. However, the transfer of information from quantities at other sites can best be done within a "homogeneous" region, warranting the development of further techniques for the accurate identification of such regions (Saf, B., 2009).

Hydrological regionalization techniques have come up with different methods. Durrans & Tomic. (1996) considered these methods to be of two different kinds. The first method is the prediction in ungauged basins (PUB) (Sivapalan et al., 2003). In this method, certain flood peak discharge or low flow is revealed in comparison with the physiographic and climatic characteristics of the gauged basins. Further, it can be applied to ungauged basins to predict the hydrological characteristics by using the already measured physiographic and climatic characteristics. This multi-regression method toward this aim has been applied for many years (Mazvimavi et al., 2004). With the advancement of geo-information technologies such as geographic information systems (GIS) and remote sensing, an increasing amount of physiographic information is being made available (Lakshmi., 2004). The other form of regionalization is regional frequency analysis, where the assessment is at gauged sites, but related information from other gauged sites with records of longer duration within a homogenous region is introduced into the estimation (Chen, Y. D. et al., 2006).

The IFM called the index-flood method, evolved by the US Geological Survey (Dalrymple., 1960), is commonly employed for deriving regional flood frequency models at ungauged sites or gauged sites for which sufficient hydro-meteorological information is not available to reliably estimate extreme events. Examples of studies that document the application benefit of the index-flood method are those of Cunnane. (1988), Potter & Lettenmaier. (1990), and Pitlick. (1994). The country Canada has twelve studies, embracing various regions of the country, which have applied the index-flood method (Watt et al., 1989). The index-flood method has been applied to Portugal using the observatories of annual maximum flood series in 120 stream gauging stations (Portela & Dias., 2005). Six homogeneous regions were delimited and models were created to each region in order to have the flood quantiles estimation (Saf, B., 2009).

The advances in regional frequency analysis are about the incorporation of L-moments in index-flooding, according to Hosking & Wallis. (1997). This methodology is applied successfully in modeling floods in many case studies in the US (Vogel et al., 1993, Vogel & Wilson., 1996), in New Zealand (Pearson., 1991, 1995, Madsen et al., 1997), in southern Africa (Mkhandi & Kachroo 1997, Mkhandi., 1995, Kjeldsen et al., 2000, 2001), in India (Parida et al., 1998, Kumar et al., 2003), in Australia (Pearson et al., 1991), Malaysia (Lim & Lye., 2003), and Turkey (Saf et al., 2007, Saf, B., 2009). L-moments offer considerably more advantages over ordinary product moment methods in the treatment of environmental data sets, for these reasons:

- L-moment ratio estimation of location, scale, and shape are nearly unbiased, irrespective of the probability distribution from which observations are derived (Hosking., 1990).
- Estimators of L-moment ratios, for example, L-coefficient of variation, L-skewness, and L-kurtosis, can be less biased than product moment ratio estimators in a highly skewed scenario.
- L-moment ratio estimators of L-coefficient of variation and L-skewness have no dependency on sample-size-bound constraints, unlike the ordinary product moment ratio estimators of coefficient of variation and skewness.
- Being linear combinations of the observations, L-moments show less sensitivity towards extreme observations in a sample than product moments, as the latter involve squaring or cubing the observations.
- In comparison with ordinary product moment diagrams, which are nearly useless for the task, L-moment ratio diagrams are very efficient in identifying distributional properties of highly skewed data.

In the regional data, the at-site data do not enter into the estimation of local parameters of a statistical distribution model; instead, the information in the sample record is summarized by a set of robust sample statistics (the L-moments), which are then regionalized. This is a kind of generalization of the widely accepted index-flooding approach. To pass information from gauged locations to ungauged ones, one must summarize the time series data. The L-moments and dimensionless ratios can be used as regional variables; specifically, the first L-moment (mean) and coefficient of L-variation (LCV) and L-skewness (L- coefficient of asymmetry, LCA) of the record are picked. After L-moments have been regionalized, one can reconstruct the entire flood frequency curve. Index-flood framework can be interpreted as the choice among mean, LCV, and LCA as hydrological signatures in a regional framework (Dalrymple., 1960) in which the mean would serve as a scale factor while the L-moments ratios would be descriptors of the dimensionless growth curve. The proposed method also makes it possible to eliminate the uncertainty about the choice of the distribution function especially, where short samples are involved making it possible for short samples, which would otherwise be thrown away, to contribute toward better consistency of the database. Separate regional models were adopted for each of the L-moments in the transfer of information to ungauged basins, being based on a very rigorously structured multiple regression approach, by selecting from numerous geomorphological and climatic descriptors of the catchment. Every regression model is calibrated by non-standard least-squares techniques on the whole dataset without any grouping procedure creating sub-regions.

In Along stream approach the main goal of regional models is to transfer information from gauged sites toward the specific ungauged basin. A variety of models, theoretical and experimental, have emerged for this purpose in literature; however, they share a common philosophy of adopting a descriptor space approach to address the lack of hydrologic information. This descriptor space is a set of catchment characteristics usually comprising topographic, morphological, or climatic indices computed without applying any hydrologic data for every basin. Then, one constructs the appropriate relationships to relate those catchment characteristics with the expected hydrological variable.

The majority of standard statistical methods used for the estimation of flood frequency curves in ungauged basins are limited. Such limitations are mainly in two ways: first, by subdividing the area of interest into homogeneous regions, and second, by predefining an a priori probability distribution for the sample data. This goes beyond the peculiarity of the estimations that would accompany the region itself with abrupt changes because distributions prove not to maintain their properties within and between the regions. This is a limitation for making these estimations (Laio. Et al., 2011).

Regional models do not retain information regarding the natural hierarchy between the gauged stations that derive from where they are located along the river network. This information is particularly important as runoff is to be estimated at a site situated immediately upstream or downstream from a gauged station. An alternative to this method might be estimating this variable directly, against the corresponding statistics measured at the gauged station. The closest to the gauged station is the estimate point; thus, this method should give a greater expected quality.

The founding principle of the model developed in this thesis is that of transferring hydrological information to an ungauged site located upstream or downstream of the gauging station. The information that we are interested in, i.e., that which we transfer along the stream network, is the one used to reconstruct the flood frequency curve, such as the L-moments. This transfer strategy integrates hydrologic data, and at the same time, defines the structure of the drainage network such that points are directly connected to one another. In other words, it might as well be said that the two basins are nested. Along-Stream (AS) approach involves at least one variable calculated in a gauged (or donor) basin and is propagated, bringing the information toward the ungauged (target) site where the variable of interest.

In conclusion, this method, defined earlier as the Along-Stream estimation method, to signify that it is applied to points along a stream network. It requires a formula to compute the variable at the ungauged site. This formula could rely on a series of basin characterizations or, alternatively, a regional estimate (local estimation combined with a regional model). Then there is defined a criterion for assessing the reliability of the stream model and its domain of application, and, finally, the accuracy of the approach is assessed through the evaluation of the standard deviation of the estimates. In this way, it is possible to compare the variance of the stream estimates against the variance of other models, if such are available, and thus choose the most accurate method (or to combine different estimates).

Although there are some notable examples, the problem of hydrological variable prediction or interpolation over a river network is usually not discussed in the literature. From Gottschalk (1993a and b), the problem represents the correlation and covariance of runoff and its interpolation along the river, using the theory of stochastic processes with the structuring hierarchy of nested catchments. It has been extended by Gottschalk et al. (2006), and the same concepts have been used by Skoien et al. (2006) in developing a kriging procedure that takes care of river structure, termed topological kriging or top-kriging. While the final aim is the same, the process developed here is built following a completely different angle.

The study carried out by Kjeldsen and Jones. (2007) in interpolation of runoff statistics. Here, the local correction of regional estimation is taken into consideration. The approach is very similar to the one developed in this thesis because of the information transfer scheme; however, a different implementation procedure is shown. The summary of the references reviewed are listed in table 1.

Table. 1. Summary of the literature reviewed in the text

Author(s)	Year	Subject	Findings
Dalrymple	1960	Index-flood method	Enabled derivation of regional flood frequency models at ungauged sites; served as a foundational method for regional flood analysis globally.
Hosking & Wallis	1993	Regional flood frequency analysis (RFFA) using L-moments	Emphasized pooling data from similar regions to improve flood estimates at ungauged sites.
Gottschalk	1993a, b	Topological kriging (top-kriging) for river networks	Developed kriging methods for predicting runoff and addressing river structure correlations.
Durrans & Tomic	1996	Regionalization methods including multi-regression and GIS-based approaches	Explored physiographic and climatic characteristics for predicting hydrological characteristics; emphasized advancements in geo-information technologies for data analysis.
Hosking & Wallis	1997	Incorporation of L-moments in regional frequency analysis	Successfully applied in multiple regions, including the US, Africa, India, and Turkey.
Madsen et al.	1997	L-moments applied in various countries including New Zealand and the US	Demonstrated global applicability of L-moments for improving the accuracy of flood frequency models.
Sivapalan et al.	2003	Prediction in Ungauged Basins (PUB) initiative; regional models	Highlighted the importance of transferring information from gauged to ungauged basins using morphoclimatic characteristics; enabled quantitative hydrological predictions.
Mazvimavi et al.	2004	Multi-regression model	Applied physiographic and climatic characteristics to predict hydrological features in ungauged basins effectively.
Portela & Dias	2005	Application of index-flood method in Portugal	Created six homogeneous regions for flood quantiles estimation, demonstrating the method's applicability in diverse geographical settings.
Uhlenbrook	2006	Catchment hydrology as an all-comprehensive approach	Highlighted the significance of basins as elemental landscape units that integrate multiple processes.
Skoien et al.	2006	Topological kriging for river structures	Developed a kriging procedure to incorporate river structure into hydrological estimations; addressed the spatial hierarchy of gauged stations.
Saf et al.	2007	L-moments and regional analysis	Successfully applied L-moments in modeling floods in Turkey; emphasized robust data summarization techniques for ungauged sites.
Kjeldsen & Jones	2007	Interpolation of runoff statistics	Local correction of regional estimates, emphasizing the transfer of information along river networks.

Some most recent research done in this topic are also summarized in table 2.

Table. 2. Summary of the recent research and literature reviewed in the text

Author(s)	Year	Subject	Findings
Ali Ahmed et al.	2023	Identification of homogeneous regions, AI-based models, climate change impacts	Various statistical tests proposed, including L-moments, Nonstationary RFFA methods needed to account for changing flood patterns
Hassan Esmaeili-Gisavandani et al.	2023	Random Forest, ANFIS, M5 decision tree, and multivariate regression.	RF performed best ($R^2 = 0.96$, NRMSE = 0.223); all models outperformed regression in RFFA.

Sabrina Ali and Ataur Rahman	2022	Regionalized Flood Frequency Analysis (RFFA) method utilizes Regional Kriging for ungauged Catchments.	(a) the developed kriging-based RFFA model is a viable alternative for flood quantile estimation in ungauged catchments, (b) the 10-year ARI model Q_{10} performs best among the six quantiles, which is followed by the models Q_5 and Q_{20}, and (c) the kriging-based RFFA model is found to outperform the 'RFFE model 2016'.

In this study the Along-stream approach introduced in 2013 by Ganora et al. will be used on the river networks and basins of the river Po for flood estimation in order to have an additional model next to the regional one to choose between the two models. This process will be explained in the methodology part later. The objective is to estimate or interpolate the hydrological variable along the river network and correcting the estimate of the regional model locally, on the river structure calculated on the gauged site (donor) towards the ungauged basin (target) which are directly connected like the basins are nested. Different definitions of distance will be used to choose the most optimum definition and evaluate the corresponding errors. All the possible pairs of connected basins will be considered for information transfer.

2- Data and Methods

2-1- Study area

The river as the longest in Italy is Po, with its main course about 652 km long (Figure 1). Moreover, it includes not only the largest watershed in Italy, covering about 71,000 km^2 at the delta, but also the time series discharge at the closure river cross-section, which has been traditionally placed at Pontelagoscuro (44°53´19.34´´N and 11°36´29.60´´E). Included in this is the minimum, mean, and maximum daily river flow recorded in Italy: 275 m^3/s, 1,470 m^3/s, and 10,300 m^3/s, respectively. The Po is comprised of main tributaries, with an average of 141 such tributaries in this area. The network of the main tributaries has an estimated length of around 6750 km, while that of artificial and natural channels accounts for 31,000 km. An annual average of 78 km^3 constitutes precipitation, from which 60% is intercepted and converted to outflow volume at the closure section. There are about 450 lakes in the Po basin. The water level of the bigger south-alpine, glacial-origin lakes is regulated through specific management schemes. But it is worth mentioning the establishment of 9 hydro-ecoregions in the Po River (Po River Basin Authority., 2006) defined as geographic areas in which freshwater ecosystems show small ranges in variation in terms of chemical, physical, and biological parameters. The spatially distributed rainfall over the catchment is illustrated in Figure 2 (Montanari., 2012).

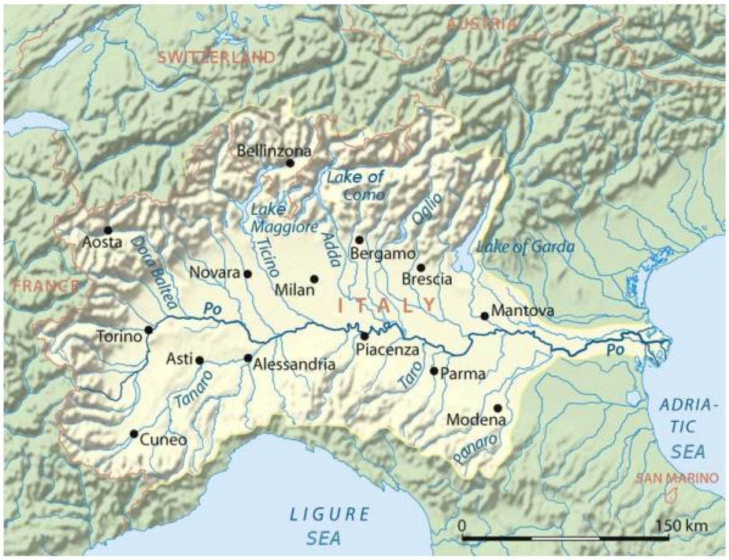


Fig. 1. Po River basin map (from Wikipedia)

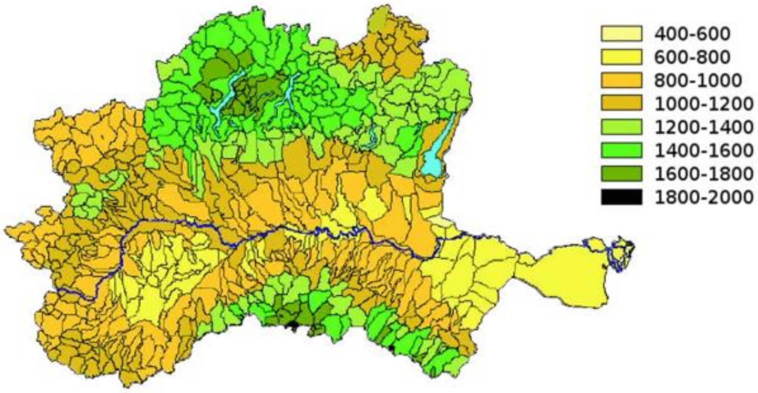


Fig. 2. Rainfall (Mean annual) in the Po River (from Montanari., 2012).

This synthetic description paints the complete picture of the Po River basin's complexity. Different hydrological behaviors and ecosystems coexist and coevolve within the basin. Interestingly enough, the Po River Basin Authority. (2006) has identified 12 different fluvial regimes in the Po catchment.

2-1-1- Po River Hydrological Behavior

The hydrological behaviors of the Po River have been explored more rigorously, particularly with regard to the flood regime (Piccoli., 1976, Marchi., 1994, Zanchettini et al., 2008, Visentini., 1953). However, many pertinent questions remain open in the context of Po hydrology, particularly in relation to the huge impact that intense human activity has had on its catchment over the course of the 20th century and climate change itself. This is because long periods of such abundance or scarcity of river flows give rise to scientific questions that remain largely undiscovered.

Hydrological fluxes for the annual average of the Po River basin are indicated in Figure 3 (Po River Basin Authority., 2006). To be specific, the much-discussed average volume of annual precipitation, as per the above, complements outlet river discharge, anguished by the annual infiltration into the underground aquifer (~9 km^3) and evapotranspiration from vegetation (~20-25 km^3). This is the same as saying that the withdrawal from the aquifer is about 6.5 km^3, that is, groundwater resources are nearing their critical point of overexploitation (deep percolation is nearly about 1 km^3, and there is some groundwater flow to the sea). Annual water withdrawals for irrigation that are contributing to evapotranspiration add up to approximately 17 km^3, with respect to industrial and domestic water withdrawal.

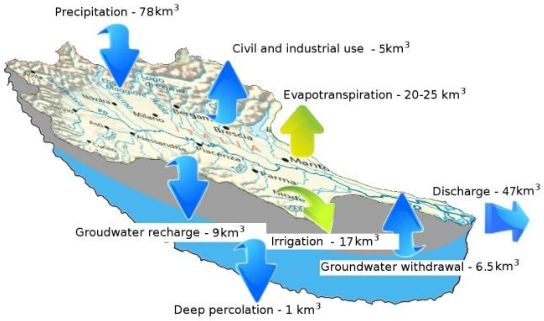

Precipitation - 78km^3
Civil and industrial use - 5km^3
Evapotranspiration - 20-25 km^3
Discharge - 47km^3
Groundwater recharge - 9km^3
Irrigation - 17km^3
Groundwater withdrawal - 6.5km^3
Deep percolation - 1 km^3

Fig. 3. Hydrological fluxes (Mean annual) for the Po River basin (from Montanari., 2012)

2-1-2- Po River discharge variability

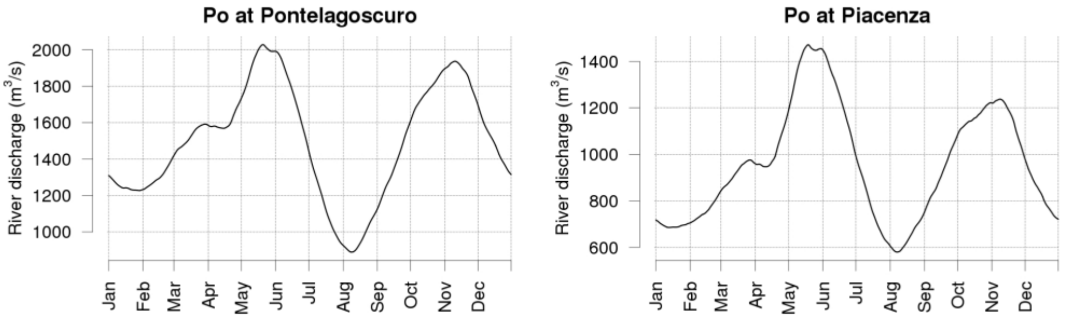
Intra-annual variations

Hydrologists commonly study the variability of river discharge with the consideration of the intra-annual period. It is also named the seasonal regime (the progress of annual average river flow). Daily river flow time series measured at Piacenza, Pontelagoscuro, and Moncalieri along the Po River, the Dora Baltea River at Tavagnasco, the Tanaro River at Farigliano, and the Stura di Lanzo River at Lanzo. The observation period of the series, their mean and standard deviation values together with catchment area and synthetic information on the dominant fluvial regime are displayed in Table 3.

Table. 3. Observation period, mean value μ and standard deviation σ of the observed time series, along with the catchment area A at the considered location according to the Po River Basin Authority. (2006)

Location	Period	μ (m^3/s)	σ (m^3/s)	A (km^2)
Po at Pontelagoscuro	1920–2009	1470	1007	71 000
Po at Piacenza	1924–2009	959	773	42 030
Po at Moncalieri	1942–1984	80	89	4885
Tanaro at Farigliano	1944–1973	39	49	1522
Stura di Lanzo at Lanzo	1946–1981	19	27	582
Dora Baltea at Tavagnasco	1951–1989	91	78	3314

Figure 4 shows the discharges for the different locations.



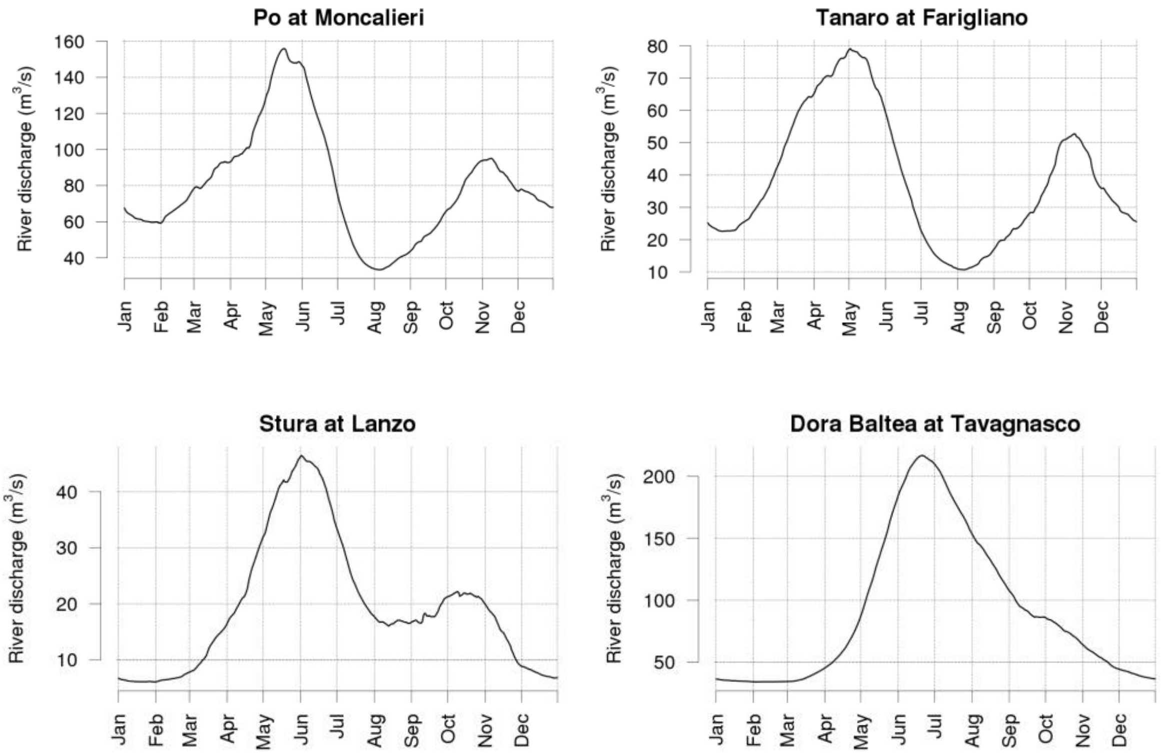


Fig. 4. River discharges for the different locations (from Montanari., 2012)

Inter-annual variations

Annual maximum and minimum value progress are shown in Figure 5, respectively, for the daily river flows at Pontelagoscuro of the Po River and the linear regression line relating that maximum or minimum value throughout the entire time period.

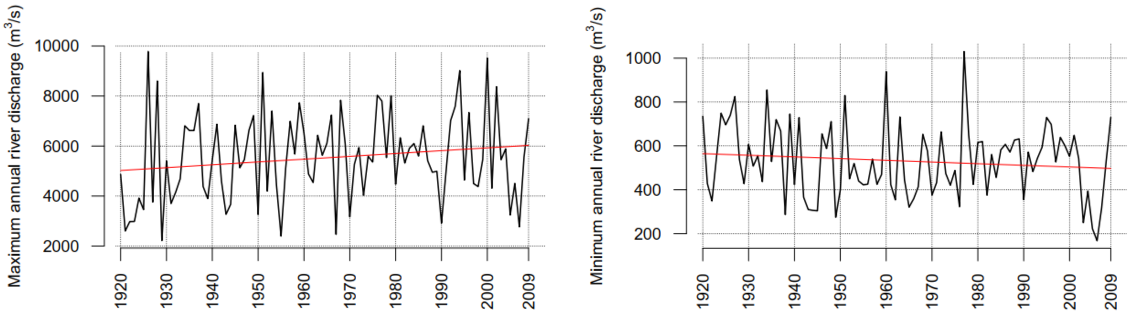


Fig. 5. Annual maxima (left) and minima (right) of the Po River at Pontelagoscuro daily discharge series (1920–2009) and linear regression line (from Montanari., 2012)

It must be noted, however, that the above trends are scarcely relevant from the statistical standpoint. In fact, assuming that the data are independent and that the null hypothesis of no-trend is true, one finds the p-values of 11% and 26% for the slope of the linear regressions applied to annual maxima and minima, respectively.

2-2- Methods and Hypotheses

Given below is the Along-Stream (AS) procedure, which is devised and presented for the statistics in the form of the index flood, which summarize the vital statistics as far as the estimation of flood quantiles is concerned. The AS model will then become additional to the regional procedure to predict the same variables. Then it would also include results from two different approaches that could be combined in order to provide more reliable final estimates in ungauged sites. In general, when two or more models are available for the same goal, following scenarios can be considered:

- Different models (AS and regional prediction for this work) could be evaluated to see which one would give a better modeling of the variable of interest in this case study. AS and regional prediction are supposed to have different reliability related to the target site location and specifically due to its distance from the donor site. Accordingly, AS is viewed as another method that may be more relevant for ungauged basins.

- Another model is defined as the use of output from a model to initialize the other model. The regional estimate could be, in this work, further considered as an additional parameter for the along-stream estimation function hence contributing to the final AS prediction. Hence the same could also be interpreted as follows: AS may help globally, but in localized areas, it may be better than the regional model estimate corrected based on the specific information available in a close donor site.

- Various estimates can be combined using appropriate relations with an aim to further minimize variance in the resulting estimator.

Along-stream approach will be used to estimate some hydrological variable *P* by propagating the information from a donor site *d* to that of the destination site *t*. This approach has a few assumptions. In particular:

Proximity: the target site is always located on the same stream path of the donor station, upstream or downstream, i,e. the two basins *d* and *t* are nested;

Transferability: the variable S_d, computed at the donor site, must be used in the information transfer, i.e.

$$P_t = f\,(S_d;\,\boldsymbol{\theta})$$

where θ is an additional (optional) set of parameters and *f* is a function to be defined;

Congruence: when the distance between the donor and the ungauged catchment becomes zero, AS estimate (variance) at the ungauged site must coincide with at-site estimate (variance) at the gauged basin, i.e.

$$P_t \rightarrow S_d \text{ for } t \rightarrow d$$

Schematic representation of the proximity and transferability hypotheses is shown in the sketch in Figure 6, a. where the arrows show possible directions for the information transfer. In the approximate sense, the function that transfers the information is not known but could be approximated by any function that satisfies to the hypotheses raised. This function must be a good approximation to the real unknown transfer function at least within a validity domain that includes a set of points close to the donor station; then different functions have, in general, different validity domains (see Figure 6, b). The validity domain hypothesis is very important in assessing the reliability of the AS method and will deal with it quite intuitively. In particular, a threshold on the distance between donor and target basins will be defined to separate the domain of validity of the selected transfer function from the remaining part of the drainage network.

The distance is intended with a general meaning, and it does not necessarily mean a geographic distance or the length of the drainage path. Moreover, given a specific information transfer function, and its corresponding validity domain, the variance of the AS prediction is expected to increase moving away from the donor site, but still within the validity domain. Beyond this, there are no reliable AS predictions, and there is no need to compute their variance. A sketch representing this aspect is shown in Figure 6, c.

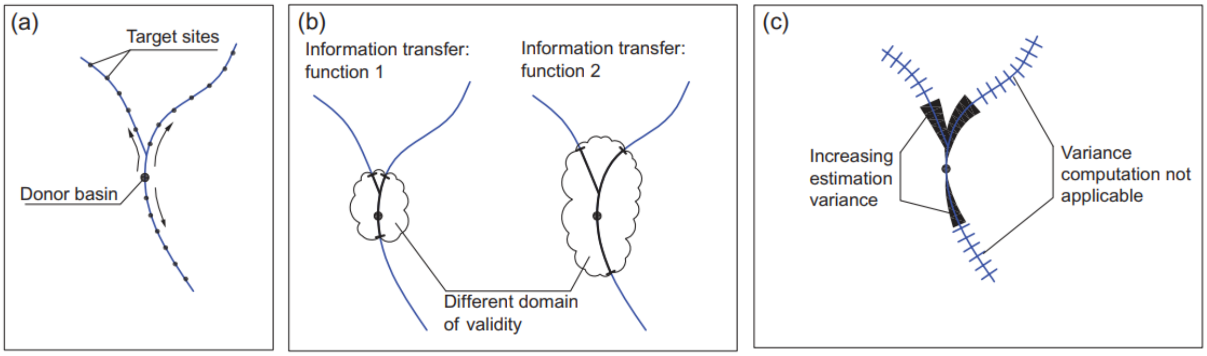

(a)
Target sites
Donor basin
(b)
Information transfer: function 1
Information transfer: function 2
Different domain of validity
(c)
Increasing estimation variance
Variance computation not applicable

Fig. 6. Sketch of the along-stream propagation of information (from Ganora et al., 2013)

The application of an along-stream modeling approach in this work also involves a regional model; to be specific, the regional model is meant to capture the 'global' variability of hydrological variables and does not include the 'local' structure of the river. The AS estimates are then calculated based on the regional ones. However, the reliability of the AS predictions diminishes with increasing distances between the donor and the target basins. Thus, the procedure remains to be defined, which allows for deciding if the AS estimates can be considered reliable or whether to prefer the regional one.

The first step in carrying out the estimation procedure of the along-stream is to come up with a function that would give the variable *P* at a target site *t* according to all the assumptions initialization made.

Let *f* be the function used for the along-stream information transfer, that reads

$$f_{t,d} = \frac{R_t}{R_d} . S_d \tag{1}$$

where the symbol *R* refers to the regional estimates and *S* is the at-site variable. The equation was suggested in the Flood Estimation Handbook (Institute of Hydrology 1999) and re-analyzed by Kjeldsen and Jones (2007) is used (Ganora, 2013).

The propagated estimate can be simply written as:

$$P_t = [f_{t,d}]\ for\ D \leq D_{lim} \tag{2}$$

Where *D* is the defined generalized distance relating *t* with *d*, and D_{lim} is the defined threshold distance beyond which the function becomes ineffective. The symbol $D \leq D_{lim}$ emphasizes that over the boundary of its validity, the transfer formula can be applicable. It should be borne in mind that all *P*, *R*, and *S* symbolically represent a generic hydrological variable-indices flood, LCV, and LCA in the particular context. This can be interpreted in a simple way with the help of the equation: the correction factor will be just the relative error that the regional model produces in *d* that is $(\frac{S_d}{R_d})$. In practice, it assumes that the regional model has the same error magnitude when assessing these two close locations. For *D* → 0 it is straightforward to verify that $P_t \rightarrow S_d$.

For example. if there exist two different functions available for the transfer of information along the stream network (similar to the representation in Figure 6, b). The first one is defined as:

$$P_t^{(1)} = S_d \; for \; D \leq \mathrm{D}_{lim}^{(1)} \tag{3}$$

Where D is the generalized distance between t and d and $\mathrm{D}_{lim}^{(1)}$ is the threshold distance beyond which function 1 is no longer effective. The second function is:

$$P_t^{(2)} = f_{t,d} \; for \; D \leq \mathrm{D}_{lim}^{(2)} \tag{4}$$

This first function simply indicates that the propagated estimate $P_t^{(1)}$ is equal to the at-site variable calculated in *d*. Obviously, equation (3) can be considered valid only in a very limited neighborhood of *d*, i.e. the threshold $\mathrm{D}_{lim}^{(1)}$ is supposed to be very low, and thus $\mathrm{D}_{lim}^{(1)} \leq \mathrm{D}_{lim}^{(2)}$.

Depending on the distance *D* there are three different possibilities:

- $D \leq \mathrm{D}_{lim}^{(1)} \leq \mathrm{D}_{lim}^{(2)}$: both the AS models are valid, the most appropriate can be selected on the basis of the prediction variance;
- $\mathrm{D}_{lim}^{(1)} \leq \mathrm{D} \leq \mathrm{D}_{lim}^{(2)}$: only model 2 can be used to propagate the information along the stream network;
- $\mathrm{D} > \mathrm{D}_{lim}^{(2)}$: neither model can be used.

2-3- Organization of Nested Basins and Definitions of Distances

The complete dataset was comprised of 227 basins with some of them were without a connection to the other basins and some of them were not included in the network data to determine the connection. Overall, this method is used for a case study on a set of 175 basins in northwestern Italy, shown in Figure 7. This dataset constituted by the catchments already used for the regional analysis. Here, it is more appropriate to work in terms of pairs of basins, {*t, d*}, rather than single catchments, at least in this context. Figure 8 shows a schematic representation of the hierarchical dependence of nested catchments, representing the connection with a line. Note that there are also multi-connected basins and basins with no connections. All the connected (nested) catchments have been considered as possible pairs of donor-target sites, characterized by a generalized distance among them.

Considering all possible connections of two stations on the same path of drainage (within nested basins), there are 270 connections (e.g.: from Figure 8, basins A021 is nested to basin A018 even if there are the intermediate basins A199 or A020). Although all the basins involved are gauged basins, all the connections are considered "in both directions"; for instance, if basin A041 is "upstream" basin A051, conditions are first drawn up regarding basin A041 being a donor site and basin A051 a target (ungauged) site; the same is then repeated concerning basin A051 as donor station and basin A041 being the target (ungauged) site. This way, 540 becomes the overall number of available connections $\{t, d\}$. Complete list of the catchments and their characteristics are accessible in appendix I.

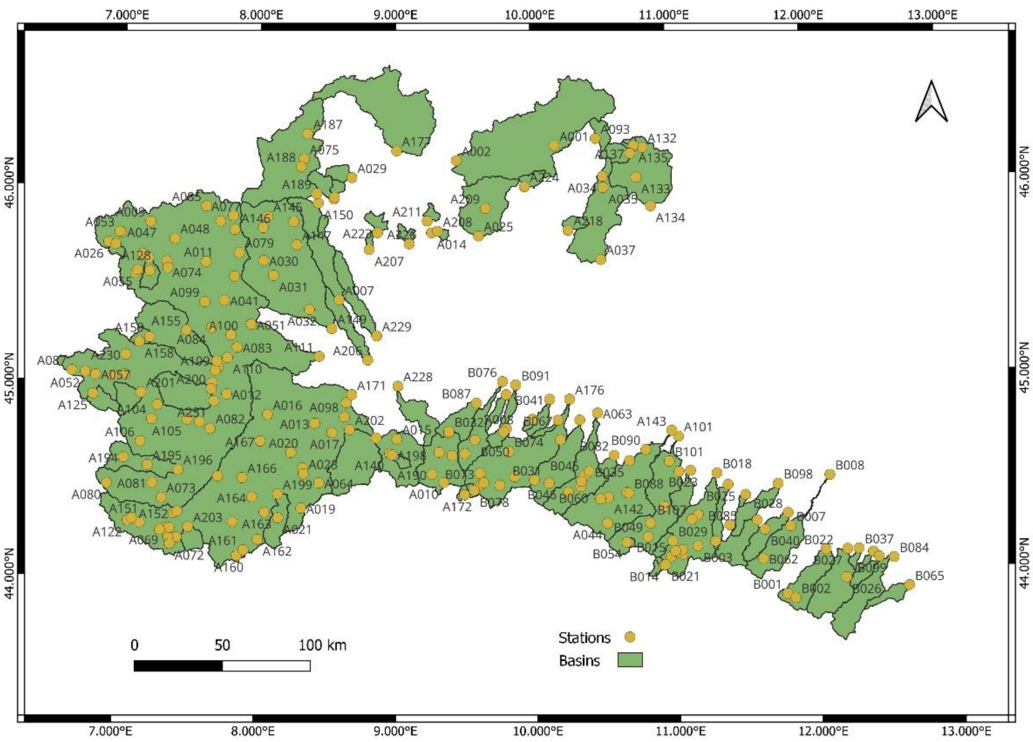


Fig. 7. Basins in the whole dataset

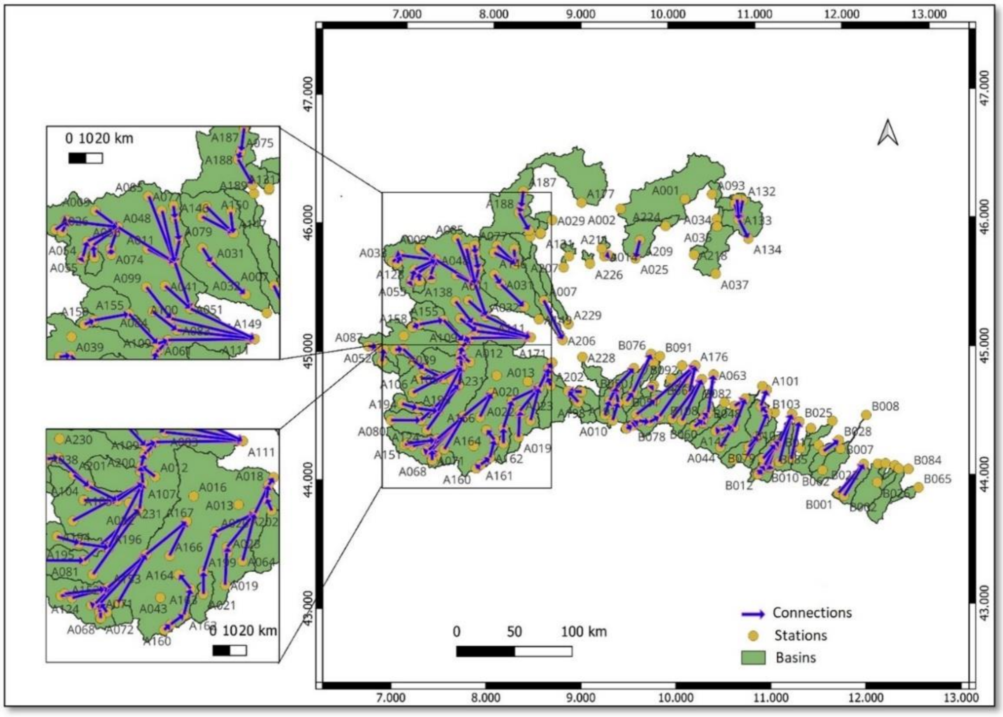


Fig. 8. Schematic representation of the basins and the connections

Among the many definitions that can be used to characterize the difference between two basins, a definition of distance based on the basin area A, Euclidean distance between two basins, basin mean elevation H and a combined formula of the two A and H is proposed. In this case,

$$D = \log\,(A_{max}/A_{min}) \tag{5}$$

With $A_{max} = \max\,[A_t, A_d\,]$ and $A_{min} = \min\,[A_t, A_d\,]$ is the first type of definition. According to the proximity assumption (but not generally), two basins with the same area have a null distance (they are the same basin), so their estimates must show coincident results (congruence hypothesis).

The second method is based on the Euclidean distance:

$$D = \log\,(Distance_{pair}) \tag{6}$$

Other simple definitions of distance, which employ the basin mean elevation H, and A are as follows:

$$D = \log\,(\,H_{max}/H_{min}) \tag{7}$$

$$D = \log\,(A_{max}/A_{min}\,.\,\,H_{max}/H_{min}) \tag{8}$$

The latest definition is expected to perform satisfactorily in cases where the mean basin elevation and area are independent, such as when the data consist of basins from both mountainous and flat states.

2-4- Model Fitting and Reliability

2-4-1- Uncertainty of the propagated estimate

The foundation of the AS approach could be summarized in two phases, (i) choice of an appropriate equation for transferring information and (ii) establishment of the threshold distance D_{lim} which is directly related to the equation adopted above. This particular case study relies on a practical equation (equation (2)), for the adoption and non-quantitative assessment of D_{lim}. This section, therefore, considers the applicability of the simplified approach for quantifying D_{lim} and the overall performance of the AS approach. The AS procedure as applied to index-flood through Equation (2) based on the 540 pairs of catchments.

The formula for calculating the uncertainty of P_t is now explained. In simplified approach, a model of P_t uncertainty is:

$$CV_{P_t} = (1 + \alpha \cdot D) \cdot CV_{S_d} \tag{9}$$

where *CV* stands for coefficient of variation, that is, the ratio of standard deviation to mean of the variable. Considering the definition of P_t provided in equation (2), and the definition of *CV* as the ratio between the standard deviation and mean, we obtain

$$\sigma_{P_t} \cdot \frac{R_d}{R_t \cdot S_d} = (1 + \alpha \cdot D) \cdot \frac{\sigma_{S_d}}{S_d} \tag{10}$$

And thus:

$$\sigma_{P_t} = (1 + \alpha \cdot D) \cdot \sigma_{S_d} \cdot \frac{R_t}{R_d} \tag{11}$$

This model could be interpreted as the estimate of σ_{P_t} in that the standard deviation of P_t is simply the standard deviation of the at-site estimate in the gauged site, increased proportionally by a factor f' accounting for the not perfectness of the AS transfer function and for the uncertainty of all the variables involved in equation (2). Moreover, for *D* → 0, it is straightforward to verify that $\sigma_{P_t}\rightarrow \sigma_{S_d}$, thus confirming the congruence hypothesis.

2-4-2- Parameter Estimation

To obtain the uncertainty of the AS estimate using eq. (11), it would require, first of all, an estimate of the parameter α already calibrated according to the available dataset rearranged for donor-target correspondence. For pairwise basins, the residual of P_t with respect to its corresponding at-site value S_t is

$$\delta_t = P_t - S_t \tag{12}$$

and, since both P_t and S_t are independent random variables, the supposed distribution of these residuals is

$$\delta_t \sim \mathcal{N}\left(0, \sigma_{P_t}^2 + \sigma_{S_t}^2\right) \tag{13}$$

Substituting eq. (11) into eq. (13), we find the final expression for the residual variance parametric in α as

$$\sigma_\delta^2 = (1 + \alpha \cdot d)^2 \cdot \sigma_{S_d}^2 \cdot \left(\frac{R_t}{R_d}\right)^2 + \sigma_{S_t}^2 \tag{14}$$

The coefficient α can be estimated using the maximum likelihood approach. If a set of *n* independent observations δ_1, δ_2, …, δ_n is taken into consideration, each follows a normal distribution:

$$\delta_i \sim \mathcal{N}\left(\mu_\delta, \sigma_\delta^2\ \right) \tag{15}$$

The likelihood function $\mathcal{L}$ of the residuals is the joint probability of observing the data given the parameters μ_δ and σ_δ and is supposed to follow a normal distribution of eq. (13) is as in

$$\mathcal{L}(\delta) = \prod \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp\left[-\frac{1}{2}\left(\frac{\delta - \mu_\delta}{\sigma_\delta}\right)^2\right] \tag{16}$$

that can be handled more easily after a logarithmic transformation:

$$\log\,\mathcal{L}(\delta) = -\frac{1}{2}\sum\left[2\pi\sigma_\delta^2 + \frac{\delta^2}{2\sigma_\delta^2}\right] \tag{17}$$

The maximum likelihood estimator can also be computed numerically through maximization of equation (17) or setting to zero its first derivative.

This approach was adopted with the appropriate functions in R language.

2-4-3- Validity of the Approach

The main objective of the procedure described above was to employ either regional or AS approaches in building ungauged site projects to result in final prediction. Thus, in practical terms, it calls for defining the operational (O) prediction as the estimate from an AS or a regional procedure, depending on which is appropriate under these rules shown in table 4.

Table. 4. Rules for the choice between regional model and operational approach

	$\sigma_{P_t} \leq \sigma_{R_t}$	$\sigma_{P_t} > \sigma_{R_t}$
$D \leq D_{lim}$	ASE	Regional
$D > D_{lim}$	Regional	Regional

The correct value of D_{lim} cannot be known a priori, but can be evaluated with an iterative process:

- A tentative value of D_{lim} is empirically defined.
- The AS estimate P_t and the regional one, R_t, would then be evaluated.
- Residuals of the AS estimates and parameter α are computed under max-likelihood applicable only to the pair basins within D_{lim}.
- Based on α, AS prediction variance would be determined using equation (11) and compared with variance for regional prediction at that location.
- The operational estimate is constructed by choosing the model with lower uncertainty.
- The operational estimate errors would be compared to those of the regional model considered the reference model.
- The procedure is repeated by changing the tentative D_{lim} value.

The mean error, named ME, is computed averaging the errors

$$E_{\{t,d\}} = \frac{(\text{ prediction })_d - S_t}{\sigma_{S_t}} \tag{19}$$

obtained for each pair $\{t,\ d\}$, where "prediction" indicates one of the three possible models. A representation of the iterative procedure is illustrated in Figure 9.

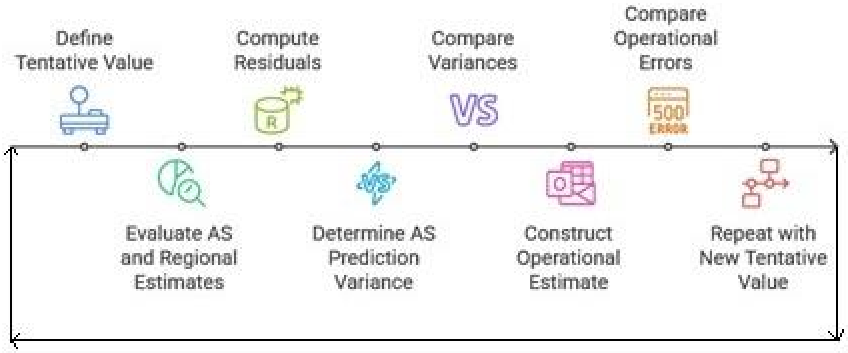

Define Tentative Value
Evaluate AS and Regional Estimates
Compute Residuals
Determine AS Prediction Variance
Compare Variances
Construct Operational Estimate
Compare Operational Errors
Repeat with New Tentative Value

Fig. 9. Process of iterative procedure

Mean error will be indicated for every pair uncertainty analysis and the Root Mean Square Error (RMSE) will be used for the performance of the methods.

3- Results and discussions

3-1- Exploratory data analysis

An early phase in data analysis, which involves searching the dataset structure to identify patterns and possible anomalies before initiating complex modeling or statistical analysis, is done for the main variables needed for this study. In Figure 10. a, the trend in data points between the basin area and index flood in local values show a positive correlation, with an increase in one resulting in an increase in the other. Furthermore, values are apparently scattered across more than one order of magnitude. For example, there is a difference between 5 and 5000 on the x-axis and 5 and more than 2000 on the y-axis. This indicates a power-law-type relationship in which the local index flood rises at a decreasing rate in proportion to area increase. For the Figure 10. b, however, between the mean elevation against local index flood, there is no evident trend for upward or downward between the variables. At almost all elevations, the local index flood values are diverged over a fairly wide range, indicating that elevation in itself does not have a major influence on flood magnitude. There exists a small possibility of a trend whereby higher elevations (elevations above ~1500m) appear to associate with lower floods, although not strong. Perhaps there are other geographical or climatic factors interfering.

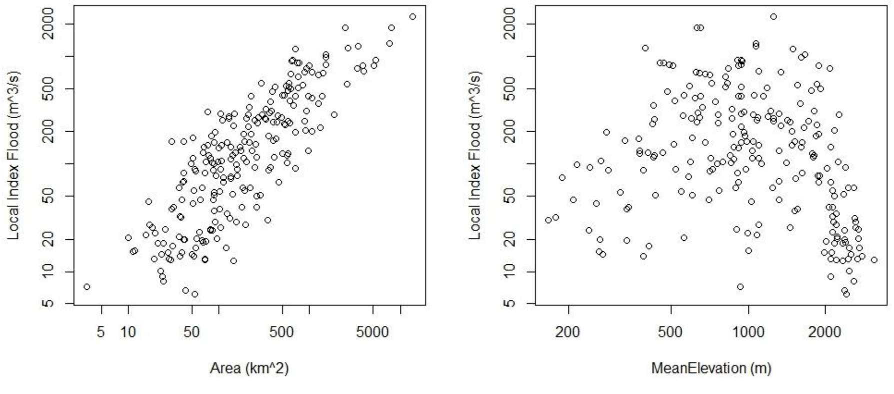


Fig. 10. Scatter plots of the local index flood versus the basin area and mean elevation

A box plot for the regional and local index floods is represented in Figure 11. a. Local and Regional display very similar distribution characteristics regarding spread and central tendency. The median values (bold horizontal lines in boxes) seem to have very similar measures. The inter-quartile ranges (IQRs) (shaded boxes) are also comparable. Both distributions maintain a remarkably broad range of about 5 m³/s to over 2000 m³/s. Some data points have been found to be outliers (represented with open circles) above the upper end of the whiskers, which indicates a positively skewed distribution. The whiskers extend to relatively similar ranges in both groups. But there are some outliers which are more distributed in the regional model compared to the local. Figure 11. b. shows the scatter plot of the regional model estimate and the local values. Local index flood has a clear positive linear correlation with the regional index flood. Therefore, it tends to rise proportionally along with an increase in the local index flood. The data points are aligned closely around the fitted line indicating that a linear equation is a good model for the relationship of the two variables. Most points tend to follow the fitted line, but some are scattered. This means regional floods are very good estimators of local flooding.

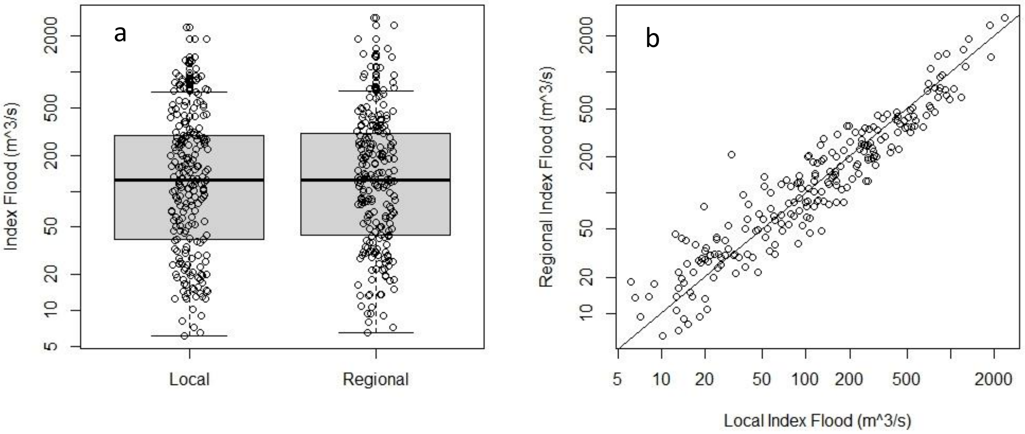


Fig. 11. (a) Box plots of the index flood for the local and regional estimate, (b) scatter plot of the regional model and local values

3-2- Along-Stream approach

At first, eq. 1 is applied to the data in order to calculate P_t. In Figure 12 the standard and log-log plot of the propagated estimate in the target sites and corresponding at-site estimates are illustrated, respectively. This is for the whole pairs of the catchments without applying any distance limit. It can be identified multiple P_t values relative to the same S_t which represents multiple sources for local estimation because there is multiple P_t estimates corresponding to the same S_t value. By increasing the distance between donor and target, higher estimation uncertainty is easily recognizable. So, defining a distance threshold has to be defined for the estimation based on the uncertainty between the regional estimate and the along-stream approach.

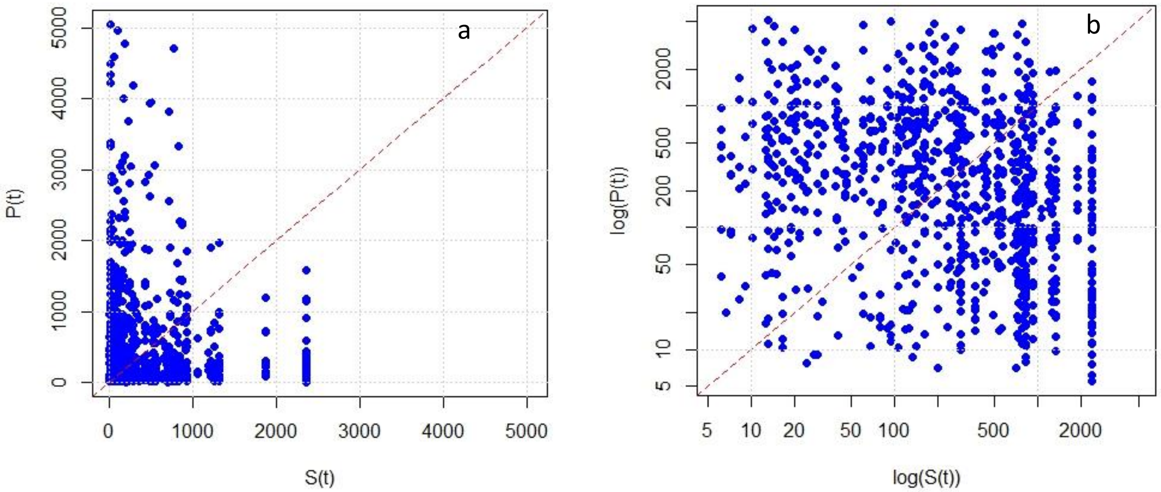


Fig. 12. Plot of the propagated estimate P_t and the local values S_t, (a) standard (b) log-log

The iterative trial-error approach mentioned earlier was applied to the index flood statistic with the different equations for distance measurements (eq. 5 to 8).

3-2-1- Distance based on Basin Area

In Figure 13 the fitting results, shown only for this distance definition, as an example of the log-likelihood and the corresponding alpha values and its maximum is reported for the first four distance limit in the case of the distance D defined as $D = \log(A_{max}/A_{min})$. The main results are reported in Table 5.

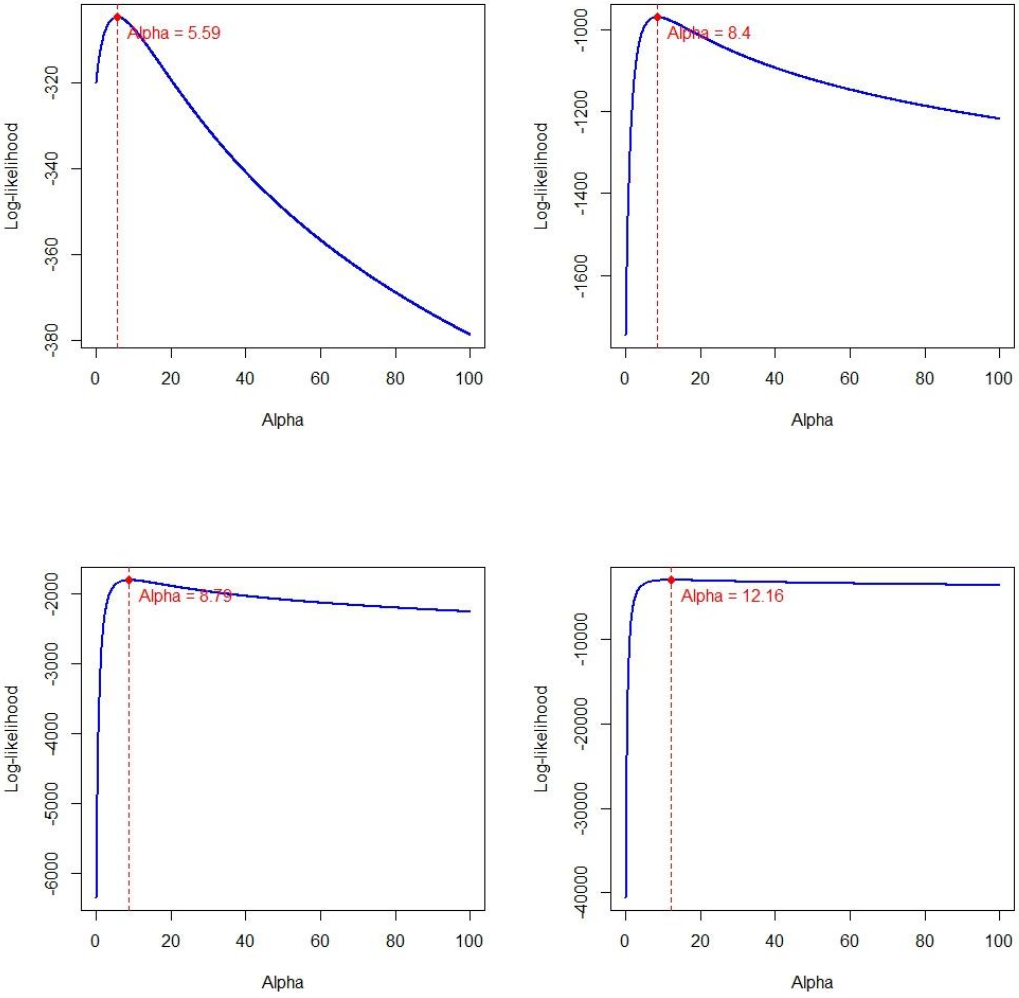


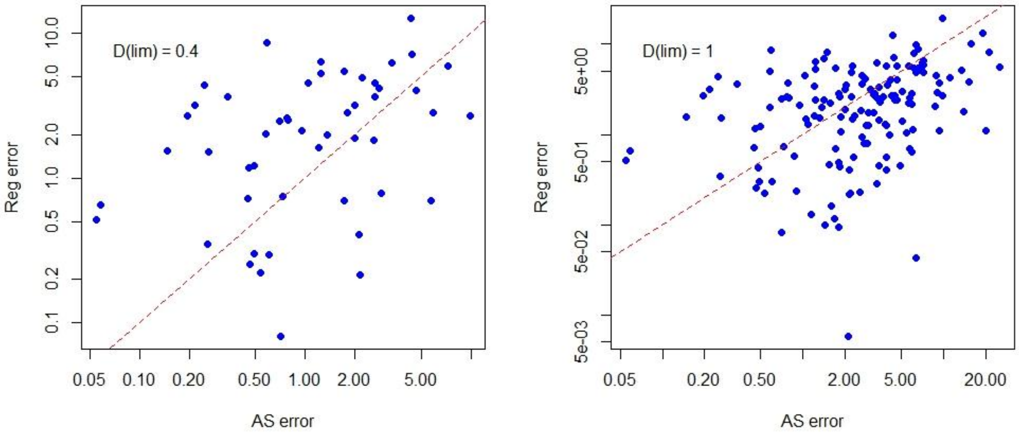
Fig. 13. The maximum likelihood estimator of α on the loglikelihood plot for $D \leq D_{lim}$, $D = \log(A_{max}/A_{min})$

Table. 5. Results obtained from the first definition of distance, $D = \log(A_{max}/A_{min})$

	Results										
$\boldsymbol{D_{lim}}$	0.4	1	1.6	2.2	2.8	3.4	4	4.6	5.2	5.8	6.4
$\boldsymbol{\alpha}$	5.58	8.4	8.78	12.16	15.24	19.29	26.08	36.91	48.92	61.71	69.53
$\%\boldsymbol{Basins}$	5.58	16.74	29.24	44.20	59.60	71.43	81.25	89.95	96.43	98.88	100
$\%\boldsymbol{Basins}\ (\boldsymbol{\sigma_{P_t} \leq \sigma_{R_t}})$	82	36.67	20.99	9.84	4.50	3.43	1.92	1.24	0.70	0.45	0.44

In this table percentage of the basins that lie on each distance limit and the percentage of the basins that have lower standard deviation for the propagation than the regional approach is illustrated. About 6% of the basins are in the $D_{lim} = 0.4$ corresponding to the area ratio of 4.05 and 16.74% for the area ratio equal to 7.39. Lower distances to the donor correspond to the lower uncertainties meaning lower standard deviation for the propagation compared to the regional model. In this case lower number of basin pairs are placed in the distance. For the whole dataset ($D_{lim} = 6.4$) that means unbounded validity domain that includes all the basin pairs, only 0.44% of the propagation estimate can result in lower standard deviation than the regional which can be easily ignored.

Scatter plots of the mean error for the whole pairs that lie on the distance limit are shown in Figure 14. By increasing the distance limit the number of pairs in that limit increases but the uncertainty goes up as well. The up-diagonal points in the graph are basins that have proved to be an improvement over the regional estimates. On the other hand, the point placed below the line show higher error of the propagation in relation to the regional model.



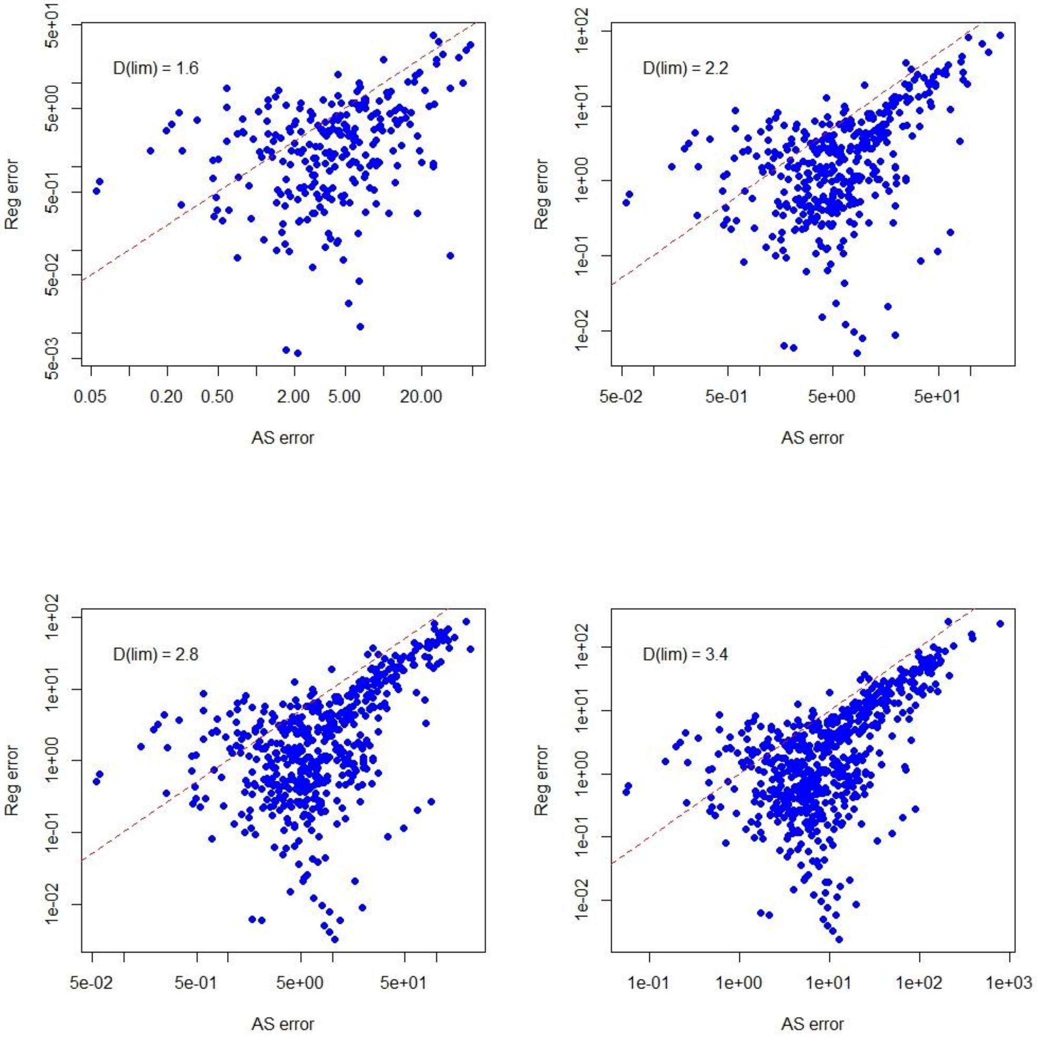


Fig. 14. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log{(A_{max}/A_{min})}$

3-2-2- Distance based on Euclidean distance

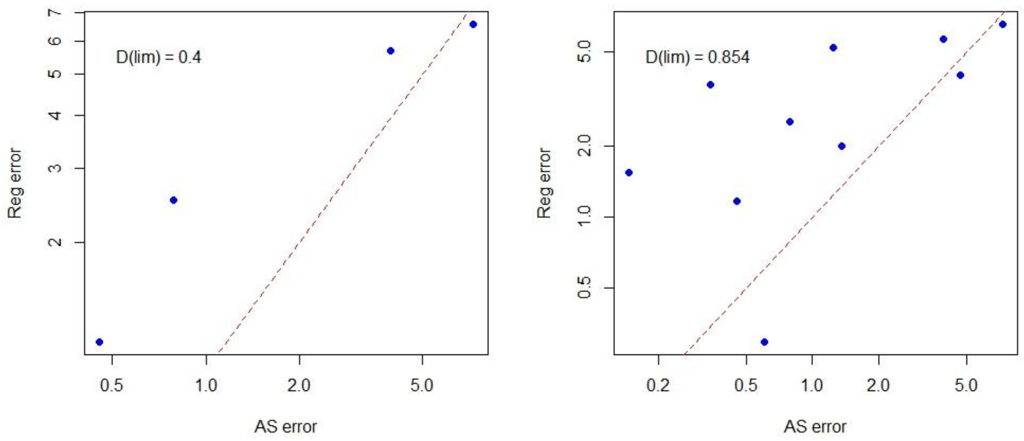
The same procedure of trial-error is carried out for the distance limit defined by the Euclidean distance of the basin pairs. The outputs for the different D_{lim} are mentioned in Table 6.

Table. 6. Results obtained from the first definition of distance, $D = \log\left(Distance_{pair}\right)$

	Results										
$\boldsymbol{D_{lim}}$	0.4	0.85	1.30	1.76	2.21	2.67	3.12	3.57	4.03	4.48	4.94
$\boldsymbol{\alpha}$	3.77	8.06	13.67	12.76	17.98	14.33	15.28	19.42	28.10	41.55	88.59
$\%\boldsymbol{Basins}$	0.46	1.17	2.34	4.68	9.13	18.03	30.44	48.47	69.55	91.34	100
$\%\boldsymbol{Basins}\ (\boldsymbol{\sigma_{P_t} \leq \sigma_{R_t}})$	100	100	25	12.5	5.12	3.24	1.92	0.96	0.67	0.51	0.47

Same starting distance limit value (0.4), like the first distance definition, is chosen for this type so as be comparable. Only 0.46% of the basins lie in this limit but all of them have standard deviation lower than the regional model. The same is true for the 1.17% of the basin. But the standard deviation percentage drop substantially from the third limit and continues to reach almost 0.47% for the whole basins.

Scatter plots of the mean errors are illustrated in Figure 15. The higher errors for the AS approach are shown in this threshold definition with increasing the limit but the point are more or less closer to the diagonal line showing relatively lower errors.



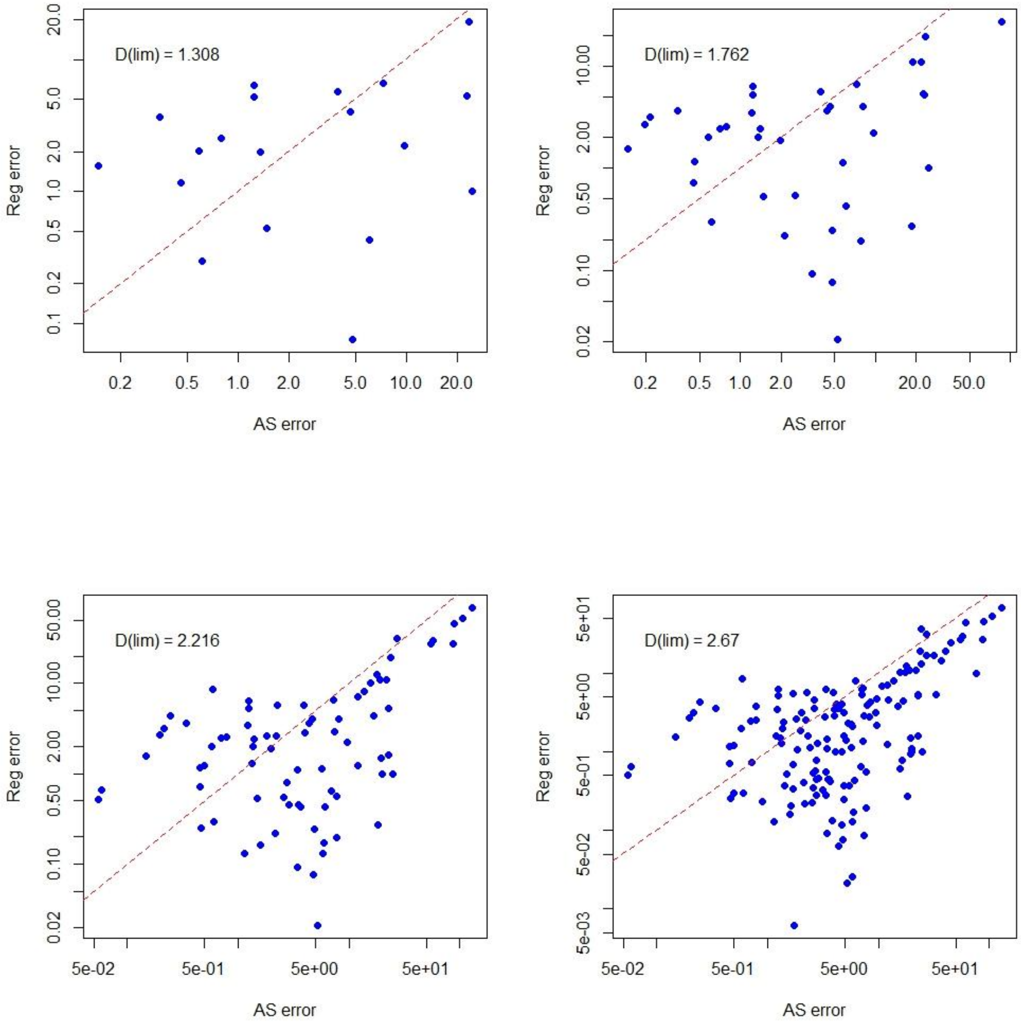


Fig. 15. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log{(Distance_{pair})}$

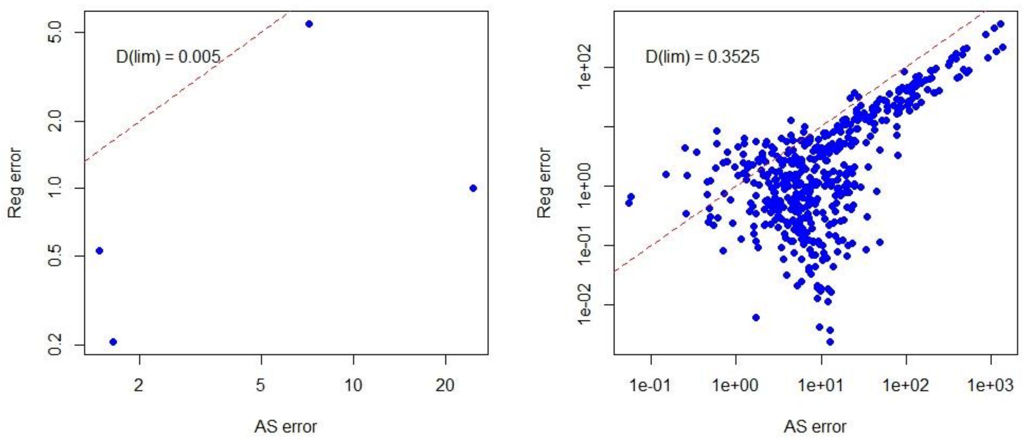
3-2-3- Distance based on Basin Mean Elevation

Starting value for distance limit for this type needs to be chosen in small numbers since from the distance limit of 0.35 more than half of the pairs lie in this distance as seen in Table 7. For the $D_{lim} = 0.005$ only 0.44 percent of the basin pairs stand in this limit, as also can be illustrated in the scatter plot in Figure 19. In addition, just 25 percent of the basins lie with the standard deviation of the predicted target values smaller than regional ones suggesting a high error would arise for this distance definition. This definition for the distance results in different behavior for the points (pair) distribution that may result from the average elevation considered that cannot properly account for the plain and mountain areas.

Table. 7. Results obtained from the first definition of distance, $D = \log{(H_{max}/H_{min})}$

	Results										
$\boldsymbol{D_{lim}}$	0.005	0.35	0.7	1.04	1.39	1.74	2.09	2.43	2.78	3.13	3.48
$\boldsymbol{\alpha}$	1817	1355	1234	1206	1187	1161	1147	1137	1134	1131	1129
$\%\boldsymbol{Basins}$	0.44	54.91	79.01	88.16	90.84	94.64	96.87	98.43	98.88	99.33	100
$\%\boldsymbol{Basins}\ (\boldsymbol{\sigma_{P_t} \leq \sigma_{R_t}})$	25	0.2	0.14	0.126	0.122	0.118	0.23	0.226	0.225	0.224	0.223

Scatter plots in Figure 16 also proving this fact with the highest percentage of the pairs stay below the diagonal line indication higher errors for the AS method. With increasing the threshold distance, the number of pairs increases but result in higher errors for AS approach confirming the non-usability of the AS approach for large distances and instead the appropriateness of regional method for the ungauged basin.



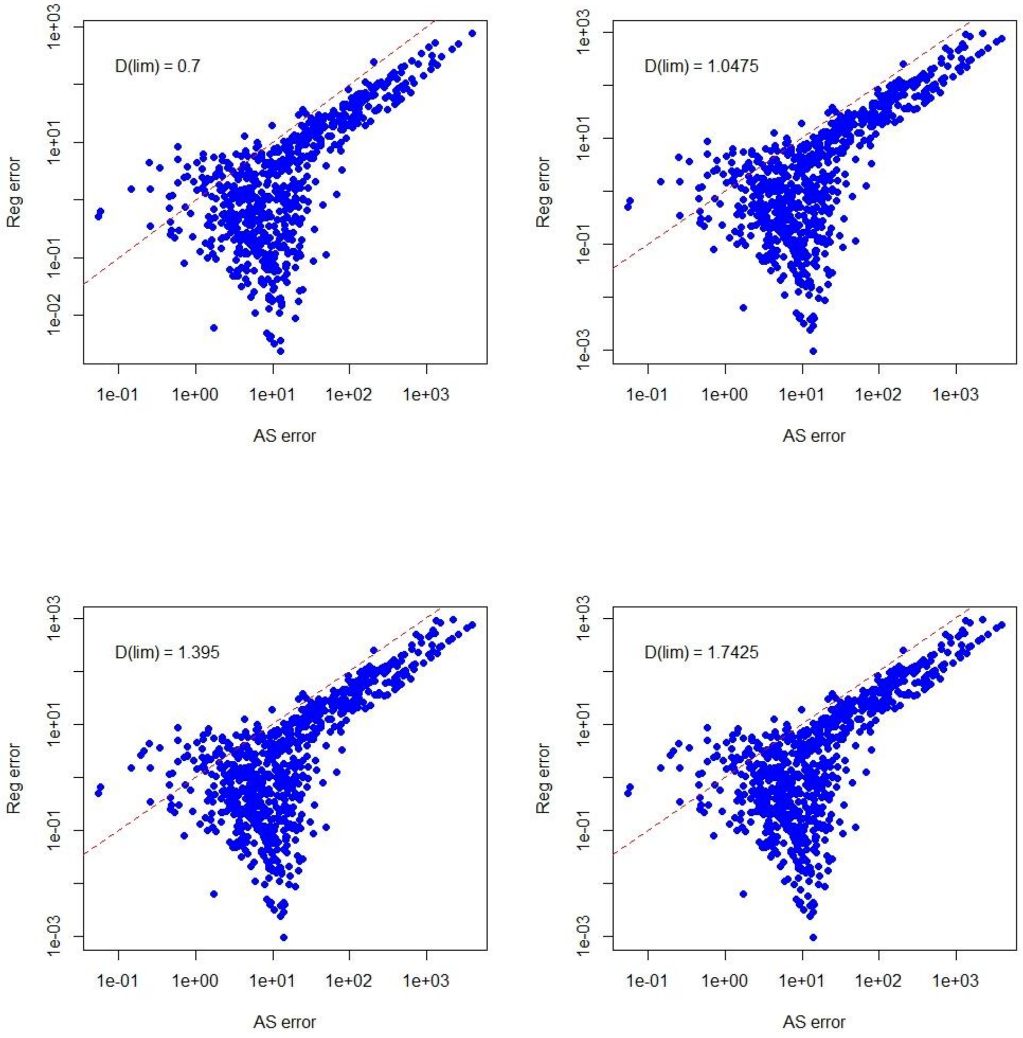


Fig. 16. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log{(H_{max}/H_{min})}$

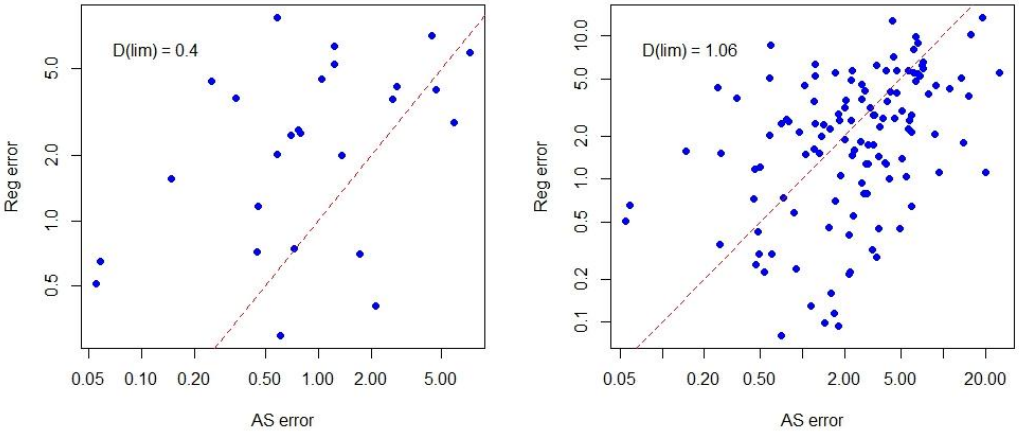
3-2-4- Distance based on Basin Area and Mean Elevation

For the last type which is a mixture of basin area and elevation, distance limits are defined starting from the 0.4 which involves 2.90 percent of the basins with almost 89 percent of them having propagated standard deviation lower than the regional in Table 8. A decreasing trend is seen for the standard deviations, so, for the whole basins only 0.44 percent having smaller propagation standard deviation.

Table. 8. Results obtained from the first definition of distance, $D = \log(A_{max}/A_{min} \cdot H_{max}/H_{min})$

	Results										
$\boldsymbol{D_{lim}}$	0.4	1.06	1.72	2.38	3.04	3.7	4.36	5.02	5.68	6.34	7.0
$\boldsymbol{\alpha}$	5.53	6.93	7.09	10.15	13.04	16.82	24.94	31.97	41.36	53.83	59.27
$\%\boldsymbol{Basins}$	2.90	13.83	25	37.05	50.22	65.17	76.56	85.49	93.97	98	100
$\%\boldsymbol{Basins}\ (\boldsymbol{\sigma_{P_t} \le \sigma_{R_t}})$	88.46	38.70	21.42	8.13	4.22	2.91	1.45	0.78	0.47	0.45	0.44

These trends can also be found in Figure 17, scatter plots of normalized error defined in equation 19. For the distance limit 0.4 most of the pairs lie above the diagonal line showing higher error for the regional. For the higher limits the errors stabilize and by increasing the limit, the normalized error increases for the AS approach. There are also some points far from the other scattered points around the diagonal line. These can be the pairs with the donor very far from the target indicating the higher error for AS method in estimating the discharge value.



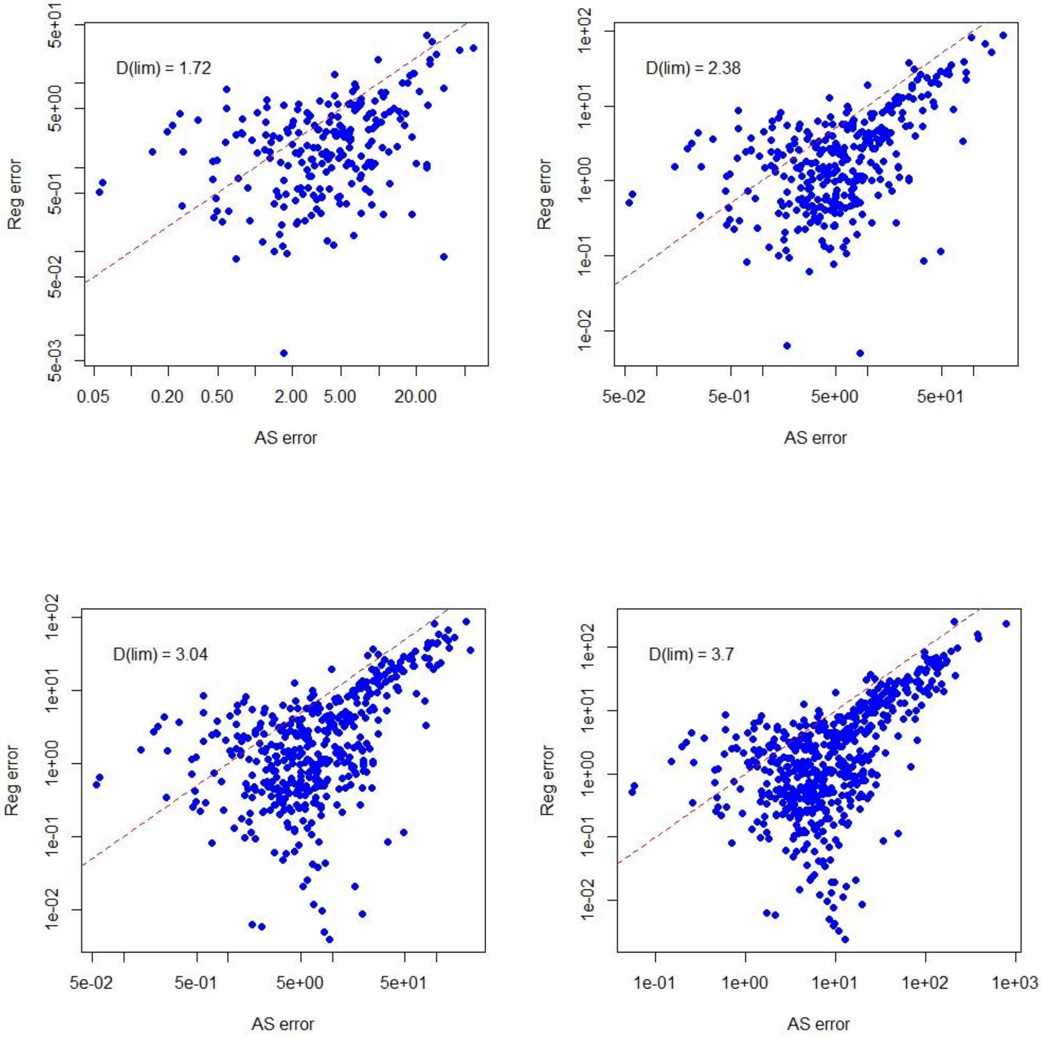


Fig. 17. Scatter log-log plots of the regional error and the Along-Stream (AS) approach for different D_{lim} for $D = \log\,(A_{max}/A_{min} \cdot H_{max}/H_{min})$

3-3- Comparison of the Results

In order to find the optimum value for the distance limit for the operation the multiplication of the basins lie within the limit and the basins with the STD of the target lower than the regional is carried out. In figure 18 the graphs for the different distance limit definitions are shown. As seen in this figure, for the limit defined by the basins' average elevation, the behavior is completely different. It is related to fact that the distribution of the distance defined by average elevation is not homogenous, shown in Figure 19, and most of the basins lie in the almost two limits.

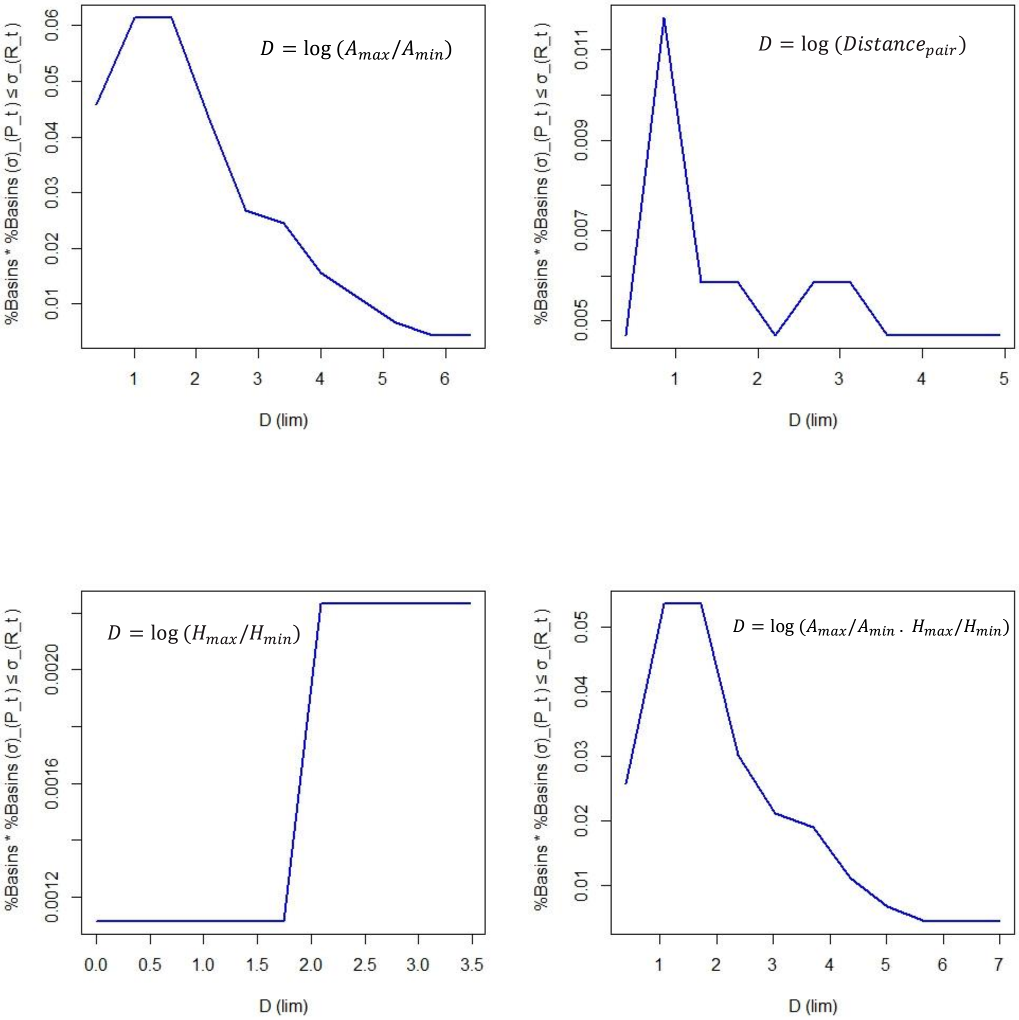


Fig. 18. Graph of percentage of basins multiplied by the percentage of basins with propagation standard deviation lower than the regional model for different D_{lim} and different distance definitions

This can also be seen in Table 7 that from the limit of 0.35 a sudden jump happens for it. This can be relevant to the elevations of the basins that results in *D* stays between 0 and 0.5 (Elevation ratio between 1 and 1.64) meaning mountainous areas cannot be taken into account related to the plain areas. So, this definition can be considered as unrepresentative of the basins. As a result. the optimum distance limit and the error analysis is not carried out for it.

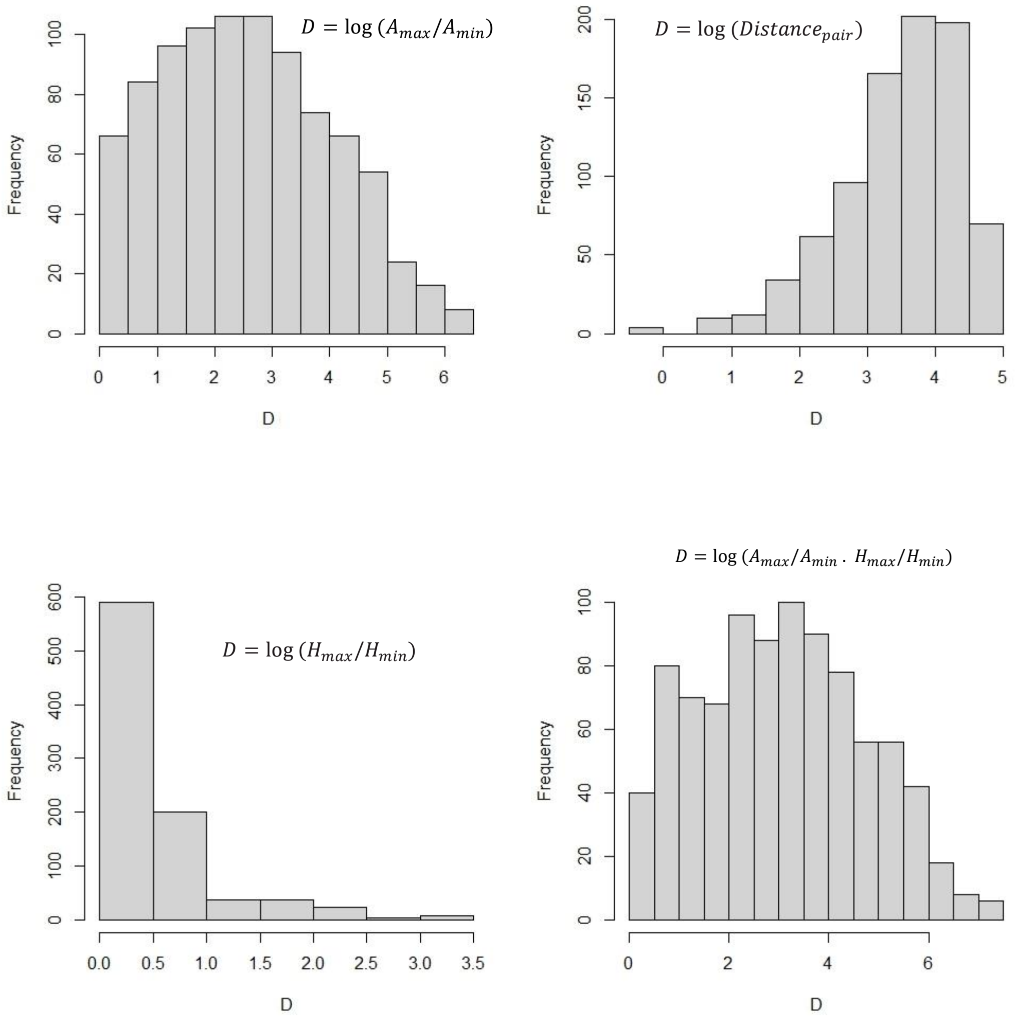


Fig. 19. Distribution of the distance definitions

From the graphs in Figure 18, the optimum values of D_{lim} are 1.6, 0.85, 1.72 for the area, Euclidean distance and multiplication of elevation and area, respectively. This implies the area ratio of 4.95, distance of 2.34 km and area ratio multiplied by elevation ratio of 5.58. The Root Mean Square Error of the different approaches for distance definitions are illustrated in Figure 20. In order to comparison the RMSE taking into account the regional prediction over the whole data set (Regional_mean) was considered. The curves provide a clear increasing behavior as the threshold distance increases. The operational method indeed proves itself to work better than simple propagation in that the operational always yields lower results than the propagated, meaning the selection criterion based on standard deviation of the propagated and regional estimates, works well. In other words, this is confirmation that, on average, the operational model is a good representative of a method to choose from the two approaches (regional or propagated). These results have underscored the improvement in effectiveness that results from using a limited validity domain in propagation of information and hence in the whole AS framework. However, it restricts the applicability of the whole AS approach to the basins that are close targets. So, the most suitable distance threshold is a compromise between two contradicted effects. The first effect penalizes smaller negative D_{lim} values, which allow better estimation results, since the applicability of the AS method would then be limited to smaller percentages of basins. The second effect enforces that the greater the domain of validity, the more pronounced would be the errors and therefore quite the opposite in terms of the operational estimator's effect. In conclusion, these graphs show the same results for the optimum value for the distance definitions, as shown in figure 18, but even better enhancement for the D_{lim} based on Euclidean distance providing D_{lim} = 1.3 is obtained by the RMSE. Overall, the definition based on the blending of basin area and elevation result has the highest D_{lim}, following by the definition solely by the area and then the Euclidean distance.

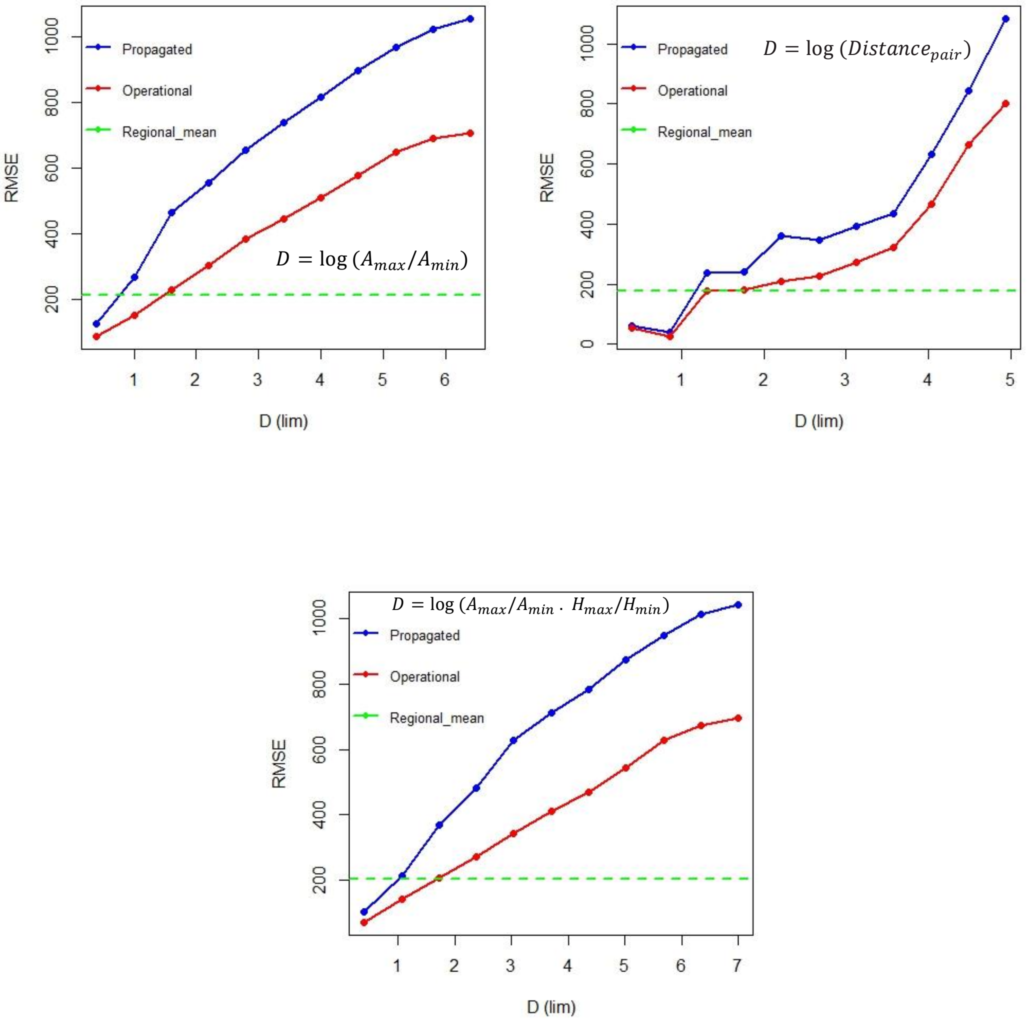


Fig. 20. Root Mean Square Error of the propagated, operational and mean of regional for different distance thresholds

In Figure 21, the RMSE of all the distance definitions are illustrated by considering the percentage of the basins that each definition can provide. As basin percentage increases, the RMSE generally rises for all lines. This suggests that as more of the basin data is considered, the error (RMSE) increases, which shows that the approaches may struggle with higher basin coverage. For the area ratio limit, multiplications of the area ratio and elevation ration and Euclidean distance, about 30, 25 and 10 percent of the basins are included, respectively. These percentages are different with the ones mentioned in the result tables since these are obtained by considering the mean RMSE of the regional model in Figure 21.

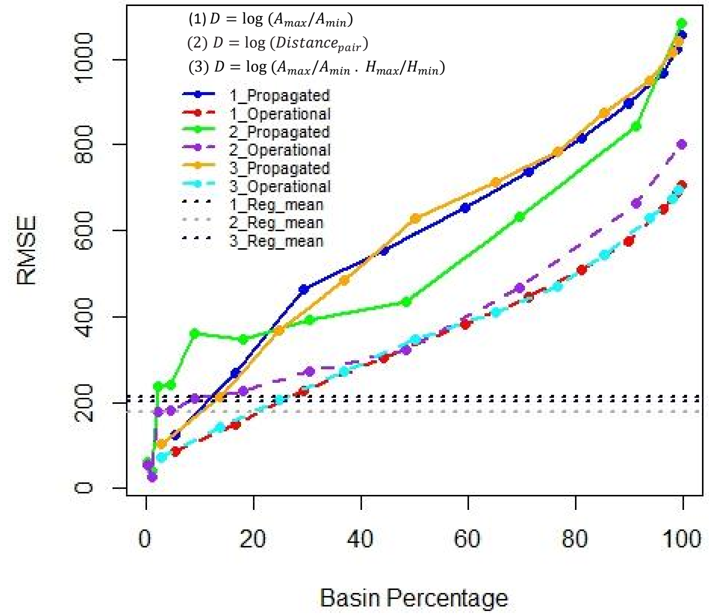


Fig. 21. Root Mean Square Error of the propagated, operational and mean of regional for different basin percentage

Scatter plots for the Normalized errors are shown in Figure 22, after choosing the optimum distance limits. These graphs show the errors and also the percentage of the basins, visually, which were mentioned earlier. All points show pairs with RMSE less than the regional mean. Points under the line are those basins that fall inside D_{lim} and where the propagated estimate has smaller errors than the regional model. Conversely, the areas above the line are basins that lie within D_{lim}, where the error in the propagated estimate exceeded that of the regional one. Most of the off-diagonal points lie at the bottom part of the plot, indicating that when the propagated estimate is considered applicable, it performs better than the corresponding regional estimate, resulting in the lower RMSE for the propagation method than the RMSE for the regional approach.

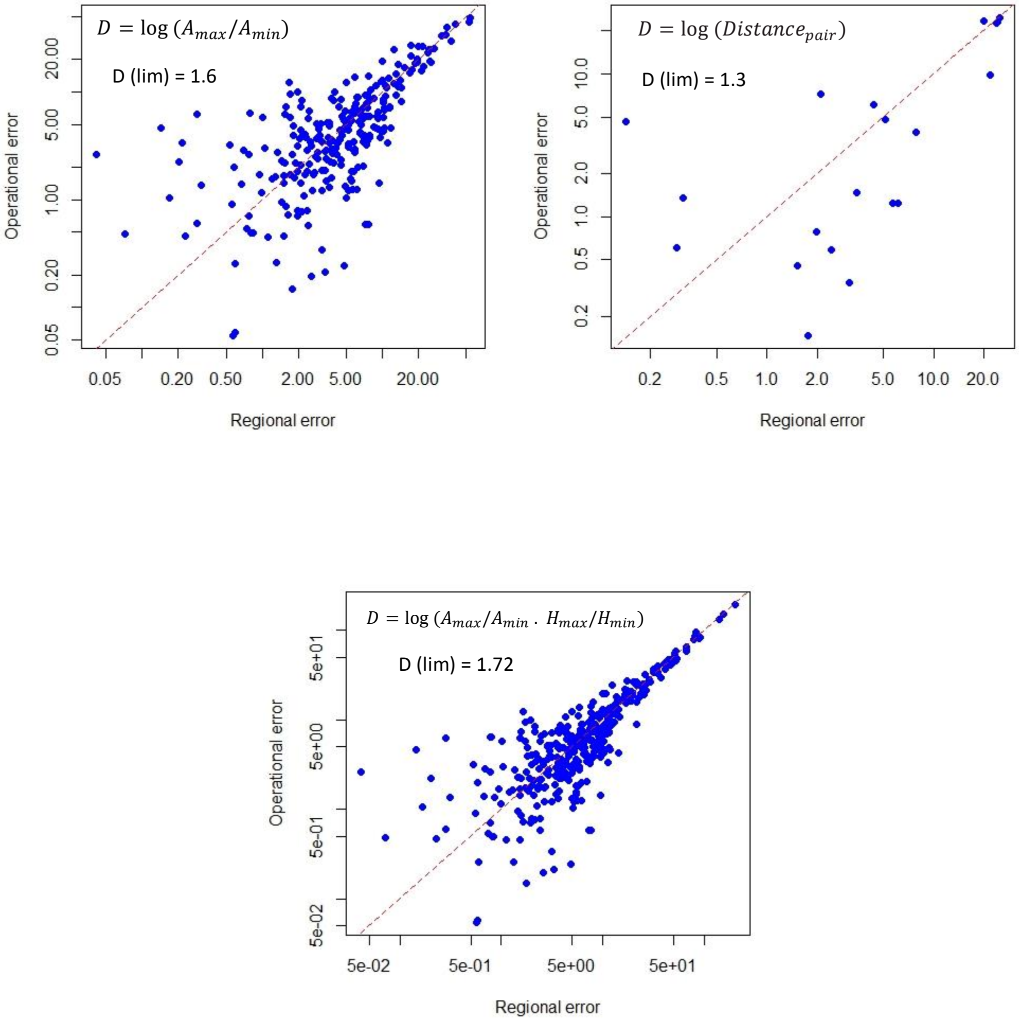


Fig. 22. Absolute Normalized Mean Error of the operational and mean of regional for optimum distance thresholds

4- Conclusion

Using an Along-Stream Estimation approach, this study aimed at the improvement of flood prediction methods for ungauged basins with a particular focus on identifying optimal distance thresholds and finally reducing the uncertainty propagated in their flood estimates. The comparison results of the different distance definitions provided valuable insights into the efficiency of various available methods for defining the valid domain for flood estimation.

The comparison of the distance definitions, as detailed in Section 4, showed that the choice of distance threshold had a great impact on the performance of the ASE method. The three different distance definitions based on basin area, Euclidean distance and combined area-elevation distance showed different behavior in terms of percentage of basins incorporated within the distance limits and standard deviation of the propagated estimates.

For the area ratio definition, the optimum distance limit turned out to be D_{lim}=1.6, area ratio 4.95. This limited distance allowed trade-off between reduced estimation errors and the adoption of a reasonable percent of basins in the valid domain to secure about 30% of basins falling within this limit. This definition also yielded the least RMSE for the ASE method as it minimized the propagation error.

The Euclidean distance definition registered a different trend where the optimum distance limit was D_{lim}=1.3 representing a distance of 3.67 km. About 10% of the basins fell under this threshold.

On the contrary, the combined area and elevation definition produced the validity domain and optimum distance limit D_{lim}=1.72, which equals area ratio times elevation ratio of 5.58. This distance limit conferred around 25% of the basins to fall within the limit. Thus, despite it including more basins in the estimation process, the RMSE for this method was slightly more than that for the area ratio distance definitions.

5- References

Ahmed A, Yildirim G, Haddad K, Rahman A (2023). Regional Flood Frequency Analysis: A Bibliometric Overview. *Water*. 15(9):1658. https://doi.org/10.3390/w15091658

Ali, S., Rahman, A (2022). Development of a kriging-based regional flood frequency analysis technique for South-East Australia. Nat Hazards, 114, 2739–2765. https://doi.org/10.1007/s11069-022-05488-4

Chen, Y. D., Huang, G., Shao, Q., & Xu, C. Y (2006). Regional analysis of low flow using L-moments for Dongjiang basin, South China. *Hydrological Sciences Journal*, *51*(6), 1051–1064. https://doi.org/10.1623/hysj.51.6.1051

Cunnane C (1988). Methods and merits of regional flood frequency analysis. J Hydrol 100:269–290.

Dalrymple T (1960). Flood frequency analyses. U.S. Geological Survey Water Supply Paper 1543A.

Dalrymple T (1960). Flood frequency methods. U.S. Geological Survey Water Supply Paper 1543A.

Durrans SR, Tomic S (1996). Regionalization of low-flow frequency estimations: an Alabama case study. Water Resour Bull 32(1):23–37.

Esmaeili-Gisavandani H, Zarei H, Fadaei Tehrani MR (2023). Regional flood frequency analysis using data-driven models (M5, random forest, and ANFIS) and a multivariate regression method in ungauged catchments. Appl Water Sci 13:139. https://doi.org/10.1007/s13201-023-01940-3

Ganora D, Laio F, Claps P (2013). An approach to propagate streamflow statistics along the river network. Hydrol Sci J 58(1):41–53. https://doi.org/10.1080/02626667.2012.745643

Gottschalk L (1993a). Correlation and covariance of runoff. Stochastic Hydrol Hydraul 7(2):85–101.

Gottschalk L (1993b). Interpolation of runoff applying objective methods. Stochastic Hydrol Hydraul 7(4):269–281.

Gottschalk L, Krasovskaia I, Leblois E, Sauquet E (2006). Mapping mean and variance of runoff in a river basin. Hydrol Earth Syst Sci 10(4):469–484.

Hosking JRM, Wallis JR (1993). Some statistics useful in regional frequency analysis. Water Resour Res 29(2):271–281.

Hosking JRM, Wallis JR (1997). Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge, UK.

Institute of Hydrology (1999). Flood estimation handbook. Wallingford, UK: Institute of Hydrology.

Kjeldsen TR, Jones D (2007). Estimation of an index flood using data transfer in the UK. Hydrol Sci J 52(1):86–98.

Kjeldsen TR, Smithers JC, Schulze RE (2000). Regional flood frequency analysis in KwaZulu-Natal Province, South Africa, using the index flood method. Unpublished report, School of Bioresources Eng. and Environ. Hydr., University of Natal, RSA.

Kjeldsen TR, Smithers JC, Schulze RE (2001). Flood frequency analysis at ungauged sites in the KwaZulu-Natal Province, South Africa. Water SA 27(3):315–324.

Kumar R, Chatterjee C, Kumar S, Lohani AK, Singh RD (2003). Development of regional flood frequency relationships using L-moments for Middle Ganga Plains Subzone 1(f) of India. Water Resour Manag 17(4):243–257.

Laio F, Ganora D, Claps P, Galeati G (2011). Spatially smooth regional estimation of the flood frequency curve (with uncertainty). J Hydrol 408:67–77. https://doi.org/10.1016/j.jhydrol.2011.07.022

Lakshmi V (2004). The role of satellite remote sensing in the prediction of ungauged basins. Hydrol Process 18:1029–1034.

Lim YH, Lye LM (2003). Regional flood estimation for ungauged basins in Sarawak, Malaysia. Hydrol Sci 48:1.

Madsen M, Pearson CP, Rosbjerg D (1997). Comparison of annual maximum series and partial duration methods for modelling extreme hydrologic events. 2. Regional modelling. Water Resour Res 33(4):759–769.

Marchi E (1994). Hydraulic aspects of the Po River flood occurred in 1951. In: Proceedings of the XVII Conference on Historical Studies, Rovigo, 2–24 November 1991. Minnelliana Editions, Rovigo.

Mazvimavi D, Meijerink AMJ, Stein A (2004). Prediction of base flows from basin characteristics: a case study from Zimbabwe. Hydrol Sci J 49(4):703–715.

Mkhandi S (1995). Choosing a distribution for flood frequency analysis. 7th South African National Hydrology Symposium, Grahamstown, RSA.

Mkhandi S, Kachroo S (1997). Regional flood frequency analysis for Southern Africa. Southern African FRIEND, Technical Documents in Hydrology No. 15, UNESCO, Paris, France.

Montanari A (2012). Hydrology of the Po River: looking for changing patterns in river discharge. Hydrol Earth Syst Sci 16:3739–3747. https://doi.org/10.5194/hess-16-3739-2012

Parida BP, Kachroo RK, Shrestha DB (1998). Regional flood frequency analysis of Mahi-Sabarmati basin (subzone 3-a) using index flood procedure with L-moments. Water Resour Manag 12:1–12.

Pearson CP (1991). New Zealand regional flood frequency analysis using L-moments. J Hydrol (NZ) 30(2):53–64.

Pearson CP (1995). Regional frequency analysis of low flows in New Zealand rivers. J Hydrol (NZ) 33(2):94–122.

Pearson CP, McKerchar AI, Woods RA (1991). Regional flood frequency analysis of Western Australian data using L-moments. In: International Hydrology and Water Resources Symposium, Australia, pp 631–632.

Piccoli A (1976). The most important floods of the Po River from 1900 to 1970. Accademia Nazionale dei Lincei, Rome.

Pitlick J (1994). Relation between peak flows, precipitation, and physiography for five mountain regions in the Western USA. J Hydrol 158:219–240.

Portela MM, Dias AT (2005). Application of the index-flood method to the regionalization of flood peak discharges on the Portugal mainland. In: Brebbia CA, Antunes do Carmo JS (eds) River Basin Management III. WIT Press, Southampton.

Potter KW, Lettenmaier DP (1990). A comparison of regional flood frequency estimation methods using a resampling method. Water Resour Res 26(3):415–424.

Saf B (2009). Regional flood frequency analysis using L-moments for the West Mediterranean Region of Turkey. Water Resour Manag 23:531–551. https://doi.org/10.1007/s11269-008-9287-z

Saf B, Dikbas F, Yasar M (2007). Determination of regional frequency distributions of floods in West Mediterranean River Basins in Turkey. Fresenius Environ Bull 16(10):1300–1308.

Sivapalan M, Takeuchi K, Franks SW, Gupta VK, Karambiri H, Lakshmi V, Liang X, McDonnell JJ, Mendiondo EM, O'Connell PE, Oki T, Pomeroy JW, Schertzer D, Uhlenbrook S, Zehe E (2003). IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol Sci J 48(6):857–880.

Skoien JO, Merz R, Bloschl G (2006). Top-kriging – geostatistics on stream networks. Hydrol Earth Syst Sci 10(2):277–287.

Uhlenbrook S (2006). Catchment hydrology – a science in which all processes are preferential – Invited commentary. Hydrol Process 20(16):3581–3585. https://doi.org/10.1002/hyp.6564

Visentini M (1953). The latest floods of the Po River. In: Proceedings of the XVIII International Conference on Navigation, Rome.

Vogel RM, Thomas WO Jr, McMahon TA (1993). Flood-flow frequency model selection in Southwestern United States. J Water Resour Plann Manag 119(3):353–366.

Vogel RM, Wilson I (1996). Probability distribution of annual maximum, mean, and minimum streamflows in the United States. J Hydrol Eng 1(2):69–76.

Watt WE, et al. (1989). Hydrology of floods in Canada: a guide to planning and design. National Research Council Canada, Associate Committee on Hydrology.

Zanchettini D, Traverso P, Tomasino M (2008). Po River discharge: a preliminary analysis of a 200-year time series. Climatic Change 88:411–433. https://doi.org/10.1007/s10584-008-9395-z

6- Appendices

6-1- Appendix I: List of the basins in the whole dataset

Code	Name	E_delimita	N_delimita	Years of data	Area	Mean Elevation
A001	Adda a Tirano	590211	5118616	13	903.55	2158
A002	Adda a Fuentes	534567	5110177	100	2576.09	1847
A007	Agogna a Novara	467981	5030915	21	400.69	332
A008	Arda a Mignano (diga del serbatoio)	563096	4957643	77	88	755
A009	Artanavaz (Dora Baltea) a St Oyen	360765	5075765	14	70.02	2210
A010	Aveto (Trebbia) a Cabanne	527598	4927096	35	41.09	983
A011	Ayasse (Dora Baltea) a Champorcher	392220	5052946	22	41.53	2364
A012	Banna a Santena	403910	4977515	23	351.6	286
A013	Belbo a Castelnuovo Belbo	454039	4960718	23	422.82	373
A014	Bevera a Colombaio	520474.2	5069077.2	17	36.95	390
A015	Borbera (Scrivia) a Baracche	500788	4951811	22	202.24	864
A016	Borbore a San Damiano d'Asti	427132	4965795	21	84.59	240
A017	Bormida a Cassine (Caranzano)	463786	4955433	40	1516.04	490
A018	Bormida ad Alessandria	472015	4972514	27	2657.6	398
A019	Bormida di Mallare a Ferrania	446089	4912508	24	51.2	603
A020	Bormida di Millesimo a Cessole	440472	4944272	20	494.96	644
A021	Bormida di Millesimo a Murialdo	432976	4907028	19	135.18	878
A022	Bormida di Spigno a Mombaldone	447206	4935280	22	390.65	485
A023	Bormida di Spigno a Valla	447475	4932026	52	66.84	464
A024	Bousset a Tetti Porcera	375046	4896152	6	38.48	1983
A025	Brembo a Ponte Briolo	547267	5067317	79	749.04	1181
A026	Breuil a Alpette	336980	5064201	14	27.7	2442
A027	Bucera a Ponte Rovine	370555	4896060	6	27.59	2117
A029	Cannobino a Traffiume	475222	5100470	27	106.45	1096
A030	Cervo (Sesia) a Passobreve	425027	5053441	32	75.62	1484
A031	Cervo a Vigliano Biellese	430621	5044910	19	129.65	1254
A032	Cervo a Quinto Vercellese	451013	5025503	19	995.23	508
A033	Chiavanne a Alpette	336855	5064354	14	22.71	2476
A034	Chiese Malga Bissina	617738.1	5101290.6	38	47.86	2447
A035	Chiese a Malga Boazzo Interbacino	618050.7	5094937.6	31	100.41	2262
A037	Chiese a Gavardo	616818	5053974	75	886.07	1248
A038	Chisone a Soucheres Basses	339221	4988138	18	94.58	2212
A039	Chisone a Fenestrelle	346006	4988947	19	153.51	2147
A040	Chisone a S. Martino	364406	4971647	61	581.67	1724
A041	Chiusella a Gurzia	402638	5030792	34	143.48	1344

A043	Corsaglia (Tanaro) a Presa Centrale Molline	407096	4904936	25	89.03	1520
A044	Dolo (Secchia) a Fontanaluccia (diga del serbatoio)	620789.2	4903965.7	21	40.99	1314
A047	Dora Baltea a Ponte di Mombardone	344193	5069807	14	369.68	2395
A048	Dora Baltea ad Aosta	374507	5065996	19	1861.51	2250
A049	Dora Baltea a Tavagnasco	408421	5044504	97	3309.22	2081
A051	Dora Baltea a Mazz?	417900	5017085	71	3845.9	1874
A052	Dora di Bardonecchia a Beaulard	323449	4990582	12	197.86	2202
A053	Dora di Courmayeur a San desiderio Terme	343253	5070108	7	219.92	2443
A054	Dora di Rhemes a Palaud	352495	5046587	5	55.26	2714
A055	Dora di Rhemes a Notre Dame	353389	5048222	14	70.1	2653
A056	Dora di Rhemes a Saint Georges	356155	5057480	6	119.62	2475
A057	Dora Riparia a Ulzio (Oulx)	329337	4988965	53	260.6	2161
A061	Dora Riparia a Torino Ponte Washington	399113	4992257	27	1320.11	1645
A063	Enza a Sorbolo	614932	4966627	32	661.61	426
A064	Erro (Bormida) a Sassello	456380	4926885	18	96.81	593
A065	Evancon a Champoluc	400645	5075837	21	103.7	2625
A068	Gesso della Barra a San Giacomo	371027	4892416	6	19.34	2099
A069	Gesso della Valletta (Stura di Demonte)	368094	4901309	11	111.57	2095
A071	Gesso di Entracque (Stura di Demonte)	371071	4901660	12	159.16	1870
A072	Gesso di Monte Colombo a San Giacomo	371309	4892550	5	24.51	2171
A073	Grana a Monterosso	366585	4918708	65	102.7	1554
A074	Grand'Eyvia a Cretaz	369982	5053186	10	181.85	2569
A075	Isorno a Pontetto	448127	5111126	16	69.88	1622
A077	Lys a d'Ejola	407796	5078893	10	29.27	3092
A078	Lys (Dora Baltea) a Gressoney St. Jean	408689	5071087	18	91.16	2624
A079	Lys a Guillemore	411124	5058000	29	202.32	2242
A080	Maira a Saretto	335576	4927181	17	54.19	2419
A081	Maira a San Damiano Macra	361439	4927199	57	452.16	1888
A082	Maira a Racconigi	394538	4957991	20	976.52	1316
A083	Malone a Brandizzo	409949	5004166	18	330.45	431
A084	Malone a Front	395360	5015295	25	122.58	677
A085	Marmore a Perreres	392605	5084497	15	56.43	2692
A086	Mastallone (Sesia) a Ponte Folle	442190	5075482	53	147	1318
A087	Melezet a Melezet	315867	4991349	15	43.19	2380
A088	Meris a Sant'Anna Valdieri	365656	4900453	5	23.55	2103
A092	Nontey a Valnontey	370540	5049494	5	52.48	2783
A093	Oglio a Tem?	613626	5122789	5	123.63	2204
A098	Orba a CasalCermelli	470871	4964452	24	766.79	454
A099	Orco a Pont Canavese	391506	5030005	49	611.98	1919

A100	Orco a San Benigno Canavese	406313	5011059	21	835.87	1550
A101	Panaro a Bomporto	661070	4953498	60	1064.27	619
A102	Parma a Ponte Bottego	604868.6	4962637.2	45	596.89	661
A104	Pellice a Luserna San Giovanni	361138	4963467	8	215.62	1618
A105	Pellice a Villafranca Piemonte	381346	4963166	20	975.4	1141
A106	Po a Crissolo	354624	4950851	24	36.56	2223
A107	Po a Carignano	396682	4973652	26	3934.39	1094
A108	Po a Moncalieri (Meirano)	395654	4983807	78	5114.72	918
A109	Po a Torino Murazzi	397499	4990961	24	5353.17	905
A110	Po a S.Mauro Torinese	404376	4998217	26	7682.01	1073
A111	Po a Casale Monferrato	456575	4998852	31	13694.72	1250
A122	Rio Bagni (Stura di Demonte) a Bagni di Vinadio	347299	4905908	20	61.4	2128
A124	Rio Freddo a Rio Freddo	354141	4904744	7	36.69	2124
A125	Ripa a Bousson	327797	4977949	5	145.77	2339
A128	Rutor (Dora Baltea) a Promise	340966	5063082	61	50.39	2508
A130	San Bernardino a Trobaso	464983	5088294	50	129.67	1188
A131	San Giovanni a Possaccio	465266	5089297	14	54.01	979
A132	Sarca a Ponte Plaza	640275	5117680	27	72.01	2009
A133	Sarca Saone	636752.6	5100794.5	10	538.12	1856
A134	Sarca Nago	645059.1	5084439.8	14	1065.81	1477
A135	Sarca di Nambron a Pian di Nambron	635264	5118637	44	21.15	2338
A137	Sarca di Val Genova opera presa	633312	5114358	21	141.43	2353
A138	Savara a Eau Rousse	360294	5047742	18	81.16	2686
A139	Savara a Fenille	359591	5055032	6	132.11	2601
A140	Scrivia a Isola del Cantone	496640	4943533	13	217.17	662
A141	Scrivia a Serravalle	488989	4952417	50	615.98	681
A142	Secchia a Ponte Cavola	621268	4918665	32	348.7	967
A143	Secchia a Ponte Bacchello	657230	4956783	53	1389.53	651
A145	Sermenza a Rimasco	427374	5078429	44	82	1833
A146	Sesia a Campertogno	424762	5072125	44	171.33	2098
A147	Sesia a Ponte Aranco	443803	5062352	17	695.98	1497
A149	Sesia a Palestro	463769	5014565	48	2446.38	630
A150	Strona di Omegna a Gravellona Toce	456135	5086079	20	227.49	870
A151	Stura di Demonte a Pianche	349597	4907221	18	179.5	2066
A152	Stura di Demonte a Gaiola	373755	4909963	45	559.18	1811
A153	Stura di Demonte a Roccasparvera	375650	4910905	9	580.75	1780
A154	Stura di Demonte a Fossano	398628	4930850	21	1241.1	1588
A155	Stura di Lanzo a Lanzo	380982	5013879	74	580.89	1755
A156	Stura di Lanzo a Torino	398253	4996117	20	884.97	1345

A158	Stura di Vi? a Malciaussia	354204	5007535	49	24.28	2583
A159	Stura di Vi? a Usseglio	360063	5010116	11	79.81	2362
A160	Tanaro a Ponte di Nava	409392	4885563	56	135.63	1611
A161	Tanaro a Ormea	413231	4888845	13	193.49	1511
A162	Tanaro a Garessio	421310	4894772	25	249.53	1420
A163	Tanaro a Nucetto	425191	4910271	39	374.7	1220
A164	Tanaro a Piantorre	418169	4918794	24	501.21	1057
A166	Tanaro a Farigliano	412709	4929894	79	1512.76	939
A167	Tanaro ad Alba	422966	4950663	26	3391.26	1067
A171	Tanaro a Montecastello	475104	4977074	87	8023.68	651
A172	Taro a S.Maria	539339	4920075	18	29.72	1049
A173	Taro a Piane di Carniglia	548372	4925569	29	91.02	959
A174	Taro a Pradella	559398	4925585	29	295.61	830
A175	Taro a Ostia	567898	4930371	28	413.01	815
A176	Taro a S. Quirico	599021	4974500	24	1414.79	646
A177	Ticino a Bellinzona	500666	5115656	31	1525.68	1609
A187	Toce a Cadarese	450177	5125409	15	189.41	2130
A188	Toce a Domodossola	446394	5106791	18	921.87	1806
A189	Toce a Candoglia	455210	5091406	78	1520.83	1667
A190	Trebbia a due Ponti	520769	4931707	21	75.11	959
A191	Trebbia a Valsigiara	524861	4944296	63	223.6	937
A192	Trebbia a S. Salvatore	530364	4955188	17	640.44	944
A194	Varaita a Castello	345132	4941739	56	67.24	2383
A195	Varaita a Rore	358666	4937361	58	262.49	2137
A196	Varaita a Rossana	376220	4934616	21	401.55	1784
A198	Vobbia a Vobbietta	497957	4942898	14	55.56	709
A199	Bormida di Millesimo a Camerana	432905	4920644	21	263.33	762
A200	Chisola a La Loggia	395104	4980550	14	500.67	376
A201	Germanasca a Perrero	355062	4978653	19	189.23	1888
A202	Orba a Basaluzzo	473870	4957161	20	730.59	469
A203	Vermenagna a Robilante	381711	4902040	12	134.87	1532
A206	Agogna a Lomello	484222	4996705	8	691.54	280
A207	Arno a Cavaria	485026	5059567	16	31.76	342
A208	Bevera a Molteno	523872	5069985	12	31.03	410
A209	Brembo a Camerata Cornello	551079.18	5082983.36	17	395.13	1460
A211	Lambro a Caslino	518046	5075784	17	53.12	766
A218	Mella a Bovegno	598195	5070534	7	86.21	1313
A223	Olona a Pte Vedano	489791	5069163	6	87.77	436
A224	Serio a Grabiasca	573386.13	5095336.98	11	92.03	1897
A226	Seveso a Cantu Asnago	507809	5062778	14	64.99	337

A228	Staffora a Voghera	501379	4981871	13	287.49	606
A229	Terdoppio a Gambolo	489275	5010486	8	340.2	167
A230	Dora Riparia a Susa	346587	5000202	15	694.159	2033
A231	Varaita a Polonghera	388396	4961803	10	560.208	1406
B001	Bidente di Corniolo a Campigna	723259.5	4864070.4	18	19.69	991
B002	Bidente di Ridracoli a Ridracoli (diga del serbatoio)	727948.9	4861745.6	31	36.41	892
B003	Brasimone (Setta) a Santa Maria (diga del serbatoio)	672076.5	4891092.9	29	25.61	900
B007	Lamone a Sarna	724964.5	4902564.3	43	255.89	515
B008	Lamone a Grattacoppa	747176.9	4931719.3	15	528.99	424
B009	Limentra di Riola (Reno) a Stagno	663593	4886374	28	67.71	879
B010	Limentra di Sambuca (Reno) a Pavana (diga del serbatoio)	660342.5	4887059.7	41	39.53	912
B011	Limentra di Treppio (Reno) a Suviana (diga del serbatoio)	663257.1	4888925.2	22	77.52	852
B012	Orsigna (Reno) a Setteponti	653531.6	4880590.1	9	15.64	1081
B013	Quaderna (Reno) a Palesio	699264.5	4920349.2	26	22.96	272
B014	Reno a Pracchia	652825.9	4880140.7	79	41.02	908
B015	Reno a Molino di Pallone	656956	4884951.4	26	88.73	936
B016	Reno a Ponte della Venturina	659351.1	4888190.4	6	100.21	916
B017	Reno a Calvenzano	672249.4	4908845	14	588.66	722
B018	Reno a Casalecchio	682940.8	4932566.8	97	1069.54	624
B021	Rio Faldo (Reno) a Setteponti	653761	4880489	6	3.45	929
B022	Ronco (Fiumi Uniti) a Meldola Casa Luzia	745048.2	4889350.8	44	441.37	555
B023	Samoggia (Reno) a Calcara	667857.9	4934037.1	57	173.6	379
B024	Savena (Reno) a Castel dellAlpi	682163	4893848.2	21	11.79	1001
B025	Savena (Reno) a S. Ruffillo	689254.1	4926051.7	12	160.36	519
B026	Savio a Mercato Saraceno	756766.1	4873501.8	11	361.06	637
B027	Savio a San Vittore	757359.8	4889665.6	44	598.05	519
B028	Senio (Reno) a Castel Bolognese	723405	4910051	12	262.5	428
B029	Silla (Reno) a Silla	657759.8	4893963.9	22	84.02	854
B031	Baganza a Berceto	579024	4928486	16	16.95	1099
B032	Trebbia a Bobbio	530451	4956017	17	653.64	938
B034	Santerno a Borgo Tossignano	705897	4905814	27	318.53	600
B035	Tresinaro a Ca' de' Caroli	633174	4939584	7	150.36	407
B037	Pisciatello a Calisese	763838	4889950	8	38.91	209
B038	Scodogna a Casella Nuova	593819	4951458	6	11.35	262
B039	Arda a Case Bonini	561495	4955891	18	72.19	797
B040	Senio a Casola Valsenio	710324	4900670	14	135.62	584
B041	Stirone a Castellina Soragna	587705	4974433	19	110.52	189

B045	Tassobbio a Compiano	607577	4931205	18	100.65	546
B046	Parma a Corniglio	587509	4926783	9	110.21	1078
B048	Enza a Currada	610157	4933620	6	430.07	632
B049	Leo a Fanano	643829	4896288	19	64.5	1245
B050	Nure a Farini	545204	4951328	15	208.96	940
B051	Nure a Ferriere	539545.31	4943491.92	12	48.68	1128
B054	Acquicciola a Fiumalbo	631822.5	4893114.2	18	18.4	1459
B056	Secchia a Gatta	616736	4917833	12	233.61	1040
B060	Lonza a Vetto	605498	4924793	12	62.63	703
B061	Secchia a Lugo	631693	4921525	19	694.64	915
B062	Lamone a Marradi	709378	4883992	13	105.23	728
B064	Riglio a Montanaro	562988	4977073	13	87.72	316
B065	Conca a Morciano	792509	4869017	10	139.8	432
B067	Recchio a Noceto	592560	4962555	11	40.29	264
B073	Ceno a Ponte Ceno	548002	4932259	10	51.07	1101
B074	Ceno a Ponte Lamberti	564477	4944700	14	331.38	861
B076	Nure a Pontenure	560857	4984492	9	373.32	745
B077	Panaro a Ponte Samone	653305	4913477	21	584.6	936
B078	Sissola (Taro) a Pontestrambo	544887.73	4922603.88	5	16.72	1024
B079	Scoltenna a Ponte Val Sasso	645235	4903956	15	272.29	1093
B082	Crostolo a Puianello	624465	4942580	16	85.83	390
B084	Marecchia a Rimini SS16	783772	4884995	13	521.51	559
B085	Rio Cella a Querceto	690277.5	4902966.1	13	10.02	561
B087	Trebbia a Rivergaro	546040	4972249	19	916.18	838
B088	Rossenna a Prignano sulla Secchia	632814	4921190	19	185.64	715
B090	Secchia a Rubiera SS9	642457	4945882	17	1250.2	708
B091	Chiavenna a Saliceto	568420	4982637	19	90.06	215
B092	Ghiara a Salsomaggiore	577973	4963067	16	30.07	335
B093	Aveto a Salsominore	532230	4942430	17	208.73	1043
B094	Rubicone a Savignano	771960	4888100	9	38.36	177
B095	Cedra a Selvanizza	598416	4921685	15	79.92	1038
B098	Sillaro a Sesto Imolese	717740	4926669	8	247.32	242
B099	Uso a Santarcangelo	775373	4885408	15	106.83	267
B101	Tiepido a San Donnino	655846	4939318	13	51.36	256
B103	Panaro a Spilamberto	661370	4933387	11	746.5	814
B106	Taro a Tornolo	550216	4926859	18	104.53	935
B107	Reno a Vergato	668681	4906209	14	552.17	592
B108	Enza a Vetto	605825	4927738	14	298.7	763

6-2- Appendix II: R Environment and Scripts for the index flood with D = log(A(max)/A(min))

General description

The entire work and its results have been performed using R. R is a language implemented through the creation of codes and scripts in the 1990's first by Robert Gentleman and Ross Ihaka at the University of Auckland. This language is closely tied to an earlier language called S for Statistical Computational created by John Chambers, Rick Becker, and others at Bell Labs in the mid-1970s and made available further in the 1980s. Though it began back in 1995 when the introduction of the R project was established under entirely free open-source software without requiring any specific license to use and publish results achieved using the software. After an introductory phase of work under development by the R core group (where core members have access to the source code), the first public release was made in February 2000. Most researchers refer to R as a statistical system, yet the developers in the official 'Introduction to R' would prefer it to be qualified as an environment 'within which many classical and modern statistical techniques have been implemented'. Therefore, the use of 'environment' denotes its characterization as a well-planned and cohesive system, rather than as an aggregating collection of very specific and inflexible tools. In fact, R is well-known not only for the statistical analyses that can be performed but also for the graphical tools provided for high-quality plotting. Along with that, Base R offers an effective data handling and storage facility, operators for calculation on arrays (in particular, matrices), a collection of intermediate tools for data analysis, and graphical facilities for data display and the programming language itself for the implementation of conditionals, loops, and cycles.

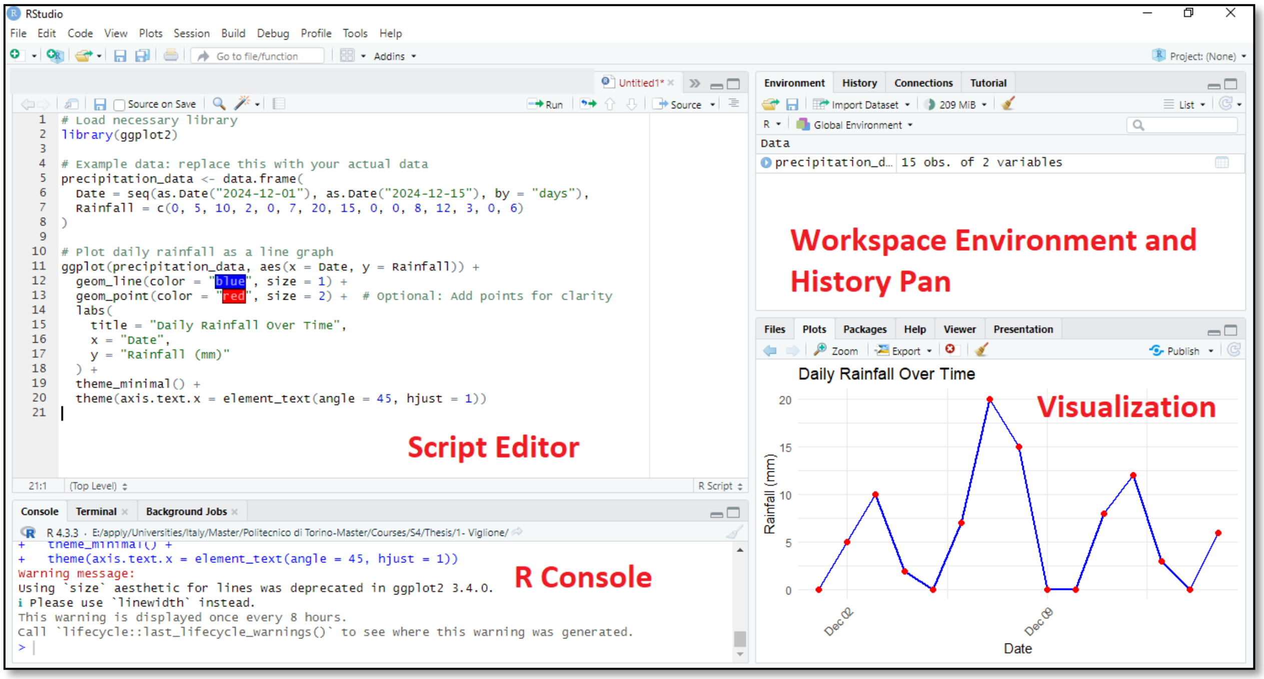
One, if not the most, powerful advantage of using R is the extended possibility of expanding the available functions and tools inside the software installation of external packages: packages are nothing except collections of additional functions and/or functionalities as developed by independent authors who decided spontaneously to contribute to the growth of open-source software. Also, the base functions contained in base R are stored inside the so-called standard packages, which are automatically loaded when the booting of the software occurs. Another set of packages (the recommended packages) are also already present in base R, but they are not automatically loaded. Every additional package has, instead, to be downloaded first and installed from one of the official sources using the function "install.packages" (the biggest repository is CRAN- Comprehensive R archive network) and then loaded for use with the dedicated function "library".

The realization of an R package has to comply with some rather stringent rules to meet the necessary requirements for a quick and easy handling by the users. In its most restricted sense, a package (which practically is a directory) comprises the actual code for the additional functions (inside a directory named R), a description of the package (containing information such as name, version, authors, and other metadata) inside a plain text file with no extension and a distilled description of the function included (that may be visualized with the "help" command in the software). Other additional elements can be present (and must be present to be compliant with the requirements to be published in CRAN) like example data or indications of required or suggested complementary packages, but they will not be here described in detail. A complete description of the creation of an entire package can be found in the "Writing R extensions" official guidelines.

R Studio

Very important to realize that R is a language that doesn't carry any graphical interface on its own (in the same way as the LINUX operative system works from line command). Since most calculations have been conducted on a device running on the Windows operating system, some additional software has been used. R Studio is just an integrated development environment for R, which has a full graphical interface for easier and more direct use (not different from other software like MATLAB).

It seamlessly integrates into a console for writing and modifying code, together with a syntax-highlighting editor and tools for plotting (with direct visualization), debugging, and workspace management (Figure 10).



```
# Load necessary library
library(ggplot2)

# Example data: replace this with your actual data
precipitation_data <- data.frame(
  Date = seq(as.Date("2024-12-01"), as.Date("2024-12-15"), by = "days"),
  Rainfall = c(0, 5, 10, 2, 0, 7, 20, 15, 0, 0, 8, 12, 3, 0, 6)
)

# Plot daily rainfall as a line graph
ggplot(precipitation_data, aes(x = Date, y = Rainfall)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +  # Optional: Add points for clarity
  labs(
    title = "Daily Rainfall Over Time",
    x = "Date",
    y = "Rainfall (mm)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
R 4.3.3 · E:/apply/Universities/Italy/Master/Politecnico di Torino-Master/Courses/S4/Thesis/1- Viglione/
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
Warning message:
Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
>
```

```
Data
precipitation_d… 15 obs. of 2 variables
```

Fig. 23. R-Studio visual interface

Scripts used for the index flood and for the $D = \log{(A_{max}/A_{min})}$.

```
library(xlsx)

Basin_data <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 13)

Basin_name <- apply(Basin_data, 1, function(x) x[!is.na(x)])

S_data <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 8)
R_data <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 10)

S_data <- apply(S_data, 2, function(col) c(col[col != ""]))
R_data <- apply(R_data, 2, function(col) c(col[col != ""]))

S_data <- lapply(S_data, function(col) as.numeric(col))
R_data <- lapply(R_data, function(col) as.numeric(col))
```

```
names(S_data) <- NULL
names(R_data) <- NULL

Basin_name1 <- unique(unlist(Basin_name))
Basin_name2 <- sort(Basin_name1)

n <- length(Basin_name2)
matrix_P_t <- matrix(NA, nrow = n, ncol = n)

rownames(matrix_P_t) <- Basin_name2
colnames(matrix_P_t) <- Basin_name2

for (i in seq_along(Basin_name)) {
  basins <- Basin_name[[i]]
  qins <- S_data[[i]]
  rs <- R_data[[i]]

  for (j in seq_along(basins)) {
    donor <- basins[j]
    qin_donor <- qins[j]
    r_donor <- rs[j]

    for (k in seq_along(basins)) {
      if (j != k) {
        target <- basins[k]
        r_target <- rs[k]
        # Apply the formula
        matrix_P_t[target, donor] <- (r_target / r_donor) * qin_donor
      }
    }
  }
}


matrix_st <- matrix(NA, nrow = n, ncol = n)
rownames(matrix_st) <- Basin_name2
colnames(matrix_st) <- Basin_name2

for (i in seq_along(Basin_name)) {
  basins <- Basin_name[[i]]
  qins <- S_data[[i]]
  rs <- R_data[[i]]

  for (j in seq_along(basins)) {
    donor <- basins[j]
    qin_donor <- qins[j]
    r_donor <- rs[j]

    for (k in seq_along(basins)) {
      if (j != k) {
        target <- basins[k]
        target_index <- match(target, Basin_name2)
        donor_index <- match(donor, Basin_name2)

        if (!is.na(matrix_P_t[target, donor])) {
          matrix_st[target_index, donor_index] <- qin_donor
        }
      }
    }
  }
}



# Plotting P(t) vs S(t) with a 45-degree line
p_values <- as.vector(matrix_P_t)
s_values <- as.vector(matrix_st)

valid_indices <- !is.na(p_values) & !is.na(s_values)
p_values <- p_values[valid_indices]
```

```
s_values <- s_values[valid_indices]

# Create the scatter plot
plot(s_values, p_values,
     main = "Scatter Plot of S(t) vs P(t)",
     xlab = "S(t)", ylab = "P(t)",
     pch = 16, col = "blue",
     xlim = range(c(s_values, p_values)),
     ylim = range(c(s_values, p_values)))

abline(a = 0, b = 1, col = "red", lty = 2)
grid()


# scatter plot
plot(s_values, p_values,
     log = "xy",
     main = "log-log Scatter Plot of S(t) vs P(t)",
     xlab = "log(S(t))", ylab = "log(P(t))",
     pch = 16, col = "blue",
     xlim = range(c(s_values, p_values)),
     ylim = range(c(s_values, p_values)))

abline(a = 0, b = 1, col = "red", lty = 2)
grid()


####  D  ######

Area <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 12)
Area <- apply(Area, 2, function(col) c(col[col != ""]))
Area <- lapply(Area, function(col) as.numeric(col))
names(Area) <- NULL

Basin_name1 <- unique(unlist(Basin_name))
Basin_name2 <- sort(Basin_name1)

basin_areas <- unlist(lapply(seq_along(Basin_name), function(i) {
  setNames(Area[[i]], Basin_name[[i]])
}))
area_map <- basin_areas[Basin_name2]


##### Area target #####

Area_data <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 12)

matrix_area <- matrix(NA, nrow = n, ncol = n)

rownames(matrix_area) <- Basin_name2
colnames(matrix_area) <- Basin_name2

for (i in seq_along(Basin_name)) {
  basins <- Basin_name[[i]]
  qin_values <- Area_data[[i]]

  for (j in seq_along(basins)) {
    donor <- basins[j]
    qin_reg_donor <- qin_values[j]

    for (k in seq_along(basins)) {
      if (j != k) {
        target <- basins[k]
        matrix_area[donor, target] <- qin_reg_donor
      }
    }
  }
}

matrix_area <- matrix(as.numeric(matrix_area),
                      nrow = nrow(matrix_area),
           ncol = ncol(matrix_area),
           dimnames = list(rownames(matrix_area), colnames(matrix_area)))


# Initialize the D matrix
n <- length(Basin_name2)
matrix_D <- matrix(NA, nrow = n, ncol = n)
rownames(matrix_D) <- Basin_name2
colnames(matrix_D) <- Basin_name2

for (i in seq_along(Basin_name)) {
  basins <- Basin_name[[i]]
  for (j in seq_along(basins)) {
    donor <- basins[j]
    for (k in seq_along(basins)) {
      if (j != k) {
        target <- basins[k]

        A_donor <- area_map[donor]
        A_target <- area_map[target]
        A_max <- max(A_donor, A_target)
        A_min <- min(A_donor, A_target)

        matrix_D[target, donor] <- log(A_max / A_min)
      }
    }
  }
}

max_value <- max(matrix_D, na.rm = TRUE)

min_value <- min(matrix_D, na.rm = TRUE)

cat("Maximum value:", max_value, "\n")
cat("Minimum value:", min_value, "\n")

# Threshold value
thresholds <- seq(0.4, 6.4, by = 0.6)
threshold_vector <- vector("numeric", length(thresholds))
maximum_vector <- vector("numeric", length(thresholds))
pt_list <- list()
st_list <- list()
sigma_st_list <- list()
rt_list <- list()
sigma_pt_list <- list()
sigma_rt_list <- list()

###### Big LOOP starts

for (ii in seq_along(thresholds)) {
  threshold <- thresholds[ii]

  # Function to filter D_matrix based on a threshold
  filter_D_matrix <- function(D_matrix, threshold) {
    # Create a copy of the D_matrix to preserve the structure
    filtered_matrix <- D_matrix

    filtered_matrix[filtered_matrix >= threshold] <- NA

    return(filtered_matrix)
  }

  filtered_D_matrix <- filter_D_matrix(matrix_D, threshold)

  count_non_na <- function(matrix) {
    sum(!is.na(matrix))
  }

  non_na_count <- count_non_na(filtered_D_matrix)

  # Print the result
```

```
cat("The number of cells containing a number (excluding NA):", non_na_count, "\n")

##### DELTA #########

delta <- matrix_P_t - matrix_st

##### Filtered DELTA #########

filtered_delta <- matrix(NA, nrow = nrow(delta), ncol = ncol(delta),
                         dimnames = list(rownames(delta), colnames(delta)))

for (i in 1:nrow(delta)) {
  for (j in 1:ncol(delta)) {

    if (!is.na(filtered_D_matrix[i, j])) {
      filtered_delta[i, j] <- delta[i, j]
    }
  }
}

filtered_delta <- matrix(as.numeric(filtered_delta),
                         nrow = nrow(filtered_delta),
                         ncol = ncol(filtered_delta),
                         dimnames = list(rownames(filtered_delta), colnames(filtered_delta)))

##### Filtered P(t) #########

filtered_pt <- matrix(NA, nrow = nrow(matrix_P_t), ncol = ncol(matrix_P_t),
                      dimnames = list(rownames(matrix_P_t), colnames(matrix_P_t)))

for (i in 1:nrow(matrix_P_t)) {
  for (j in 1:ncol(matrix_P_t)) {
    if (!is.na(filtered_D_matrix[i, j])) {
      filtered_pt[i, j] <- matrix_P_t[i, j]
    }
  }
}

filtered_pt <- matrix(as.numeric(filtered_pt),
                      nrow = nrow(filtered_pt),
                      ncol = ncol(filtered_pt),
                      dimnames = list(rownames(filtered_pt), colnames(filtered_pt)))

pt_list[[ii]] <- filtered_pt

##### Filtered S(t) #########

filtered_st <- matrix(NA, nrow = nrow(matrix_st), ncol = ncol(matrix_st),
                      dimnames = list(rownames(matrix_st), colnames(matrix_st)))

for (i in 1:nrow(matrix_st)) {
  for (j in 1:ncol(matrix_st)) {
    if (!is.na(filtered_D_matrix[i, j])) {
      filtered_st[i, j] <- matrix_st[i, j]
    }
  }
}

filtered_st <- matrix(as.numeric(filtered_st),
                      nrow = nrow(filtered_st),
                      ncol = ncol(filtered_st),
                      dimnames = list(rownames(filtered_st), colnames(filtered_st)))

st_list[[ii]] <- filtered_st

##### Reginal donor #####

Qin_reg <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 10)

matrix_reg_d <- matrix(NA, nrow = n, ncol = n)
```

```r
rownames(matrix_reg_d) <- Basin_name2
colnames(matrix_reg_d) <- Basin_name2

for (i in seq_along(Basin_name)) {
  basins <- Basin_name[[i]]
  qin_values <- Qin_reg[[i]]

  for (j in seq_along(basins)) {
    donor <- basins[j]
    qin_reg_donor <- qin_values[j]

    for (k in seq_along(basins)) {
      if (j != k) {
        target <- basins[k]

        matrix_reg_d[target, donor] <- qin_reg_donor
      }
    }
  }
}

matrix_reg_d <- matrix(as.numeric(matrix_reg_d),
                       nrow = nrow(matrix_reg_d),
                       ncol = ncol(matrix_reg_d),
                       dimnames = list(rownames(matrix_reg_d), colnames(matrix_reg_d)))

##### Filtered regional donor #######

filtered_reg_d <- matrix(NA, nrow = nrow(matrix_reg_d), ncol = ncol(matrix_reg_d),
                         dimnames = list(rownames(matrix_reg_d), colnames(matrix_reg_d)))


for (i in 1:nrow(matrix_reg_d)) {
  for (j in 1:ncol(matrix_reg_d)) {
    # If the cell in filtered_D_matrix is not NA, copy the value from matrix_qin_reg
    if (!is.na(filtered_D_matrix[i, j])) {
      filtered_reg_d[i, j] <- matrix_reg_d[i, j]
    }
  }
}

filtered_reg_d <- matrix(as.numeric(filtered_reg_d),
                         nrow = nrow(filtered_reg_d),
                         ncol = ncol(filtered_reg_d),
                         dimnames = list(rownames(filtered_reg_d), colnames(filtered_reg_d)))


######## regional target ######

matrix_reg_t <- matrix(NA, nrow = n, ncol = n, dimnames = list(Basin_name2, Basin_name2))

for (i in seq_along(Basin_name)) {
  basins <- Basin_name[[i]]
  qin_values <- Qin_reg[[i]]

  for (j in seq_along(basins)) {
    donor <- basins[j]
    qin_reg_donor <- qin_values[j]

    for (k in seq_along(basins)) {
      if (j != k) {
        target <- basins[k]

        matrix_reg_t[donor, target] <- qin_reg_donor
      }
    }
  }
}

matrix_reg_t <- matrix(as.numeric(matrix_reg_t),
                       nrow = nrow(matrix_reg_t),
                         ncol = ncol(matrix_reg_t),
                         dimnames = list(rownames(matrix_reg_t), colnames(matrix_reg_t)))


  ##### Filtered regional target #######

  filtered_reg_t <- matrix(NA_real_, nrow = nrow(matrix_reg_t), ncol = ncol(matrix_reg_t),
                           dimnames = list(rownames(matrix_reg_t), colnames(matrix_reg_t)))

  for (i in 1:nrow(matrix_reg_t)) {
    for (j in 1:ncol(matrix_reg_t)) {
      if (!is.na(filtered_D_matrix[i, j])) {
        filtered_reg_t[i, j] <- matrix_reg_t[i, j]
      }
    }
  }

  rt_list[[ii]] <- filtered_reg_t

  ##### sigma S donor #####

  sigma_S_d <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 15)

  matrix_sigma_S_d <- matrix(NA, nrow = n, ncol = n)

  rownames(matrix_sigma_S_d) <- Basin_name2
  colnames(matrix_sigma_S_d) <- Basin_name2

  for (i in seq_along(Basin_name)) {
    basins <- Basin_name[[i]]
    sigma_qin_values <- sigma_S_d[[i]]

    for (j in seq_along(basins)) {
      donor <- basins[j]
      sigma_qin_donor <- sigma_qin_values[j]

      for (k in seq_along(basins)) {
        if (j != k) {
          target <- basins[k]

          matrix_sigma_S_d[target, donor] <- sigma_qin_donor
        }
      }
    }
  }

  matrix_sigma_S_d <- matrix(as.numeric(matrix_sigma_S_d),
                             nrow = nrow(matrix_sigma_S_d),
                             ncol = ncol(matrix_sigma_S_d),
                             dimnames = list(rownames(matrix_sigma_S_d),
colnames(matrix_sigma_S_d)))

  ##### Filtered sigma S donor #######

  filtered_sigma_s_d <- matrix(NA, nrow = nrow(matrix_sigma_S_d), ncol = ncol(matrix_sigma_S_d),
                               dimnames = list(rownames(matrix_sigma_S_d),
colnames(matrix_sigma_S_d)))


  for (i in 1:nrow(matrix_sigma_S_d)) {
    for (j in 1:ncol(matrix_sigma_S_d)) {
      if (!is.na(filtered_D_matrix[i, j])) {
        filtered_sigma_s_d[i, j] <- matrix_sigma_S_d[i, j]
      }
    }
  }

  filtered_sigma_s_d <- matrix(as.numeric(filtered_sigma_s_d),
                               nrow = nrow(filtered_sigma_s_d),
                               ncol = ncol(filtered_sigma_s_d),
                               dimnames = list(rownames(filtered_sigma_s_d), 
colnames(filtered_sigma_s_d)))

  ##### sigma S target #####

  sigma_s_t <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 15)

  matrix_sigma_s_t <- matrix(NA, nrow = n, ncol = n)

  rownames(matrix_sigma_s_t) <- Basin_name2
  colnames(matrix_sigma_s_t) <- Basin_name2

  for (i in seq_along(Basin_name)) {
    basins <- Basin_name[[i]]
    sigma_qin_values <- sigma_s_t[[i]]

    for (j in seq_along(basins)) {
      donor <- basins[j]
      sigma_qin_donor <- sigma_qin_values[j]

      for (k in seq_along(basins)) {
        if (j != k) {
          target <- basins[k]

          matrix_sigma_s_t[donor, target] <- sigma_qin_donor
        }
      }
    }
  }

  matrix_sigma_s_t <- matrix(as.numeric(matrix_sigma_s_t),
                             nrow = nrow(matrix_sigma_s_t),
                             ncol = ncol(matrix_sigma_s_t),
                             dimnames = list(rownames(matrix_sigma_s_t), 
colnames(matrix_sigma_s_t)))

  ##### Filtered sigma S target #######

  filtered_sigma_s_t <- matrix(NA, nrow = nrow(matrix_sigma_s_t), ncol = ncol(matrix_sigma_s_t),
                               dimnames = list(rownames(matrix_sigma_s_t), 
colnames(matrix_sigma_s_t)))

  for (i in 1:nrow(matrix_sigma_s_t)) {
    for (j in 1:ncol(matrix_sigma_s_t)) {
      if (!is.na(filtered_D_matrix[i, j])) {
        filtered_sigma_s_t[i, j] <- matrix_sigma_s_t[i, j]
      }
    }
  }


  filtered_sigma_s_t <- matrix(as.numeric(filtered_sigma_s_t),
                               nrow = nrow(filtered_sigma_s_t),
                               ncol = ncol(filtered_sigma_s_t),
                               dimnames = list(rownames(filtered_sigma_s_t), 
colnames(filtered_sigma_s_t)))

  sigma_st_list[[ii]] <- filtered_sigma_s_t


  ##### Alpha ########

  valid_indices <- !is.na(filtered_delta) & !is.na(filtered_D_matrix) & !is.na(filtered_reg_t) & 
!is.na(filtered_reg_d)& !is.na(filtered_sigma_s_t)& !is.na(filtered_sigma_s_d)

  # Filter out the NA values using the valid indices
  deltas <- filtered_delta[valid_indices]
  D <- filtered_D_matrix[valid_indices]
  Rt <- filtered_reg_t[valid_indices]
  Rd <- filtered_reg_d[valid_indices]
  sigma_St <- filtered_sigma_s_t[valid_indices]
  sigma_sd <- filtered_sigma_s_d[valid_indices]

  loglik <- function(alpha, deltas, D, Rt, Rd, sigma_sd, sigma_St) {

    variance <- ((1 + alpha * D)^2 * (sigma_sd)^2 * (Rt / Rd)^2) + sigma_St^2

    log_likelihood <- sum(dnorm(deltas, mean = 0, sd = sqrt(variance), log = TRUE))

    return(log_likelihood)
  }

  #optimization to find the best alpha
  result <- optimize(loglik, interval = c(0, 150), deltas = deltas, D = D,
                     Rt = Rt, Rd = Rd, sigma_sd = sigma_sd, sigma_St = sigma_St, maximum = TRUE)

  cat("The optimal alpha is:", result$maximum, "\n")
  cat("The maximum log-likelihood is:", result$objective, "\n")

  threshold_vector[ii] <- threshold
  maximum_vector[ii] <- result$maximum

  ####### plot alpha #########

  alphas <- seq(0, 150, by = 0.05)
  log_lik_values <- sapply(alphas, function(alpha) {
    loglik(alpha, deltas, D, Rt, Rd, sigma_sd = sigma_sd, sigma_St = sigma_St)
  })

  plot(alphas, log_lik_values, type = "l", col = "blue", lwd = 2,
       main = "Log-likelihood vs Alpha",
       xlab = "Alpha", ylab = "Log-likelihood")
  abline(v = result$maximum, col = "red", lty = 2)  # Add a line for the optimal alpha

  points(result$maximum, max(log_lik_values), col = "red", pch = 19)

  text(result$maximum, max(log_lik_values),
       labels = paste0("Alpha = ", round(result$maximum, 4)),
       pos = 4, col = "red")

  legend("topright",
         legend = paste("D(lim) =", threshold_vector[ii]),
         bty = "n", cex = 1, text.col = "black",
         inset = c(0, 0.1))  # Moves the legend downward


  ####### sigma_pt #########

  optimal_alpha <- maximum_vector[ii]

  sigma_pt <- (1 + optimal_alpha * filtered_D_matrix) * filtered_sigma_s_d * (filtered_reg_t /
filtered_reg_d)

  sigma_pt_list[[ii]] <- sigma_pt

  sigma_pt_array <- simplify2array(sigma_pt_list)


  #####  sigma R target #######

  sigma_R_t <- read.xlsx('local_vs_regional_Lmoments226sites20231103.xlsx', 17)

  matrix_sigma_R_t <- matrix(NA, nrow = n, ncol = n)

  rownames(matrix_sigma_R_t) <- Basin_name2
  colnames(matrix_sigma_R_t) <- Basin_name2

  for (i in seq_along(Basin_name)) {
    basins <- Basin_name[[i]]
    sigma_qin_values <- sigma_R_t[[i]]

    for (j in seq_along(basins)) {
      donor <- basins[j]
      sigma_qin_donor <- sigma_qin_values[j]

      for (k in seq_along(basins)) {
        if (j != k) {
          target <- basins[k]

          matrix_sigma_R_t[donor, target] <- sigma_qin_donor
        }
      }
    }
  }

  matrix_sigma_R_t <- matrix(as.numeric(matrix_sigma_R_t),
                             nrow = nrow(matrix_sigma_R_t),
                             ncol = ncol(matrix_sigma_R_t),
                             dimnames = list(rownames(matrix_sigma_R_t),
colnames(matrix_sigma_R_t)))

  ##### Filtered sigma R target #######

  filtered_sigma_R_t <- matrix(NA, nrow = nrow(matrix_sigma_R_t), ncol = ncol(matrix_sigma_R_t),
                               dimnames = list(rownames(matrix_sigma_R_t),
colnames(matrix_sigma_R_t)))

  for (i in 1:nrow(matrix_sigma_R_t)) {
    for (j in 1:ncol(matrix_sigma_R_t)) {

      if (!is.na(filtered_D_matrix[i, j])) {
        filtered_sigma_R_t[i, j] <- matrix_sigma_R_t[i, j]
      }
    }
  }


  filtered_sigma_R_t <- matrix(as.numeric(filtered_sigma_R_t),
                               nrow = nrow(filtered_sigma_R_t),
                               ncol = ncol(filtered_sigma_R_t),
                               dimnames = list(rownames(filtered_sigma_R_t),
colnames(filtered_sigma_R_t)))
  sigma_rt_list[[ii]] <- filtered_sigma_R_t

}

###### Big LOOP ends

# Print the results
cat("Threshold values: ", threshold_vector, "\n")
cat("Maximum values for each threshold: ", maximum_vector, "\n")


###############

# Plot the line
plot(threshold_vector, maximum_vector, type = "l", col = "blue", lwd = 2,
     main = "D_lim and Alpha",
     xlab = "Distance Limit (log (A_max/A_min))", ylab = "Alpha")

points(threshold_vector, maximum_vector, col = "red", pch = 16)


###### percentage of basin only by threshold ####

for (j in 1:length(threshold_vector)) {
  threshold <- thresholds[j]

  # Function to filter D_matrix based on a threshold
  filter_D_matrix <- function(D_matrix, threshold) {
    # Create a copy of the D_matrix to preserve the structure
    filtered_matrix <- D_matrix
```

```r
    filtered_matrix[filtered_matrix >= threshold] <- NA

    return(filtered_matrix)
  }

  filtered_D_matrix <- filter_D_matrix(matrix_D, threshold)

  valid_cells <- !is.na(matrix_D)

  non_na_filtered_D <- sum(!is.na(filtered_D_matrix[valid_cells]))

  total_valid_cells <- sum(valid_cells)

  percentage_non_na <- (non_na_filtered_D / total_valid_cells) * 100

  # Print the result
  cat("Percentage of non-NA cells in 'filtered_D_matrix' corresponding to 'matrix_D':",
percentage_non_na, "%\n")
}

####### percentage of sigma_pt <= sigma_R_t #######

for (k in 1:length(threshold_vector)) {

  comparison <- sigma_pt_list[[k]] <= sigma_rt_list[[k]]

  valid_comparison <- !is.na(comparison)

  percentage_correct <- sum(comparison[valid_comparison]) / sum(valid_comparison) * 100

  # Print the result
  cat("Percentage of cells where sigma_pt <= matrix_sigma_R_t:", percentage_correct, "%\n")
}


######### Error Index for ASE #####

error_list_ASE <- list()
mean_error_list_ASE <- list()

for (l in 1:length(thresholds)) {
  error_ASE <- (pt_list[[l]]-st_list[[l]])/sigma_st_list[[l]]
  error_list_ASE[[l]] <- error_ASE

  mean_error_ASE <- mean(error_list_ASE[[l]], na.rm = TRUE)
  mean_error_list_ASE[[l]] <- mean_error_ASE
}


######### Error Index for Regional #####

error_list_Reg <- list()
mean_error_list_Reg <- list()

for (l in 1:length(thresholds)) {
  error_Reg <- (rt_list[[l]]-st_list[[l]])/sigma_st_list[[l]]
  error_list_Reg[[l]] <- error_Reg

  mean_error_Reg <- mean(error_list_Reg[[l]], na.rm = TRUE)
  mean_error_list_Reg[[l]] <- mean_error_Reg
}

#######

print(sapply(mean_error_list_ASE, function(x) x))
print(sapply(mean_error_list_Reg, function(x) x))


###### Error plot

for (u in seq_along(error_list_ASE)) {
  matrix_ASE <- error_list_ASE[[u]]
  matrix_Reg <- error_list_Reg[[u]]

  vec_ASE <- as.vector(matrix_ASE)
  vec_Reg <- as.vector(matrix_Reg)

  valid_indices <- !is.na(vec_ASE) & !is.na(vec_Reg)

  filtered_ASE <- vec_ASE[valid_indices]
  filtered_Reg <- vec_Reg[valid_indices]

  if (length(filtered_ASE) > 0 && length(filtered_Reg) > 0) {

    if (!interactive()) dev.new()

    # Create scatter plot
    plot(abs(filtered_ASE), abs(filtered_Reg),
         log = "xy",  # Log scale on both axes
         xlab = paste("AS error"),
         ylab = paste("Reg error"),
         pch = 16, col = "blue")
    legend("topleft",
           legend = paste("D(lim) =", threshold_vector[u]),
           bty = "n", cex = 1, text.col = "black",
           inset = c(0, 0))  # Moves the legend downward

    abline(a = 0, b = 1, col = "red", lty = 2)  # Red dashed line
  }
}


##### Extract values for OASE

extracted_pt_list <- list()
extracted_st_list <- list()
extracted_sigma_st_list <- list()

for (k in seq_along(sigma_pt_list)) {

  if (!all(dim(sigma_rt_list[[k]]) == dim(sigma_pt_list[[k]]),
           dim(sigma_pt_list[[k]]) == dim(pt_list[[k]]),
           dim(sigma_pt_list[[k]]) == dim(st_list[[k]]),
           dim(sigma_pt_list[[k]]) == dim(sigma_st_list[[k]]))) {
    stop("Matrices are not of the same size.")
  }

  row_names <- rownames(sigma_pt_list[[k]])
  col_names <- colnames(sigma_pt_list[[k]])

  comparison <- sigma_pt_list[[k]] <= sigma_rt_list[[k]]

  valid_indices <- !is.na(comparison) & comparison

  extracted_pt <- matrix(NA, nrow = nrow(pt_list[[k]]), ncol = ncol(pt_list[[k]]))
  extracted_st <- matrix(NA, nrow = nrow(st_list[[k]]), ncol = ncol(st_list[[k]]))
  extracted_sigma_st <- matrix(NA, nrow = nrow(sigma_st_list[[k]]), ncol =
ncol(sigma_st_list[[k]]))

  extracted_pt[valid_indices] <- pt_list[[k]][valid_indices]
  extracted_st[valid_indices] <- st_list[[k]][valid_indices]
  extracted_sigma_st[valid_indices] <- sigma_st_list[[k]][valid_indices]

  rownames(extracted_pt) <- row_names
  colnames(extracted_pt) <- col_names
  rownames(extracted_st) <- row_names
  colnames(extracted_st) <- col_names
  rownames(extracted_sigma_st) <- row_names
  colnames(extracted_sigma_st) <- col_names

  extracted_pt_list[[k]] <- extracted_pt
  extracted_st_list[[k]] <- extracted_st
  extracted_sigma_st_list[[k]] <- extracted_sigma_st
}


######### For operational, Adding values for which the sigma pt >= sigma rt

for (k in seq_along(extracted_pt_list)) {

  if (!is.null(pt_list[[k]]) && all(dim(pt_list[[k]]) == dim(extracted_pt_list[[k]]))) {

    na_indices <- is.na(extracted_pt_list[[k]])
    extracted_pt_list[[k]][na_indices] <- rt_list[[k]][na_indices]
  }

  if (!is.null(st_list[[k]]) && all(dim(st_list[[k]]) == dim(extracted_st_list[[k]]))) {

    na_indices <- is.na(extracted_st_list[[k]])
    extracted_st_list[[k]][na_indices] <- st_list[[k]][na_indices]
  }

  if (!is.null(sigma_st_list[[k]]) && all(dim(sigma_st_list[[k]]) ==
dim(extracted_sigma_st_list[[k]]))) {

    na_indices <- is.na(extracted_sigma_st_list[[k]])
    extracted_sigma_st_list[[k]][na_indices] <- sigma_st_list[[k]][na_indices]
  }
}

#########

for (k in seq_along(sigma_pt_list)) {
  num_non_na <- sum(!is.na(extracted_pt_list[[k]]))

  cat("Number of non-NA cells in the matrix:", num_non_na, "\n")
}


######### Error Index for OASE #####

error_list_OASE <- list()
mean_error_list_OASE <- list()

for (l in 1:length(thresholds)) {
  error_OASE <- (extracted_pt_list[[l]]-extracted_st_list[[l]])/extracted_sigma_st_list[[l]]
  error_list_OASE[[l]] <- error_OASE

  mean_error_OASE <- mean(error_list_OASE[[l]], na.rm = TRUE)
  mean_error_list_OASE[[l]] <- mean_error_OASE
}

print(sapply(mean_error_list_OASE, function(x) x))


####### RMSE

# Function to compute RMSE for two matrices
compute_rmse <- function(mat1, mat2) {
  if (!all(dim(mat1) == dim(mat2))) {
    return(NA)
  }

  valid_indices <- !is.na(mat1) & !is.na(mat2)

  values_mat1 <- mat1[valid_indices]
  values_mat2 <- mat2[valid_indices]

  # Compute RMSE
  sqrt(mean((values_mat1 - values_mat2)^2))
}
```

```r
rmse_values_prop <- mapply(compute_rmse, pt_list, st_list)

print(rmse_values_prop)

rmse_values_op <- mapply(compute_rmse, extracted_pt_list, extracted_st_list)

print(rmse_values_op)

rmse_values_reg <- mapply(compute_rmse, rt_list, st_list)

print(rmse_values_reg)
mean_rmse_reg <- mean(rmse_values_reg)

# Function to compute RMSE for each column of two matrices
rmse_per_column <- function(mat1, mat2) {
  sapply(1:ncol(mat1), function(j) {
    valid_idx <- !is.na(mat1[, j]) & !is.na(mat2[, j])

    if (sum(valid_idx) == 0) {
      return(NA)
    }

    sqrt(mean((mat1[valid_idx, j] - mat2[valid_idx, j])^2))
  })
}

# Function to compute the average RMSE for each matrix pair
average_rmse_per_matrix <- function(pt_list, st_list) {
  mapply(function(mat1, mat2) {
    rmse_values <- rmse_per_column(mat1, mat2)

    mean(rmse_values, na.rm = TRUE)
  }, pt_list, st_list)
}

average_rmse_results <- average_rmse_per_matrix(extracted_pt_list, extracted_st_list)

print(average_rmse_results)

x_values <- thresholds

if (length(rmse_values_prop) != length(x_values) | length(rmse_values_op) != length(x_values)) {
  stop("The number of RMSE values does not match the x-axis length")
}

plot(x_values, rmse_values_prop, type = "o", col = "blue", pch = 16, lwd = 2,
     xlab = "D (lim)", ylab = "RMSE",
     ylim = range(c(rmse_values_prop, average_rmse_results), na.rm = TRUE))  # Adjust y-axis
range

lines(x_values, average_rmse_results, type = "o", col = "red", pch = 16, lwd = 2)

legend("topleft", legend = c("Propagated", "Operational", "Regional_mean"),
       col = c("blue", "red", "green"), pch = 16, lwd = 2,
       cex = 0.8,  # Reduce font size
       bty = "n")  # Remove the box around the legend)

abline(h = 210.93, col = "green", lwd = 2, lty = 2)  # Green dashed line
```