

POLITECNICO DI TORINO



Ph.D. in Water and Land Management Engineering

**Runoff estimation in data-scarce
and ungauged basins with
systematic use of
morpho-climatic information**

Daniele Ganora

matr. 143732

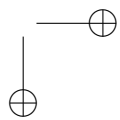
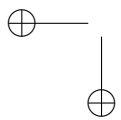
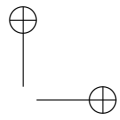
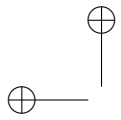
Advisors:

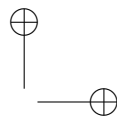
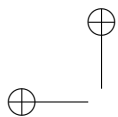
Prof. Pierluigi Claps

Dr. Francesco Laio

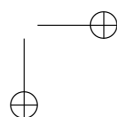
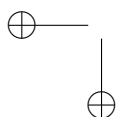
Ph.D. Dissertation

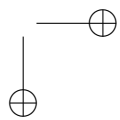
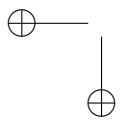
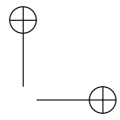
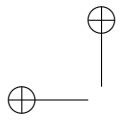
March 2010

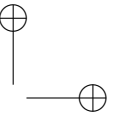
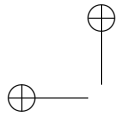




To my family





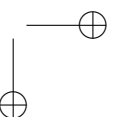
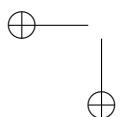


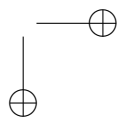
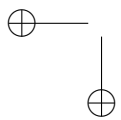
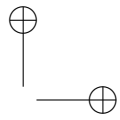
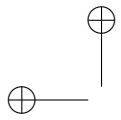
Abstract

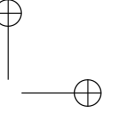
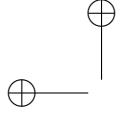
The knowledge of the hydrological behavior of a river is a valuable information useful for many purposes: from the regulation of water resources use to the management of extreme events and for many environmental issues. The primary sources of this kind of information are the streamflow measurements, that are, however, available only at a limited number of gauging stations. Despite this, nowadays, it is necessary to quantitatively extend the hydrological information to many other locations, or even to the entire river network.

This work analyzes, improves and defines statistical techniques that allow one to use the information available only at few locations, to obtain suitable estimates of hydrological variables in ungauged sites. To this aim, some limitations of previously available models are overcome and new analysis tools are developed. The main topics discussed in this work are: the information retrieval from poorly-gauged sites, the analysis of uncertainty of the regional estimates, the improvement of results by means of proximity information and the use of non-conventional data that cannot be represented by simple observations.

The proposed methodologies are applied to the flood frequency curve and the flow duration curve, in order to assess respectively the extreme and the average catchment behavior. The models are tested on the basis of different sets of measurements available for Northwestern Italy and Switzerland, demonstrating the applicability and reliability of the proposed methods.

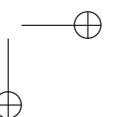
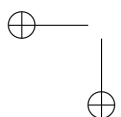






I would like to thank my advisors Prof. Pierluigi Claps and Dr. Francesco Laio for their encouragement, guidance and support.
I wish to extend my warmest thanks to all friends and colleagues who have helped me during this work.

Thesis typeset with L^AT_EX 2_ε



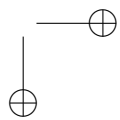
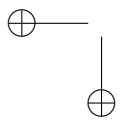
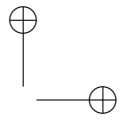
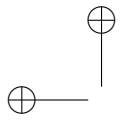
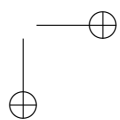
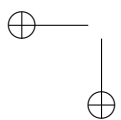
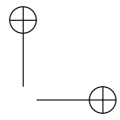
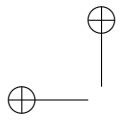


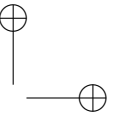
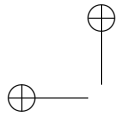
Table of Contents

1. Scope of the work: statistical hydrology in ungauged basins	1
2. Regional approach for flood quantile estimation in ungauged and data-scarce basins	7
2.1. Introduction	7
2.2. Model definition	11
2.2.1. At-site estimates: systematic and non-systematic in-formation	11
2.2.2. Regression models	15
2.2.3. Model selection	18
2.3. Selection of the probability distribution	20
2.4. Case study	23
2.4.1. Data availability	23
2.4.2. Model definition	25
2.4.3. Regression results	29
2.4.4. Quantile estimation	39
2.4.5. L-moments estimates in data-scarce stations	41
2.5. Final remarks	42
3. Along-stream estimation approach	47
3.1. Introduction	47
3.2. Extension of the regional procedure	50
3.2.1. Regional covariance and correlation	50
3.2.2. Formulae for log-transformed data	52

3.3. Along-Stream information propagation method	53
3.3.1. Methods and hypotheses	53
3.3.2. Example about functions and assumptions	58
3.3.3. Organization of nested basins	59
3.4. Model reliability: simplified approach	61
3.4.1. Uncertainty of the propagated estimate	61
3.4.2. Assessment of the variance parameter	63
3.4.3. Validity of the simplified approach	65
3.5. Model reliability: analytical approach	71
3.6. Final remarks	73
4. Distance-based regional approach for flow duration curves	75
4.1. Introduction	75
4.2. Distance-based method	77
4.2.1. (Dis)similarity between curves	78
4.2.2. Distance matrices, linear regression and Mantel test	81
4.2.3. Cluster analysis	84
4.3. Case study: distance-based method application	88
4.3.1. Hydrological and geomorphologic data	88
4.3.2. Procedure setting	90
4.3.3. Regions definition	92
4.4. Comparison with parametric models	93
4.5. Final remarks	99
5. Summary and conclusions	101
References	103
A. Hydrological data summary	113
A.1. Hydrological data for flood flows modelling	113
A.2. Hydrological data for flow duration curves modelling	116

B. Morpho-climatic information	119
B.1. Parameters description	119
B.1.1. Geomorphological parameters	120
B.1.2. River network parameters	121
B.1.3. Soil use and permeability parameters	123
B.1.4. Climatic parameters	124
B.2. Parameters list	124



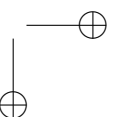
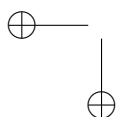


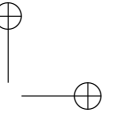
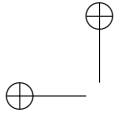
Chapter 1.

Scope of the work: statistical hydrology in ungauged basins

Hydrology deals with the water cycle, and the water cycle profoundly affects all the aspects of the life on Earth. The necessity of a better knowledge and understanding of the hydrological mechanisms is not only a scientific pursuit, but also a need for a correct management of water resources.

Undoubtedly, the water cycle is hard to understand, primarily because it involves a great variety of physical processes and also because these processes take place on a great range of scales, both from the spatial and the temporal viewpoint. Far from a complete understanding and coupling of the elementary components of the water cycle, hydrological issues are usually referred to different macro-areas of expertise. Under this broad context, this thesis can be considered a contribution to the field of “catchment hydrology” [Uhlenbrook, 2006] that deals with all the components of the terrestrial water cycles that interact over the basin domain. Basins, in fact, can be considered as fundamental landscape units [Sivapalan et al., 2003] that integrate the hydrological cycle with geochemical, ecologic, morphological and other processes. All these processes are strictly related to the fluxes through the basin boundaries and, in particular, from and toward the atmosphere and groundwater. Despite this, one of the major elements of interest lies in the water flux through the basin outlet, i.e. the streamflow or runoff, due to the importance that this variable has in many practical applications. Streamflow





2 Scope of the work: statistical hydrology in ungauged basins

is often also an easily interpretable index that summarizes, in some way, all the catchment processes.

A basin is a complex system whose behavior can be studied through macro-characteristics that are, for instance, the magnitude, frequency and duration of a particular type of event. These macro-characteristics can be handled by means of statistical procedures, that try to interpret hydrological patterns without resorting to the description of the physical processes. This is the case of this thesis in which extreme and mean runoff are investigated respectively through the flood frequency curve and the flow duration curve. The former allows one to define the probability to have a flood equal or greater than a predefined threshold, while the second summarize the time during which a certain discharge is available.

The hydrological behavior of a basin has all along held a fundamental role in the organization of human societies. An early well known example is the strong relationship between the Egyptian civilization and the river Nile that is documented, for instance, on the North side of the Karnak temple complex [Lauro, 2009, see also figure 1.1] where hieroglyphics represent propitiatory values of flood level in different location along the river [Lacau and Chevrier, 1956]. This relationship has become much more important in the last decades: the increasing exploitation of water resources, as well as changes in land use and urbanization, highlighted the need for a quantitative and reliable characterization of the surface water flows, in order to correctly manage this resource. Problems arise because of different types of users with conflicting necessities (e.g. agriculture, industry, energy plants, etc) that compete for water exploitation. Moreover, the increased attention to environmental problems related to water quantity and quality claims for practical management tools. In addition to the problem of protection of water, it is also important to protect communities and goods from water extreme events, in particular extreme floods and droughts. As a consequence, nowadays we need quantitative and extended information about several hy-

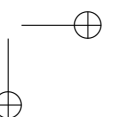
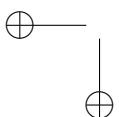




Figure 1.1.: Particular of the North side of the Karnak temple complex (*Courtesy of Mario Lauro - www.cartigli.it*)

drological variables to cope with these pressing necessities.

The more straightforward way to analyze the catchment behavior is to study its streamflow time series and, eventually, other related variables, like precipitation, soil characteristics, vegetation, etc. However, this approach requires the discharge time series to be known at the site of interest. When a direct river flow monitoring is not available or collected data are not adequate for the analysis, the basin is referred to as ungauged, and indirect methods are required to study its hydrological characteristics. An example of application of an indirect procedure is reported in figure 1.2. The map shows the Dora Baltea river basin at Tavagnasco, that is one of the largest basins analyzed in this thesis, with some gauging stations present on the catchment. The map is an example of how the hydrological-information available only at some locations should be extended to the whole drainage network.

Indirect procedures are based on the concept of information transfer from gauged to ungauged basins, that has been summarized by the principle “substitute time for space” proposed by the US National Research Council [1988] for hydro-meteorological modelling. This principle underlines the idea of compensating the lack of time serie records by using the data relative to



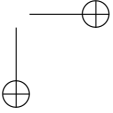
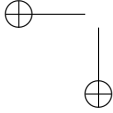
4 Scope of the work: statistical hydrology in ungauged basins

other sites. This topic is particularly important in the area of catchment hydrology, and a demonstration to this is the decennial Prediction in Ungauged Basins (PUB) initiative [Sivapalan et al., 2003].

This thesis focuses on several points of interest for PUB, among which:

- the improvement and generalization of current procedures, with particular attention to data-scarce stations, that represent cases that are usually discarded by classic approaches;
- the analysis of uncertainty of the predicted values;
- the correction of estimates yield by large-scale models by means of local information;
- the use of non-conventional data and procedures to handle hydrological information.

In particular, in chapter 2, the identification of a suitable flood frequency curve in ungauged basins is analyzed. The study aims at implementing a regional procedure that overcomes some of the limitations of the classic approaches and adds a clearer quantitative description of the uncertainty components of the estimation. To do so, the at-site data are not used to estimate local parameters of a statistical distribution model, but the information in the sample record is summarized by a set of robust sample statistics (the L -moments), that become the variables to be regionalized. This leads to a generalization of the widely used *index-flood* approach. The proposed approach allows one to eliminate the uncertainty related to the choice of the distribution function, in particular when short samples are involved. As a consequence short samples, that are usually discarded, can still be used to contribute to the improvement of the consistency of the database. To transfer the information to the ungauged basins, we adopt a separate regional model for each of the L -moments considered, based on a comprehensive multiple regression approach, selecting the independent variables among many geomorphoclimatic catchment descriptors. Each regression model is calibrated



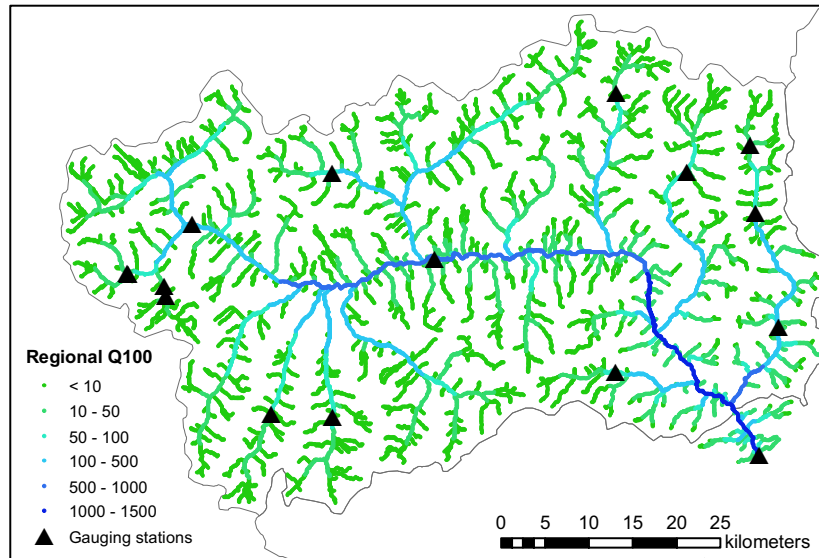


Figure 1.2.: Map of the predicted flood flow (in m^3/s) with return period of 100 years mapped over the river network upstream Tavagnasco on the river Dora Baltea.

using non-standard least-squares techniques over the whole dataset, without resorting to any grouping procedure to create sub-regions. The flood frequency features related to each catchment are thus allowed to vary smoothly from site to site, following the variation of the descriptors selected in the regression models.

Regional models, however, do not preserve the information related to the natural hierarchy between gauged stations deriving from their location along the river network. This information is particularly important when one wants to estimate runoff at a site located immediately upstream or downstream a gauged station. In this case, a possible alternative is to estimate the variable directly, on the basis of the corresponding statistics calculated at the gauged station. The closer the estimation point is to the gauged station, the greater is the expected quality of this estimation procedure. This idea is discussed in chapter 3 where this procedure is defined as Along-Stream estimation

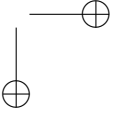
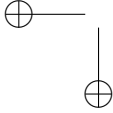


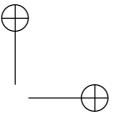
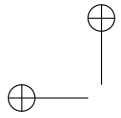
6 Scope of the work: statistical hydrology in ungauged basins

method, to underline that it is applied to points along a stream network. It requires to identify a suitable formula to compute the variable at the ungauged site. This formula can be based on a set of basin characteristics, or, in alternative, on a regional estimate (local estimation coupled with a regional model). Then, a criterion to assess the reliability of the stream model and its domain of application is defined and, finally, the accuracy of the approach is evaluated through the assessment of the standard deviation of its estimates. In this way it is possible to compare the variance of the stream estimates against the variance of other models, if any, and thus to choose the most accurate method (or to combine different estimates).

For cases in which the hydrological information to reconstruct is fairly complex, such as for flow duration curves (FDC), a new regional model is developed, as described in chapter 4. Although the method is still referred to as “regional”, the basic approach is very different from that of chapter 2 because it aims at representing the FDC as a non-parametric object, rather than providing a parametric representation and trying to relate the parameter values to basin descriptors. This approach considers the (dis)similarity between all possible pairs of curves, and uses distance measures to quantify the dissimilarity. The regional model uses this concept of dissimilarity to recognize similar catchments and thus to perform the prediction in ungauged basins. The main characteristic of this approach is that it allows one to describe the variable of interest without recurring to an analytic function. This can be very useful when variables of interest are particularly “complex” objects, for instance, images or spectra [e.g. Pekalska and Duin, 2005] or when dealing with ecological data [e.g. Legendre and Legendre, 1998]. This is also useful in those cases for which classic parameterizations are not satisfactory.

All the approaches developed in the thesis are tested on a group of catchments located in the Northwestern Italy and in Switzerland.





Chapter 2.

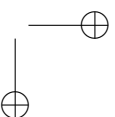
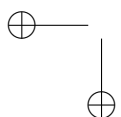
Regional approach for flood quantile estimation in ungauged and data-scarce basins

Contents

2.1. Introduction	7
2.2. Model definition	11
2.2.1. At-site estimates: systematic and non-systematic information	11
2.2.2. Regression models	15
2.2.3. Model selection	18
2.3. Selection of the probability distribution	20
2.4. Case study	23
2.4.1. Data availability	23
2.4.2. Model definition	25
2.4.3. Regression results	29
2.4.4. Quantile estimation	39
2.4.5. L-moments estimates in data-scarce stations	41
2.5. Final remarks	42

2.1. Introduction

The evaluation of the frequency of flood events in ungauged catchments is usually approached by building suitable statistical relationships between flood statistics and basins characteristics, based on a set of gauged stations. These models are used to transfer the information available from the gauged



sites to the target basin, where only morphoclimatic catchment's characteristics need to be known. This type of procedure is referred to as *regional model*, because it identifies a (homogeneous) subset of basins, called *region*, that is used later as the pooling set for the estimation at the ungauged site. In this case, the basins belonging to the region are supposed to donate their statistical properties to the ungauged ones falling in the same region.

Different methods to achieve this goal have been proposed in the literature (see for example the review by Cunnane [1988]), differing to each other mainly on the basis of the distribution used to describe the at-site data [see e.g. Hosking and Wallis, 1997, for a bouquet of distributions], and on the pooling criterion used for the delineation of regions. Several classification techniques for regions creation have been adopted, from cluster analysis to proximity pooling [Burn, 1990], hierarchical classification [Gabriele and Arnell, 1991], neural network classifiers [Hall and Minns, 1999] and mixed approaches [Merz and Blöschl, 2005], among others. The homogeneity of such regions is also an important issue [Viglione et al., 2007, Castellarin et al., 2008].

Classic statistical tools for the assessment of the flood frequency curve in ungauged basins usually presents methodologies that limit the model to further generalizations. In particular, (i) the subdivision of the domain of interest in homogeneous regions, and (ii) the choice of an a priori probability distribution to describe the sample data can be overcome through the model proposed in this chapter. The method will still be referred to as *regional*, although it relaxes the need for the definition of regions in the classical way; rather, *regional* is intended as defining the hydrological information transfer over mid-large scale areas.

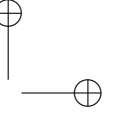
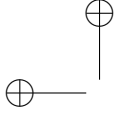
Regarding the first point, the usual approach is to create bounded regions that can have different characteristics depending on the method adopted. For instance, regions can be created splitting in separated areas the geographical space or the space of the physiographic basin's characteristics. Then, the

regions can be defined by means of fixed boundaries (e.g. cluster analysis procedures) or with a changing border as in the region-of-influence (ROI) approach. A few exceptions are the interpolation of the hydrological variable in the descriptors space [Chokmani and Ouarda, 2004], the top-kriging [Skøien et al., 2006] and the approach of Chebana and Ouarda [2008], who use a weighting procedure that accounts for all the available data. No separated regions are considered, but all the gauged basins contribute to the regional estimation. This is possible thanks to a smooth surface, over the descriptors space, that define the weight to assign to each gauged site.

The main limitation of the approaches that use a subdivision in separate regions is the assessment of the uncertainty relative to the configuration of the regions themselves (e.g. which catchments to include or not in a particular region). In fact, since there is not any prior information about the regions configuration, any algorithm used for regions delineation induces some errors. Then, the regions are tested for their statistical homogeneity, although, sometimes, slight heterogeneous regions are accepted. Also in these cases the uncertainty component is difficult to assess.

In this work, we consider the dataset as a whole, without resorting to a grouping procedure to form the regions. This idea has been already developed by Stedinger and Tasker [1985] and recently improved by Griffis and Stedinger [2007] where the benefit of using this approach is underlined. Again, with a unique region there is no longer the need for an homogeneity test: statistical characteristics of the hydrological behavior can vary from site to site and it is the model itself that tries to catch this variability.

Regarding the second point we mark another difference from former, as well as more recent, works [Griffis and Stedinger, 2007, Ouarda et al., 2008, among others]. In particular, our approach does not require, as an initial stage, any hypothesis on the at-site frequency distribution that is used to describe the data records and to estimate flood quantiles. In fact, regional estimation of flood quantiles depends on the distribution used to fit the

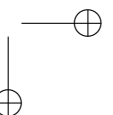
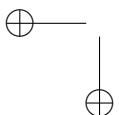


Regional approach for flood quantile estimation in ungauged and data-scarce basins

original sample, especially when dealing with short records, because many different distributions can fit equally well the data for low return periods, while they may behave in a different way when extrapolated to high return periods.

Nevertheless, to transfer information from the gauged to the ungauged sites, the time series data need to be summarized in some way. To this purpose we use the L -moments and their dimensionless ratios as regional variables; in particular we select the L -moment of order one (the mean) and the coefficient of L -variation (L_{CV}) and the L -skewness (L_{CA}) of the record. After the regionalization of L -moments it is possible to reconstruct the whole flood frequency curve. The choice of the mean, L_{CV} and L_{CA} as hydrological signatures in a regional framework can be interpreted in an index-flood framework [Dalrymple, 1960] considering the mean as the scale factor and the L -moments ratios as the descriptors of the dimensionless growth curve.

The use of mean, L_{CV} and L_{CA} instead of a quantile or the distribution parameters is particularly helpful, for both calibration and prediction purposes, when catchments with short sample records are present in the analysis. During the model calibration phase, in a traditional approach, a short sample that can not be used to fit a distribution is discarded. In this model, instead, the L -moments can be computed and used in the regional procedure, even if the accuracy of their estimators is low (but known), avoiding the information loss due to the data elimination. On the other hand, if one is interested in the prediction in a poorly gauged site, it is possible to compute, for instance, Q_{ind} and L_{CV} directly on the sample record, leaving to the regional procedure the assessment of L_{CA} . From this point of view, this approach extends the original index-flood method, in which Q_{ind} is often estimated on the basis of few at-site measurements, while the growth curve is treated by a regional model. In this way we are able to take advantage of the few available data increasing the final accuracy of the estimate.



To sum up, the model considers the L -moments and L -moments ratios, afterward also referred to as distribution-free parameters, as the variables to be regionalized, according to the definition given in section 2.2. The relationships built to transfer the information to the ungauged sites are discussed in section 2.2.2, while the definition of the distribution used for the final quantile estimation is reported in section 2.3. Finally, a case study is presented in section 2.4 with the definition of a regional model for the alpine basins located in Northwestern Italy. Final remarks are in the conclusions section.

2.2. Model definition

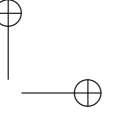
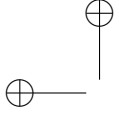
2.2.1. At-site estimates: systematic and non-systematic information

The first step in the procedure is to check the available data and use them to calculate suitable statistical indicators at the gauged sites. Common statistical analyses require the record to be a set of n systematic measures over a certain time span. Systematic measures are rearranged in increasing order

$$x_{(1)}^S \leq x_{(2)}^S \leq \dots \leq x_{(n)}^S, \quad (2.1)$$

where the subscript in round brackets indicates the sorted position.

Sometimes, however, systematic records of data can be integrated with additional data, derived from measurements of significant occasional events. This can be particularly useful when the original systematic record is short. When an additional record of occasional measurements is available, it can be merged with the systematic ones to improve the robustness of the final estimates [e.g. Bayliss and Reed, 2001]. This is done producing a new time series of equivalent “duration” m , that is the number of years between the first and the last measurement of both the systematic and the occasional record, merged together.



Regional approach for flood quantile estimation in ungauged and 12 data-scarce basins

The new set is used to calculate the probability weighted moments (PWMs), as suggested by Wang [1990]: the merged sample is again arranged in increasing order

$$x_{(1)} \leq x_{(2)} \leq \dots x_{(m-l+1)} \leq x_{(m-l+2)} \leq \dots \leq x_{(m)}, \quad (2.2)$$

where the l largest events, exceeding a threshold x_0 , are considered as a censored sample, whose elements can be either systematic or an occasional data.

When working with censored samples, the theoretical formula for the PWM of order r of a random variable X with distribution function $F(x) = P(X \leq x)$,

$$\beta_r = \int_0^1 x(F)F dF, \quad (2.3)$$

must be splitted in two components, called partial PWM or PPWM, [Wang, 1990],

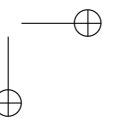
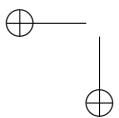
$$\beta_r = \int_0^{F_0} x(F)F dF + \int_{F_0}^1 x(F)F dF = \beta_r'' + \beta_r' \quad (2.4)$$

where $F_0 = F(x_0)$ is the non-exceedance probability relative to the censoring threshold x_0 . The unbiased estimator of β_r'' is then:

$$b_r'' = \frac{1}{n} \sum_{i=1}^n \frac{(i-1)(i-2)\dots(i-r)}{(n-1)(n-2)\dots(n-r)} x_{(i)}'' \quad (2.5)$$

where $x_{(i)}''$ is deduced from the sorted systematic sample as

$$x_{(i)}'' = \begin{cases} x_{(i)}^S & \text{if } x_{(i)}^S < x_0, \\ 0 & \text{if } x_{(i)}^S \geq x_0. \end{cases}$$



On the other hand, the estimator of β'_r is

$$b'_r = \frac{1}{m} \sum_{i=1}^m \frac{(i-1)(i-2)\dots(i-r)}{(m-1)(m-2)\dots(m-r)} x'_{(i)} \quad (2.6)$$

where $x'_{(i)}$ is the above-threshold sample obtained from the complete merged sample, i.e.

$$x'_{(i)} = \begin{cases} 0 & \text{if } x_{(i)} < x_0, \\ x_{(i)} & \text{if } x_{(i)} \geq x_0. \end{cases}$$

The unbiased estimator of β_r is $b_r = b'_r + b''_r$.

The censoring threshold represents the level above which all the extreme flood of the non-systematic record are recorded, and can be assumed equal to the smallest non-systematic measure [Bayliss and Reed, 2001] (with some exceptions discussed later in the case study). In the absence of non-systematic information, the above formulas reduces to the usual definition of PWMs.

L -moments and the dimensionless L -moments ratios are then estimated with the usual formulas [e.g. Hosking and Wallis, 1997] as linear combination of PWMs. In this work the first variable of interest is the index-flood, that reads:

$$Q_{ind} = b_0 \quad (2.7)$$

while L_{CV} and L_{CA} are computed as

$$L_{CV} = \frac{2b_1 - b_0}{b_0}, \quad (2.8)$$

$$L_{CA} = \frac{6b_2 - 6b_1 + b_0}{2b_1 - b_0}. \quad (2.9)$$

Also the coefficient of L -kurtosis will be useful onwards for the choice and estimation of appropriate probability distributions, although it is not neces-

sary in the regional model. Its definition is

$$L_{kur} = \frac{20b_3 - 30b_2 + 12b_1 - b_0}{2b_1 - b_0}. \quad (2.10)$$

In addition, the estimates of sample L -moments are integrated with an estimate of their uncertainty, i.e. their variance. The sample variance is always important in statistical analyses, but in this work it is also a key element for the regression model adopted in the regionalization procedure. Elmir and Seheult [2004] proposed a method for the computation of variances and covariances of sample L -moments and ratios of sample L -moments; however, their formulation appear to be inconsistent when applied to short samples, producing in some cases negative estimates of the variance.

The standard deviation of the index-flood is defined, following the Bulletin 17B Appendix 6 [Interagency Advisory Committee on Water Data, 1982], as

$$\sigma_{Q_{ind}} = \sqrt{\frac{1}{n^2} \sum_{x_i < x_0} (x_i - Q_{ind})^2 + \frac{1}{m^2} \sum_{x_i \geq x_0} (x_i - Q_{ind})^2} \quad (2.11)$$

where Q_{ind} is calculated with equation (2.7). It is easy to see that, in the absence of non-systematic data, equation (2.11) reduces to the usual standard deviation of the mean $\sigma_{Q_{ind}} = \sigma_Q / \sqrt{n}$.

The quality of estimates of L_{CV} and L_{CA} is more difficult to assess. In the present work, simplified formulas for the standard deviations of L -moments ratios are deduced from Viglione [2007b] where expressions for the variance are computed from many Monte Carlo simulations. These formulas are respectively

$$\sigma_{L_{CV}} = \frac{0.9 \cdot L_{CV}}{\sqrt{n}} \quad (2.12)$$

and

$$\sigma_{L_{CA}} = \frac{0.45 + 0.6 \cdot |L_{CA}|}{\sqrt{n}}. \quad (2.13)$$

Moreover, sample L_{CV} and L_{CA} are found to be correlated with cross-correlation corresponding to

$$\rho = \frac{1 - \exp(-5 \cdot L_{CA})}{1 + \exp(-5 \cdot L_{CA})}. \quad (2.14)$$

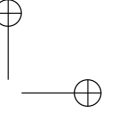
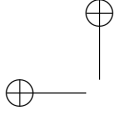
Equations (2.12)-(2.14) are approximated and cannot be easily extended to deal with occasional information. For this, we define $\sigma_{L_{CV}}$ and $\sigma_{L_{CA}}$ calculated only on the systematic sample.

2.2.2. Regression models

After the definition of the variables of interest at the gauged stations, a model to transfer the information to the ungauged sites is needed. In this work, the regional model is intended as a relation that allows one to estimate the first three L -moments in an ungauged basin on the basis of a set of basins descriptors. In the model the relationship, defined by means of a linear regression, is smooth over the whole descriptors domain, without any limitation due to the region boundaries. In this case an homogeneity test is no longer necessary, because the flood frequency curve are allowed to change site by site.

A basic element of the model is \mathbf{Y}_T , the vector containing the true values of the variable of interest, in turn index-flood, coefficient of L -variation and coefficient of L -skewness, or any transformation of these variables. The basic hypothesis is that it can be modelled through the linear relation

$$\mathbf{Y}_T = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\delta} \quad (2.15)$$



Regional approach for flood quantile estimation in ungauged and data-scarce basins

where the $(N \times p)$ matrix \mathbf{X} contains p suitable independent variables relative to N basins and $\boldsymbol{\delta}$ is the error due to the incorrectness of the linear model approximation, i.e. the model error. Nevertheless, in regional frequency analysis applications \mathbf{Y}_T is not known, while one usually can calculate its sample estimator

$$\mathbf{Y} = \mathbf{Y}_T + \boldsymbol{\eta} \quad (2.16)$$

which is affected by a sampling error $\boldsymbol{\eta}$.

Combining equation (2.15) and (2.16), the regressive model reads

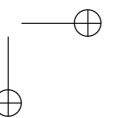
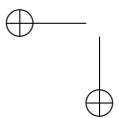
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.17)$$

where $\boldsymbol{\varepsilon}$ is the vector of the residuals that contains both the model and the sampling error.

The simplest method to compute the regression coefficients is the ordinary least squares (OLS) procedure, that is usually not appropriate in hydrological analyses. In fact, its residuals often violate the common assumption of homoscedasticity and independence, and thus the regression coefficients are no longer best linear unbiased estimators (BLUE). This is basically due to the presence of records of different length and cross-correlated [e.g. Steidinger and Tasker, 1985]. To deal with those problems, the weighted and the generalized least squares (WLS and GLS respectively) methods have been developed, although they require the definition of the covariance matrix of the observations.

The regression coefficients $\boldsymbol{\beta}$ are unknown, but the vector containing their unbiased estimators $\hat{\boldsymbol{\beta}}$ can be computed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{Y}, \quad (2.18)$$



where $\mathbf{\Lambda}$ is the covariance matrix. In particular, the ordinary least squares (OLS) are the special case in which $\mathbf{\Lambda}$ is the identity matrix, whereas the weighted least squares (WLS) involve a generic diagonal matrix. In the case of mutual covariances, $\mathbf{\Lambda}$ has non-null values also out of diagonal (GLS).

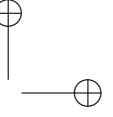
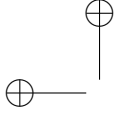
The covariance matrix is interpreted by Stedinger and Tasker [1985] as a function of two terms: the prediction precision of the true model (model error variance) and the sampling error. The method used in this work is based on this approach, where the two uncertainty components are separated and the model error variance is used as a quality index. Note that, usually, the solution for models referred to as WLS and GLS does not account for the model error. If one considers a non-exact model [Stedinger and Tasker, 1985, Griffis and Stedinger, 2007], i.e. the model as an approximation of a real unknown functional relation, the GLS approach requires an iterative solution. In such a situation, to avoid misunderstandings due to the notation, we will refer to as iGLS (or iWLS), where the “i” stands for “iterative”. In this case $\mathbf{\Lambda}$ is approximated by its estimator, defined as

$$\hat{\mathbf{\Lambda}}(\sigma_\delta^2) = \sigma_\delta^2 \mathbf{I}_N + \hat{\mathbf{\Sigma}} \quad (2.19)$$

where $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix of the previously estimated \mathbf{Y} , \mathbf{I}_N is the identity matrix and σ_δ^2 is the model error variance. The regression coefficients $\hat{\boldsymbol{\beta}}$, computed with equation (2.18), and σ_δ^2 are (jointly) estimated [Griffis and Stedinger, 2007] searching for nonnegative solution to the equation

$$\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \left[\hat{\sigma}_\delta^2 \mathbf{I}_N + \hat{\mathbf{\Sigma}}\right]^{-1} \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = N - p \quad (2.20)$$

where $\hat{\sigma}_\delta^2$ is the estimate of σ_δ^2 , N is the number of catchments and p is



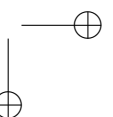
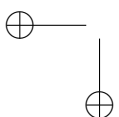
the number of independent variable used in the regression (including the intercept).

In the paper by Stedinger and Tasker [1985] and related works, a complete covariance matrix $\hat{\Sigma}$ is used, that includes covariances in the off-diagonal elements. In this study, the basins are assumed to be independent of each other, because of the highly climatic heterogeneity of the area, thus $\hat{\Sigma}$ reduces to a diagonal matrix. Strictly speaking, our model follows the iWLS approach, although all the equations are still valid for the iGLS case.

2.2.3. Model selection

In regional analyses a great number of physical descriptors at the basin scale is available nowadays, thanks to accurate digital terrain models and remotely sensed data. Despite this, only few characteristics can be used in a robust model for the estimation of the hydrological variables in ungauged sites. The problem is thus the choice of a suitable subset of descriptors to be used in the regression in order to obtain the best final estimates, i.e. to choose the most appropriate regression model among all the possible combinations. Usually, this choice is based on the analysis of the regression residuals: models with the lower coefficient of determination R^2 are favored. In the approach based on GLS, the model error itself can be directly used to select the more appropriate models.

Moreover, when R^2 (or similar metrics) is used to select the best model from a set of different candidate models, it is important to account for the number of descriptors involved in each relation (degrees of freedom of the model). In fact, a larger number of independent variables improves the prediction ability of the model, but decreases its robustness. For this purpose, the adjusted R^2 is currently used, that allows one to compare models with a different number of descriptors. Using the model error for the regression model selection, the degrees of freedom are naturally accounted for by the right hand side of equation (2.20), through the term $N - p$. Model identifica-



tion can be performed also on the basis of the average variance of prediction (*AVP*) [Griffis and Stedinger, 2007] defined as

$$AVP = \hat{\sigma}_\delta^2 + \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \left(\mathbf{X}^T \hat{\Lambda}^{-1} \mathbf{X} \right)^{-1} \mathbf{x}_i^T \quad (2.21)$$

where \mathbf{x}_i is the row vector of \mathbf{X} relative to the i -th basin, that include, in average, the effect of the sampling error.

Since no prior knowledge is considered about the processes that generate the hydrological variables involved in the regional approach, the statistical significance of the independent variables used in the model have to be tested. The method used to test parameter significance is the standard Student t -test, applied to each estimated regression coefficient $\hat{\beta}_j$. The null hypothesis $H_0 : \beta_j = 0$ is tested using the statistic

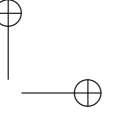
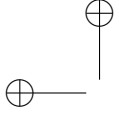
$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\text{var}(\hat{\beta}_j)}} \quad (2.22)$$

[e.g. Montgomery et al., 2001] where the variance of the regression coefficient is taken from the diagonal of the sampling covariance matrix [Reis et al., 2005]

$$\text{cov} [\hat{\boldsymbol{\beta}}] = \left(\mathbf{X}^T \hat{\Lambda} \mathbf{X} \right)^{-1}. \quad (2.23)$$

The t_0 statistic is compared against its limit value and the null hypothesis is rejected if $|t_0| > t_{\alpha/2, n-p}$, where t is the quantile of the (two-tailed) Student distribution with a confidence level α and $n - p$ degrees of freedom.

The regression is also checked against multicollinearity, in order to avoid to select descriptors that are nearly linearly-related among themselves. The test used for this purpose is the variance inflation factor (VIF) test [e.g. Montgomery et al., 2001] with a limit value of 5, that is commonly accepted



as an indicator of absence of multicollinearity.

After the choice of the most appropriate model, we calculate the predicted value and its variance in an ungauged basin. Hence forward we use the “ $\hat{\cdot}$ ” symbol to refer to the value predicted by the regression, while the symbol without any mark is the sample estimate. Let \hat{Y} be the predicted value at a site, i.e.

$$\hat{Y} = \mathbf{x}\hat{\boldsymbol{\beta}}, \quad (2.24)$$

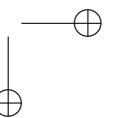
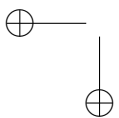
where \mathbf{x} is the row-vector of descriptor relative to the investigated basin and $\hat{\boldsymbol{\beta}}$ the regional regression coefficients (equation (2.18)); the variance of \hat{Y} is thus

$$\sigma_{\hat{Y}}^2 = \hat{\sigma}_\delta^2 + \mathbf{x} \left(\mathbf{X}^T \hat{\boldsymbol{\Lambda}}^{-1} \mathbf{X} \right)^{-1} \mathbf{x}^T \quad (2.25)$$

with \mathbf{X} taken from the calibration dataset and $\hat{\boldsymbol{\Lambda}}$ from equation (2.19).

2.3. Selection of the probability distribution

The final aim of a regional procedure is to estimate the flood quantile and its uncertainty for a specific return period at an ungauged site. So far, however, the procedure presented above focused only on modelling Q_{ind} , L_{CV} and L_{CA} leaving aside the problem of the distribution choice. Classic approaches, instead, explicitly require to adopt a probability distribution to describe the at-site data and then apply the regionalization to some quantiles or, alternatively, to the distribution parameters. The necessity of defining such a probability distribution introduces an additional source of uncertainty due to the ambiguity in the choice of the distribution, in particular when one deals with short samples. Indeed, for low return periods, there are more than one distribution that fit well the data, and the adoption of a suitable



distribution for the regional model is not trivial. We use the regional models to provide Q_{ind} , L_{CV} and L_{CA} estimates and their uncertainty at ungauged sites and then follow a model-averaging approach to define the quantile for a specific return period.

The model averaging approach is based on the idea that more than one distributions may be suitable for quantile estimation. Instead of choosing only one distribution (among those that behave well in the fitting range), it is possible to evaluate many of them and then to take their average. A numeric index that measures the fitting performance on sample data can be used as the weight in the averaging procedure (e.g. the Akaike's Information Criterion [Burnham and Anderson, 2002]). This approach can be applied in the first place to the sample sets of Q_{ind} , L_{CV} and L_{CA} in order to identify a subset of suitable distributions; thus, the model averaging is implemented using the regional estimates of \hat{Q}_{ind} , \hat{L}_{CV} and \hat{L}_{CA} for the site of interest. All the distribution frequency curves involved in the model-averaging procedure are computed only on the basis of the three parameters Q_{ind} , L_{CV} and L_{CA} . Suitable equations to obtain flood quantiles from L -moments, for a set of distributions commonly used in hydrology, can be found in Hosking and Wallis [1997].

The uncertainty of the final estimates is the last step in the regional procedure. Since regional Q_{ind} , L_{CV} and L_{CA} are equipped with their variance and covariance values, we can use a Monte Carlo simulation to define the confidence limits of the frequency curves adopted. The same can be done with the sample L -moments. The method is summarized as follow: (i) for each basin the sample or regional Q_{ind} , L_{CV} and L_{CA} are computed as well as their variances; (ii) a set of factitious Q'_{ind} , L'_{CV} and L'_{CA} are randomly extracted from the distribution of each L -moment; (iii) the parameters of the reference distributions are computed on the basis of the L -moments of point (ii) and the quantile is estimated for the required return period; (iv) points (ii) and (iii) are repeated for a great number of times so that the dis-

10 – Chisone a S.Martino

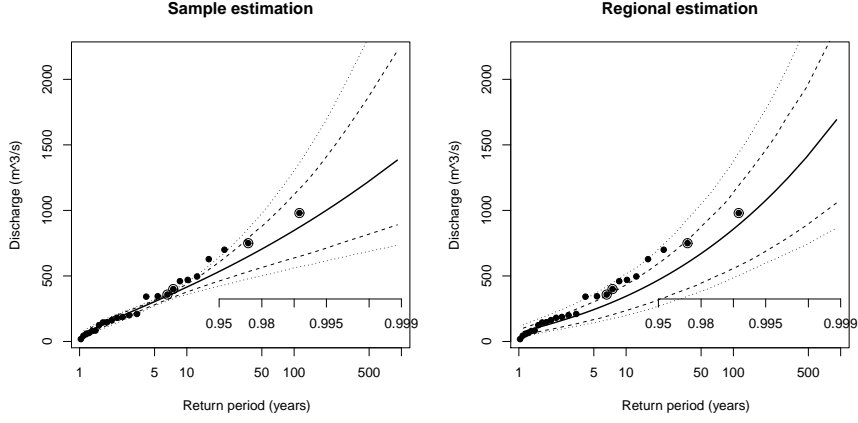


Figure 2.1.: Example of quantiles confidence bands for the river Chisone at S. Martino obtained with a monte carlo simulation. Panel (a) reports the values based on a lognormal distribution applied on sample L -moments, while panel (b) is based on the lognormal distribution fitted on regional L -moments.

tribution of the quantile can be empirically estimated; (v) confidence bands are defined on the distribution of point (iv).

Note that when dealing with regional estimates, Q_{ind} , L_{CV} and L_{CA} are independently estimated so that we can extract the index-flood from a log-normal distribution $Q'_{ind} \sim \log \mathcal{N}(\hat{Q}_{ind}, \sigma_{\hat{Q}_{ind}}^2)$ and the L -moments ratios from two independent normal distributions: $L'_{CV} \sim \mathcal{N}(\hat{L}_{CV}, \sigma_{\hat{L}_{CV}}^2)$ and $L'_{CA} \sim \mathcal{N}(\hat{L}_{CA}, \sigma_{\hat{L}_{CA}}^2)$.

Differently, the uncertainty of a quantile based on sample data depends of mutual correlated L_{CV} and L_{CA} (equation (2.14)), so the index-flood is sampled from the normal distribution $Q'_{ind} \sim \mathcal{N}(Q_{ind}, \sigma_{Q_{ind}}^2)$ while the L -moments ratios are jointly extracted from a multinormal distribution $(L'_{CV}, L'_{CA}) \sim \mathcal{N}(L_{CV}, \sigma_{L_{CV}}^2, L_{CA}, \sigma_{L_{CA}}^2, \rho)$. An example of quantile estimation with confidence bands is reported if figure 2.1.

2.4. Case study

2.4.1. Data availability

The methods described above are applied to a set of 70 catchments located in the northwestern part of Italy (see figure 2.2 and appendix A). The analysis is carried out on small basins belonging mainly to mountainous areas, with area ranging between 22 and 3,320 km² and mean elevation from 471 to 2,719 m a.s.l. To reduce any effect of upstream lakes and/or reservoirs, we discarded basins where catchment area is covered by lakes in a percentage beyond 10%.

The first step in the model building is the analysis of available data of annual streamflow maxima. To improve the sample statistics, we adopt the method described in section 2.2.1 to include non-systematic and systematic information about large flood occurred in the area. Occasional values are retrieved from reports issued by the National or Regional environment agencies and refer to unusually intense events occurred when no systematic measurements were available. Sometimes, such extreme events involved only a limited area, but the hydrological surveys were extended to a wider zone, and thus the reconstructed flood flow was reported also in sites where the flood was not extreme. In these cases, the smallest occasional event is no longer usable as censoring threshold because there is no evidence that all the larger occasional extreme floods, happened in that basin during the “un-gauged” period, are known (all the events above threshold need to be known [Wang, 1990]). Where this kind of problem is observed, we adopt a threshold equal to a greater non-systematic measure.

The method for including occasional information allows one to extend the time series length of 18 basins using a total of 36 occasional measurements. The equivalent time series are, in average, 20 years longer than those without non-systematic information.

After the data checking, the sample index-flow, L_{CV} and L_{CA} coefficients,

Regional approach for flood quantile estimation in ungauged and data-scarce basins

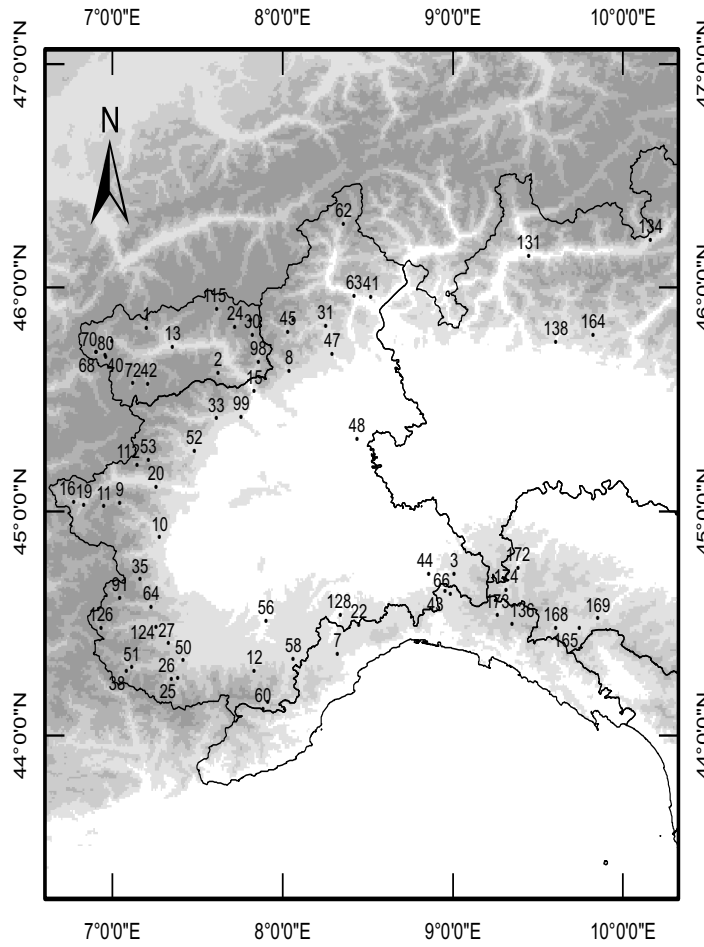


Figure 2.2.: Geographical location of the gauging stations used for the calibration and validation of the present work. This area is located in northwestern Italy and the basins are mainly on mountainous environments.

as well as their standard deviations, are computed with the equations of section 2.2.1. A short summary of the sample coefficients is shown in figure 2.3, panel (a) for the index-flood and panel (b) for L_{CV} and L_{CA} , where the filled circles highlight the catchments with non-systematic information.

A set of 40 basins descriptors has been defined for the group of catchments involved in this analysis (appendix B), using geomorphologic and climatic characteristics available in the CUBIST database [CUBIST Team, 2007].

2.4.2. Model definition

The model structure adopted in this work is linear, with parameters determined with an improved least squares procedure, as discussed in detail in section 2.2. Although this model has an additive structure (see equation (2.15)), in hydrology it is common to use also multiplicative models in the form

$$Y = \alpha_1 X_2^{\alpha_2} X_3^{\alpha_3} \dots X_p^{\alpha_p} \varepsilon \quad (2.26)$$

that can be reduced to linear additive form by means of a log-transformation of both sides of the equation.

Both additive and multiplicative model structures for each distribution-free parameter are examined; details on the variables involved and on the transformation applied are summarized in table 2.I, where a concise name is assigned to each model structure. In particular, concerning the index-flood, two additive and two multiplicative models are considered, with the dependent variable equal to Q_{ind} or to Q_{ind}/A , where A is the catchment area. These models will be referred to as Qind, QindA, lnQind and lnQindA respectively. The regional model for L_{CV} is still based on an additive model (named LCV) and a multiplicative one (lnLCV), while the L_{CA} is investigated through an additive model (LCA). A direct application of the multiplicative model to L_{CA} is not possible due to the non-positiveness of the variable that does not allow a logarithmic transformation.

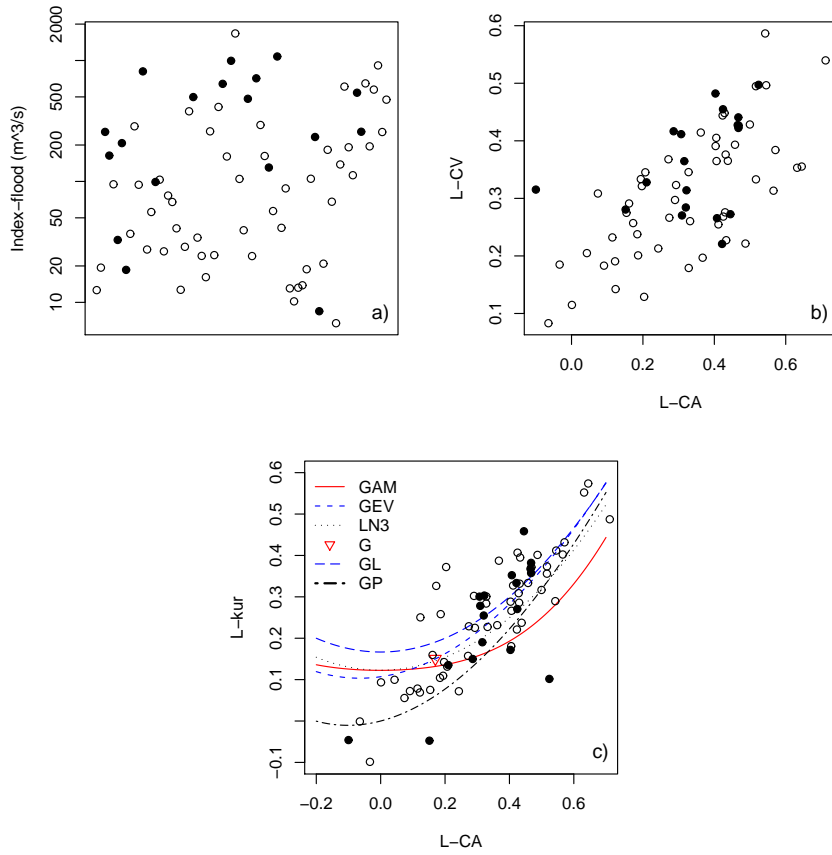
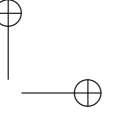
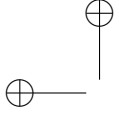


Figure 2.3.: Summary of sample estimates for the 70 basins located in Northwestern Italy. Panel (a) shows the index-flood values, while panels (b) and (c) reports the diagnostic diagrams [Hosking and Wallis, 1997] with, respectively, L_{CV} versus L_{CA} and L_{CA} versus L_{kur} . For all the panels, filled circles indicate the basins in which non-systematic information have been included in the analysis.

Table 2.1.: Different model structures used in the analysis. The first four models deal with the estimation of the index-flood, while the remaining ones are relative to L -moments ratios. The first two columns indicate the model denomination and the variable involved, followed by the transformation applied, if any, and the equation to compute the sample standard deviation. The last column provides the matrix of independent variables \mathbf{X} to be used in the linear regression, that depends on the descriptors matrix \mathbf{X}_d in which each column is a different descriptor and each row a different catchment. The symbol $\mathbf{1}$ indicates an unitary column vector introduced to account for the intercept coefficient in equation (2.17).

Model denomination	Original variable	Transformation	Sample standard deviation	Descriptors
Qind	Q_{ind}	none	from eq. (2.11)	$\mathbf{X} = [\mathbf{1}, \mathbf{X}_d]$
QindA	Q_{ind}	Q_{ind}/A	$\sigma_{Q_{ind}}/A$	$\mathbf{X} = [\mathbf{1}, \mathbf{X}_d]$
lnQind	Q_{ind}	$\log(Q_{ind})$	$\sigma_{Q_{ind}}/Q_{ind}$	$\mathbf{X} = [\mathbf{1}, \log \mathbf{X}_d]$
lnQindA	Q_{ind}	$\log(Q_{ind}/A)$	$\sigma_{Q_{ind}}/Q_{ind}$	$\mathbf{X} = [\mathbf{1}, \log \mathbf{X}_d]$
LCV	L_{CV}	none	from eq. (2.12)	$\mathbf{X} = [\mathbf{1}, \mathbf{X}_d]$
lnLCV	L_{CV}	$\log(L_{CV})$	$\sigma_{L_{CV}}/L_{CV}$	$\mathbf{X} = [\mathbf{1}, \log \mathbf{X}_d]$
LCA	L_{CA}	none	from eq. (2.13)	$\mathbf{X} = [\mathbf{1}, \mathbf{X}_d]$



Regional approach for flood quantile estimation in ungauged and data-scarce basins

The best models to be used for the regional estimation are not known a priori, but are identified among all the possible combinations of descriptors considering 1 to 5 parameters (1 to 4 descriptors in addition to the intercept). The limit of 4 descriptor is mainly due to the computational effort in investigating all the combinations ($\sim 65,000$ combinations with 35 descriptors), and considering that more than 4 descriptors usually do not improve the efficiency and the robustness of the final estimates. Anyway, a preliminary investigation can be helpful for reducing the number of useful descriptors and thus reduce the computational requirements.

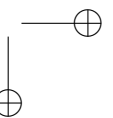
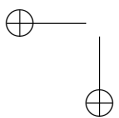
These models are then tested for significance and multicollinearity and the ones passing the Student and the VIF tests are ranked on the basis of their model variance ($\hat{\sigma}_\delta^2$) and the average variance of prediction (*AVP*). Models that appear to be more efficient are checked in order to verify the basic regression hypotheses (see diagnostic plots in figures 2.4-2.7). Finally, a set of few operational models for each variable is selected.

When the variable of interest is log-transformed, equations (2.24) and (2.25) yield estimates that are not directly usable and need to be back-transformed to their original space. In this case, if the regression residuals are normally distributed, also \hat{Y} is normally distributed and its back-transformation leads to a lognormal variable. Therefore, we evaluate the mean of the estimate as

$$\mu = \exp\left(\mu_{\hat{Y}} + \frac{1}{2}\sigma_{\hat{Y}}^2\right) \quad (2.27)$$

with $\mu_{\hat{Y}}$ equal to \hat{Y} , estimated with the regression in the logarithmic space (equation (2.24)), and $\sigma_{\hat{Y}}^2$ that comes from equation (2.25), while its variance reads

$$\sigma_\mu^2 = \mu^2 \cdot [\exp(\sigma_{\hat{Y}}^2) - 1]. \quad (2.28)$$



This back-transformation can prevent to have large biases [e.g. Seber and Wild, 1989, 2.8.7]; however, we verified that for this case study, the simple exponential transformation

$$\mu' = \exp(\hat{Y}). \quad (2.29)$$

allowed us to reconstruct the variable in its original space without appreciable errors.

2.4.3. Regression results

A number of acceptable model structures obtained after sorting the models are reported in table 2.II together with a short summary of the prediction performances errors, i.e. the RMSE, MAE and Nash efficiency. These indexes are computed after a cross-validation procedure: one basin is temporarily removed from the database and the regression coefficients are calculated on the basis of the remaining catchments. The variable of interest is computed at the “temporarily ungauged” site and the procedure is repeated for all the basins under analysis. Table 2.III is again related to these models, but reports the regression coefficients and the descriptors selected. Note that the regression coefficients refer only to the additive model, i.e. they lead to the estimation of \hat{Y} . When a transformation of the original variable is involved, it is necessary to back-transform the predicted value (e.g. to calculate μ or μ' from \hat{Y} , for the logarithmic case).

Concerning Q_{ind} , as expected, the choice of a regression model is easier and leads to a rather efficient estimation of the variable. Among the possible model structures (linear or log-transformed; normalized or not by the catchment area), a preliminary analysis showed that the most suitable model is $\ln Q_{ind}$. The log-transformation requires some attention, but avoids the problem of having negative estimates when using additive models.

Table 2.II.: Summary statistics for the selected models.

Model	tstud test	σ_0^2	AVP	σ_0	\sqrt{AVP}	Nash	RMSE	MAE	Nash ^a	RMSE ^a	MAE ^a
lnQind1	1%	0.1153	0.1248	0.340	0.353	0.913	91.5	55.0	0.894	101.2	60.1
lnQind2	1%	0.1338	0.1445	0.366	0.380	0.844	122.9	63.8	0.801	138.7	70.6
lnQind3	1%	0.1398	0.1488	0.374	0.386	0.749	155.8	70.7	0.669	178.9	77.1
LCV1	2%	0.0044	0.0049	0.066	0.070	0.238	0.092	0.070	0.159	0.097	0.074
LCV2	1%	0.0051	0.0056	0.071	0.075	0.155	0.097	0.074	0.084	0.101	0.078
LCV3	2%	0.0052	0.0057	0.072	0.075	0.137	0.098	0.075	0.070	0.102	0.078
lnLCV1	1%	0.0520	0.0570	0.228	0.239	0.368	0.084	0.066	0.293	0.089	0.070
lnLCV2	2%	0.0723	0.0785	0.269	0.280	0.173	0.096	0.074	0.093	0.101	0.078
LCA1	2%	0.0070	0.0083	0.083	0.091	0.184	0.157	0.125	0.107	0.164	0.131
LCA2	1%	0.0071	0.0083	0.084	0.091	0.172	0.158	0.122	0.086	0.166	0.129
LCA3	1%	0.0085	0.0098	0.092	0.099	0.110	0.164	0.130	0.006	0.173	0.137
LCA4	2%	0.0084	0.0098	0.092	0.099	0.116	0.163	0.129	0.025	0.171	0.136

^a after cross-validation procedure

Table 2.III.: Descriptors used as independent variables in the regression and relative coefficients for different dependent variables and different model structures. In case of log-transformation, note that the coefficients always refer to the linear model. For a short description of the independent variables, refer to table 2.IV.

Model	par1	par2	par3	par4	par5	coeff1	coeff2	coeff3	coeff4	coeff5
lnQind1	intercept	lnA	lna	lnMAP	lnc _f	-8.76E+00	7.99E-01	1.09E+00	9.53E-01	7.85E-01
lnQind2	intercept	lnY _c	ln(H/√A)	lnn	lnMAP	-1.51E+02	9.85E+00	-1.57E+00	1.63E+00	1.72E+00
lnQind3	intercept	lnΔH ₁	ln(H/√A)	lnMAP		-7.17E+00	5.75E-01	-1.31E+00	1.91E+00	
LCV1	intercept	X _c	D _d	P	H _{min}	9.35E-01	-3.94E-07	-5.14E-01	-4.79E-04	-1.48E-04
LCV2	intercept	X _c	P	H _{min}		6.44E-01	-4.28E-07	-5.00E-04	-1.44E-04	
LCV3	intercept	X _c	A	H _{min}		5.88E-01	-3.90E-07	-5.06E-05	-1.28E-04	
lnLCV1	intercept	lnY _c	lnD _d	lnc _f		1.55E+02	-1.02E+01	-1.41E+00	4.44E-01	
lnLCV2	intercept	lnY _c	lnn	lnH _{min}		1.69E+02	-1.09E+01	8.92E-01	-1.76E-01	
LCA1	intercept	MHL	D _d	MAP		1.97E+00	-1.05E-03	-9.75E-01	-2.18E-04	
LCA2	intercept	a	P	H _{min}		9.38E-01	-1.40E-02	-1.39E-03	-2.65E-04	
LCA3	intercept	A	MAP	H _{min}		6.98E-01	-9.87E-05	-1.89E-04	-1.18E-04	
LCA4	intercept	MSL	MAP	H _{min}		7.82E-01	-2.65E-03	-1.96E-04	-1.54E-04	

Regional approach for flood quantile estimation in ungauged and data-scarce basins

Table 2.IV.: Short description of descriptors involved in regional models of table 2.III. See appendix B for more detailed references.

A	Catchment area
X_c	Longitude of catchment's centroid
H_{min}	Minimum catchment elevation
P	Catchment perimeter
MSL	Main Stream Length
MHL	Mean Hillslope Length
D_d	Drainage density
c_f	Permeability index
MAP	Mean Annual Precipitation
a, n	Coefficients of the precipitation IDF curve in the form $h = ad^n$

The best among the selected models involves four descriptors: the catchment area A , the mean annual precipitation (MAP), the coefficient a of the Intensity-Duration-Frequency (IDF) curve and a permeability index c_f . Coefficient a is related to a monomial representation of the IDF curve of precipitation extremes having the form $h_{d,T} = k_T a d^n$, where $h_{d,T}$ is the precipitation quantile of duration d , k_T represents the growth factor and a and n are catchment-averaged parameters required to set the intensity-duration relation at the index-rainfall.

Figure 2.4 shows the regression diagnostic graph: in particular, panel (a) shows the comparison between sample and estimated log-transformed values. Regression residuals are checked to verify the absence of heteroscedasticity (panel (b)) and normality (panel (c)). Panel (d) shows the comparison between sample and estimated index-flood, reported in the original measurement space. As previously mentioned, the back-transformation should be done with equation (2.27) (see black dots in panel d), nevertheless a simple exponential lead to comparable results (empty circles on the same panel). This entails us to adopt the simpler approach without loss of detail.

Tables 2.II and 2.III report other two models that can be alternately used in place of of the previous one. Both the alternative models are based on

simple geomorphological characteristics and few climatic information, and still have very good residuals diagnostics, even if they present a lower efficiency; as proven by the RMSE, MAE and Nash indexes, they prove to be efficient for a practical use. All the models reported pass the Student test with a level of confidence of 1% and the VIF test with a limit value of 5.

L_{CV} is an indicator of the sample dispersion and the reliability of its assessment is investigated through an additive model as well as a multiplicative one. These approaches are respectively referred to as LCV and lnLCV. Regarding LCV, we obtain only a few models that pass the Student test with a confidence level of 1%, therefore we relax the test level to 2%. The first-ranked model (see table 2.II and 2.III) has four descriptors: the longitude of the catchment centroid, the drainage density, the basin perimeter and the elevation of the gauging station. Drainage density is, however, not easy to calculate and depends on the GIS procedure adopted for the river network delineation. For this reason, we propose to use the second model (named LCV2) that is structurally similar to the first one and has comparable diagnostic statistics, but does not require the drainage density. Diagnostic plots for this latter model are shown in figure 2.5. Tables 2.II and 2.III also report a third alternative model for L_{CV} .

L_{CV} estimation can be approached also by multiplicative models (lnLCV model structure) and tables 2.II and 2.III show the results of this approach. Model error and AVP are not directly comparable with those correspondent to LCV models, but the RMSE, MAE and Nash efficiency, calculated on the final estimated values, show that these models have a good predictive ability. Despite this, lnLCV models present problems in the regression residuals plots (see figure 2.6) with high deviations from the theoretical normal behavior. For this reason we suggest to leave aside the multiplicative models for L_{CV} in favor of the additive ones. From figure 2.5 (panel (a)) we note that predictions are systematically overestimated for small values and underestimated for large ones. On one hand, this behavior, due to the limitations of

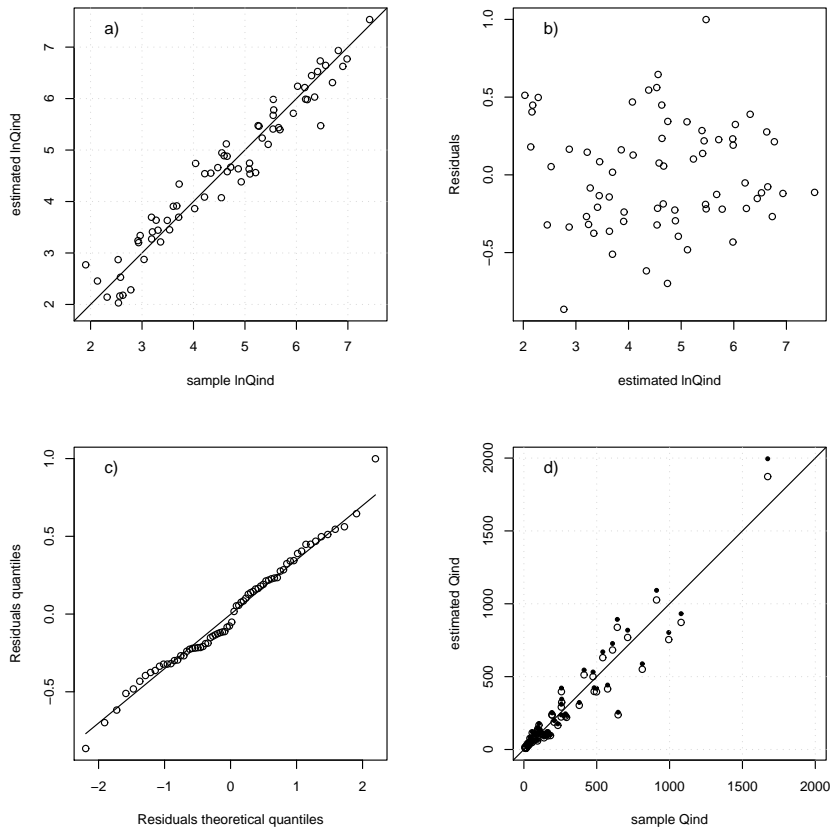


Figure 2.4.: Diagnostic diagram for model lnQind1. Panels (a), (b) and (c) refer to the log-transformed values, and show respectively the comparison between sample and estimated values, the residuals behavior and the residuals normality plot. Panel (d) shows the comparison between sample and estimated values in the original index-flow space. Empty and filled circles differ for the back-transformation used: the formers are simply the exponential of the regression estimates (eq. (2.29)), while the latter are computed as the mean of the related log-normal distribution (eq. (2.29)).

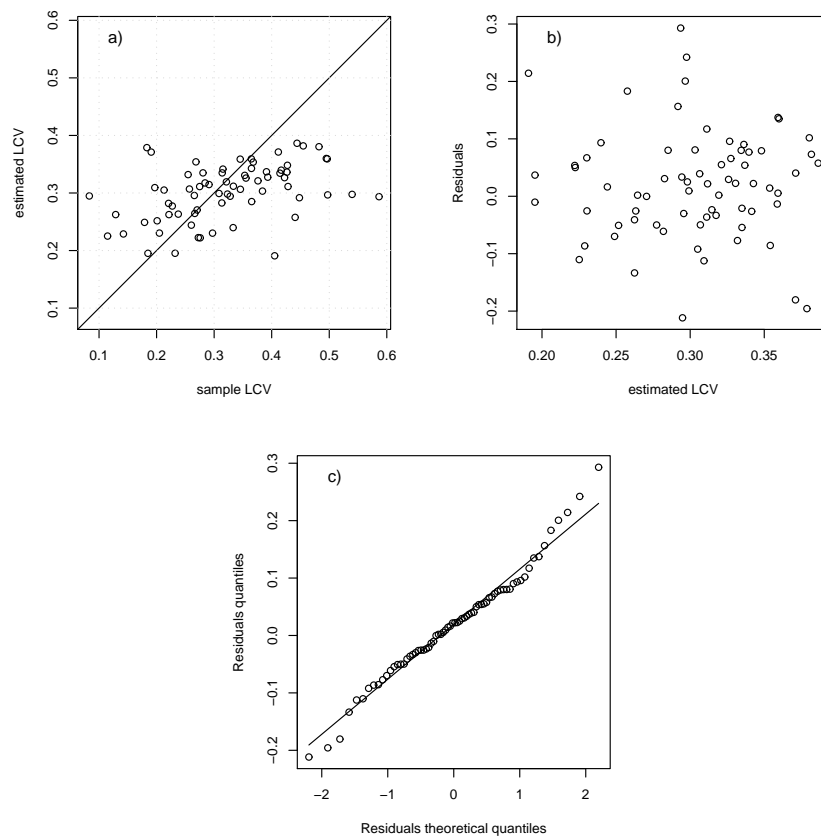


Figure 2.5.: Diagnostic plots for model LCV2. Panel (a) shows the systematic underestimation for small values and over estimation for large values, however residuals behavior confirm the absence of heteroscedasticity (panel (b)) and a good alignment to the theoretical normal distribution (panel (c)).

the multiple (linear) regressive approach based on a set of simple descriptors, can be seen as a weak result. On the other hand, this result has the merit to stem from a robust regression model, that respect the modelling hypotheses, and that yields a valuable estimation of uncertainty.

The last parameter required for building flood regionalized distributions is the coefficient of L -skewness (L_{CA}) that is investigated only by the additive model, differently from L_{CV} and Q_{ind} . General remarks about L_{CA} are very similar to those relative the L_{CV} ; in fact, the best model we obtain is characterized by three descriptors: mean hillslope length, drainage density and mean annual precipitation. The first two descriptors are yet dependent on the GIS procedure used for their computation and should be used carefully. To avoid this inconvenience, we prefer a second model (LCA2) that is analyzed in greater detail in figure 2.7; residuals diagnostics show no evidence of heteroscedasticity (panel (b)) and a good alignment to the theoretical residuals quantiles distribution (panel (c)). As for L_{CV} , panel (a) of the same figure shows that the model is not able to capture the whole sample variability.

Equation (2.21) shows that the average variance of prediction (AVP) is constituted by the sum of two terms, the model error and the site-specific error due to sample variability. The relative effect of these factors can be seen in table 2.II, where we observe that the ratio between model variance and AVP is 0.92 for both $\ln Q_{ind}$ and L_{CV} , while decreases to 0.85 for L_{CA} . This suggests that the model error dominates the total error, as expected because of the limitations of the linear regression model. This value decrease for L_{CA} and this is probably due to the greater sample uncertainty that increases the AVP value, rather than decreasing the model error. Either the model variance or the AVP can be used to classify the final models, but in this case the final ranks are equivalent because the AVP is dominated by the model variance.

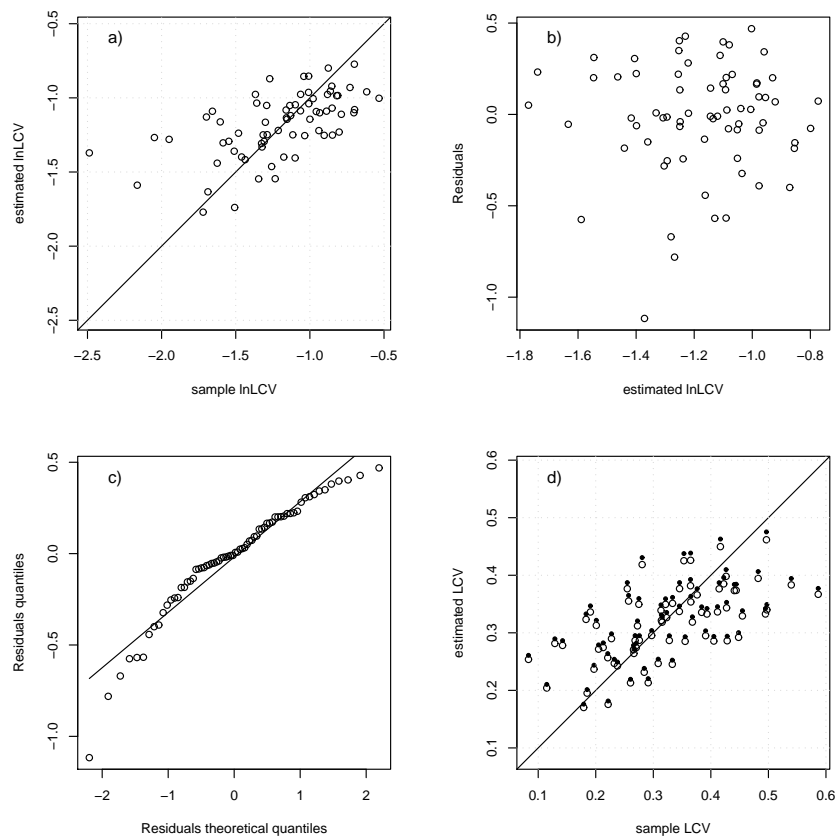


Figure 2.6.: Diagnostic plots for model lnLCV1. Panel (c) shows a weak alignment to the normal theoretical residuals distribution although predicted values have a quality comparable to that of the LCV2 model.

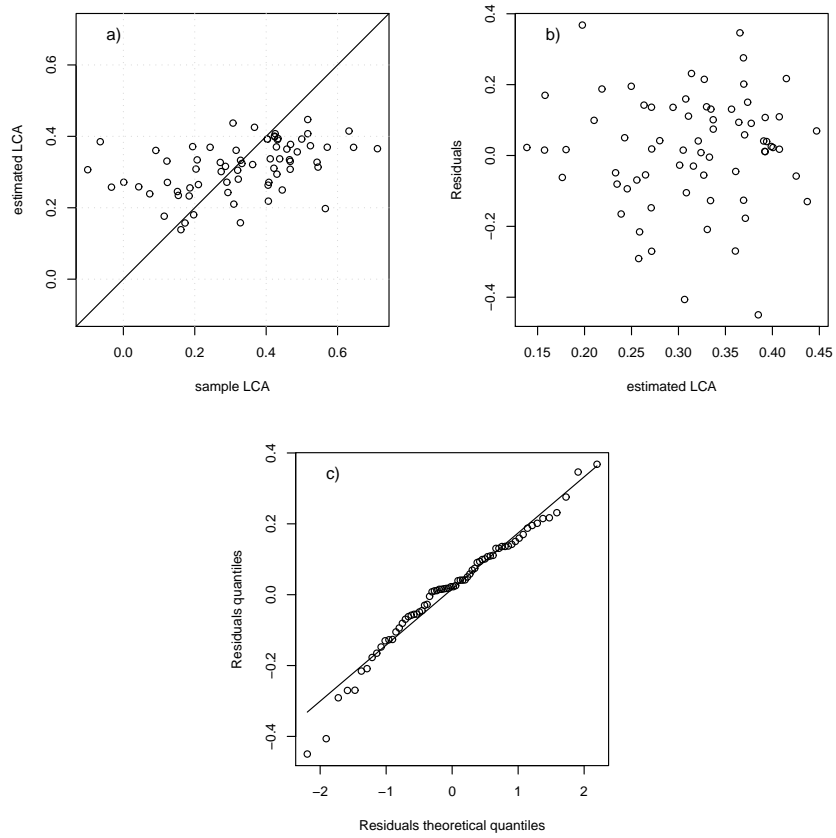


Figure 2.7.: Diagnostic plots for LCA2 model. As for L_{CV} , the model shows no evidence of heteroscedasticity (panel (b)) and a normal distribution of residuals (panel (c)), although systematic under- and overestimations are evident in panel (a).

2.4.4. Quantile estimation

As already mentioned in section 2.3 the final aim of the procedure is the assessment of the flood quantile for a specific return period (with uncertainty). In this work the regional (regression) model is applied to distribution-free parameters in order to avoid the errors induced by the preliminary choice of a distribution probability and/or a subdivision in regions. These errors are particularly difficult to estimate, especially when working with time series of different length and even more so when samples available are short, as in this case study.

In this section, the following issues are discussed: (i) the estimation of a flood-quantile applying the model averaging approach and (ii) the estimation of its uncertainty by means of monte carlo simulations. Concerning the first point we evaluated six different distributions commonly used in hydrology, fitting, for each of them, the frequency curve on the sample data relative to each of the 70 basins under analysis. The distributions are the Pearson type III or Gamma (GAM), the generalized extreme value (GEV), the lognormal with three parameters (LN3), the Gumbel (G), the generalized logistic (GL) and the generalized Pareto (GP). A resume of the relations used for estimation by means of the L -moments can be found in Hosking and Wallis [1997]. The frequency curve obtained in the gauged sites using the sample data can be plotted to check the validity of the hypotheses made. For this purpose, we assign a non-exceedance probability to each sample value by means of a plotting positions. Here the Hazen plotting position is applied, as defined by Hirsch [1987], to include the non-systematic information. Let i be the index of the complete sample (systematic and non-systematic values merged together) sorted in descending order, the non-exceedance probability of the i -th measure is defined as

$$p_i = 1 - \begin{cases} \frac{i-0.5}{m} & i = 1, \dots, k \\ \frac{k}{m} + \frac{m-k}{m} \frac{i-k-0.5}{s-e} & i = k+1, \dots, g \end{cases} \quad (2.30)$$

10 – Chisone at S.Martino

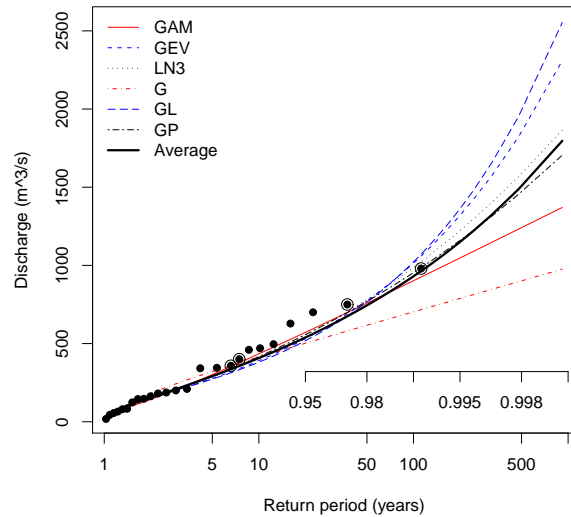


Figure 2.8.: Example of sample flood data for the river Chisone at S. Martino and superposition of different theoretical frequency distributions: Gamma (GAM), generalized extreme value (GEV), lognormal (LN3), Gumbel (G), generalized logistic (GL) and generalized Pareto (GP). Black dots represent empirical data, circled ones correspond to non-systematic events; plotting positions come from equation (2.30).

where m is the equivalent sample length (as in section 2.2.1), g is the total number of flood events available (both systematic and non-systematic), k is the total number of events that exceed the threshold x_0 , s is the length of the systematic sample and e is the number of measures of the systematic sample that exceed the threshold.

An example is shown in figure 2.8 for the river Chisone at S. Martino, where the sample points are plotted using the plotting position of equation (2.30) and the circled dots highlight the non-systematic measurements. The example shows that all the distributions have a similar behavior up to a 100-years return period, except for the Gumbel distribution that is a less flexible distribution with only two parameters.

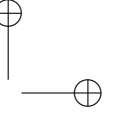
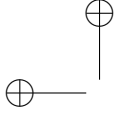
Although goodness-of-fit tests or more advanced techniques [Laio et al., 2009] can be used to select only one distribution, it is clear that all of these models are almost equally suitable; as a consequence we propose to take their average as the final frequency curve for quantile estimation (thicker line in figure 2.8). This model-averaging procedure [Burnham and Anderson, 2002] can appear operationally too complicated. For the sake of simplicity, we propose to compute the quantiles using the lognormal distribution as the one that falls closest to the average frequency curve for each basin under analysis. The second point is the evaluation of the uncertainty of the quantile estimate through the Monte Carlo procedure described in section 2.3.

2.4.5. L-moments estimates in data-scarce stations

Strictly speaking, an ungauged catchment has no data records; thus one needs to use regional models to obtain estimates of all the three L -moments under consideration. However, if only few measurements are available it is sometimes possible to estimate at least the lower-order sample L -moments with an acceptable degree of robustness. In these cases, it is possible to compute both the sample (at-site) and the regional estimate and then choose the one with the lowest uncertainty. To this end, the standard deviation of the sample estimates, computed on the available data through equations (2.11), (2.12) or (2.13), can be compared to the standard deviation of the estimates obtained by the regional model by means of equations (2.25)-(2.28).

This approach allows one to reconstruct the frequency curve in data-scarce sites using a set of L -moments derived from different sources (sample or regional), in order to improve the quality of the final estimates.

An example is shown in figure 2.9 where the index $\Psi = (\hat{\sigma} - \sigma)/\sigma$ (i.e. the regional standard deviation minus the sample one, normalized by the sample one) is reported for all the three L -moments used in this work. Positive Ψ indicates that the sample estimates have lower uncertainty than the regional ones and viceversa. The figure shows that moving to higher-order L -moments



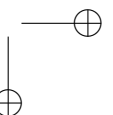
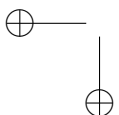
(panel (a) to panel (c)), the regional estimates become more reliable than the sample ones.

When a short record is available, the graphs of figure 2.10 can be used as a practical tool to identify the most suitable way to assess the index-flood, L_{CV} and L_{CA} . It is sufficient to know the record length n , the sample standard deviation and the sample L_{CV} and L_{CA} , to check on the graph if the sample-point falls in the shaded area. In this case, the sample estimate is preferable to the regional one and viceversa. The threshold value (thick line) is the regional prediction variance averaged over the calibration set, thus it is an indicative value. A more precise choice can be done using the formulae previously reported.

2.5. Final remarks

The approach to regional flood frequency analysis proposed in this work aims at overcoming some limitations of the classical methods and at facilitating the use of non systematic measurements that might be retrieved for some catchments. The at-site data are non directly used to build up a locally valid parametric model, however, the sample record is summarized by calculating its L -moments that are afterwards used to reconstruct the complete flood frequency curve. The L -moments become the regional variables that are related to the basins descriptors by means of a regression, that allows the predicted L -moments to vary smoothly over the whole descriptors domain without any grouping or formation of sub-regions.

The representation of sample data by L -moments avoids the uncertainty related to the preemptive choice of a probability distribution and allows one to make wise use of short samples, otherwise discarded. In this way the database can be increased without loss of information. For instance, in the present case study, eight stations out of 70 present 10-20 data they probably would be discarded in a traditional approach. Of course, the uncertainty of these short sample is accounted by the assessment of the L -moments



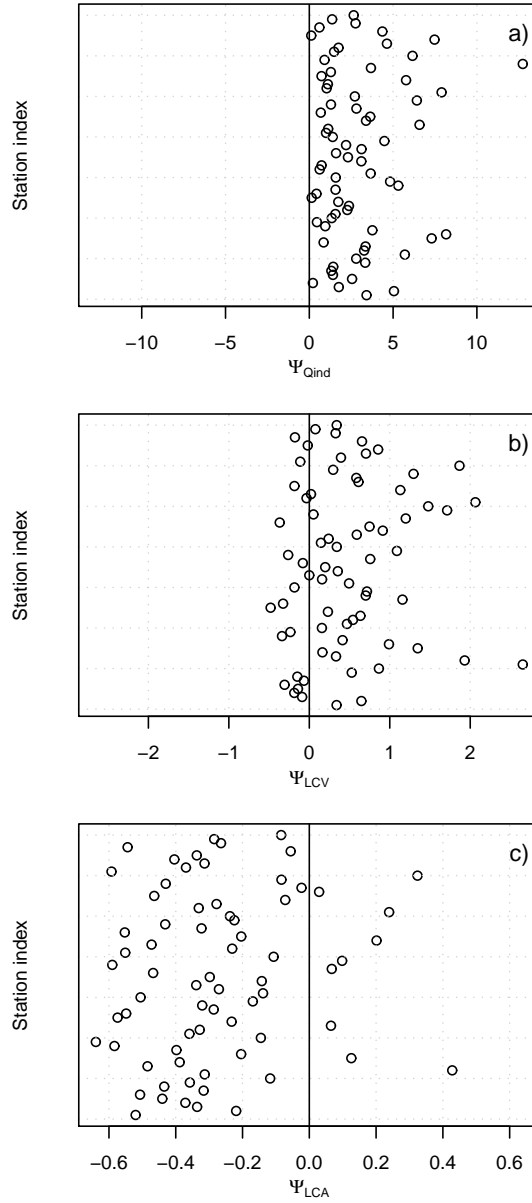


Figure 2.9.: Comparison between regional and sample standard deviations for the index-flood (panel (a)), LCV (panel (b)) and LCA (panel (c)) by means of the index Ψ . Each point greater than zero (right part of the graph) indicates that the sample estimate is preferable to the regional one; on the other hand, the points falling on the left side indicate that a regional estimates are more reliable than the sample ones because they are calculated using a larger amount of information.

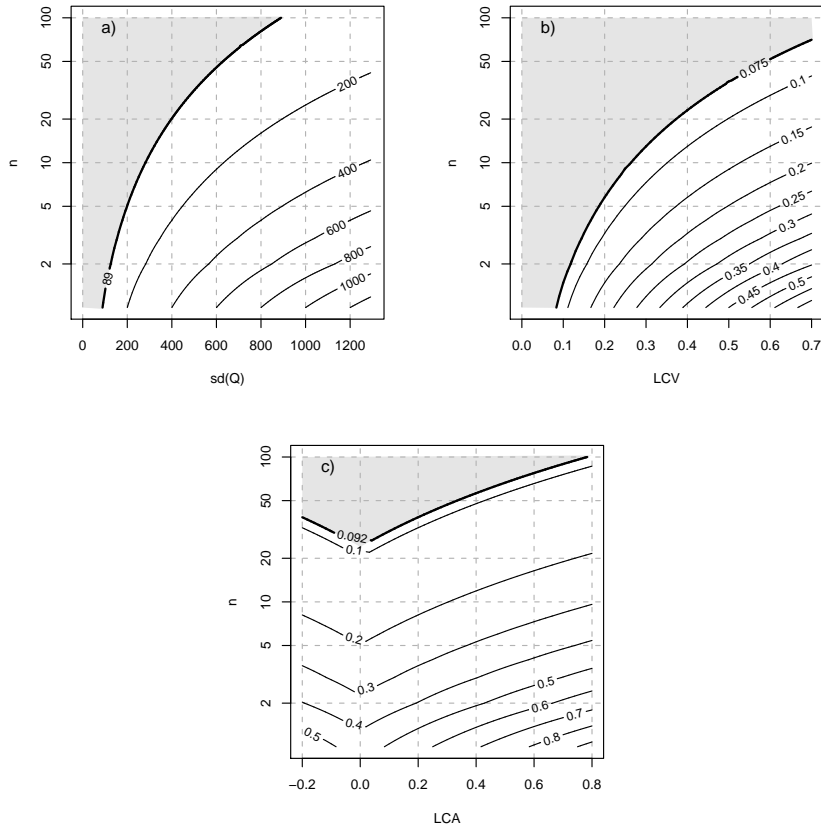


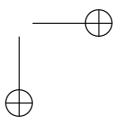
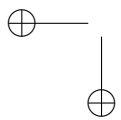
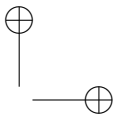
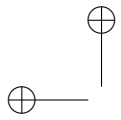
Figure 2.10.: Graphs for the choice of the sample versus regional method to predict the variable of interest, relative to the index-flood (panel (a)), LCV (panel (b)) and LCA (panel (c)). Iso-lines represent points with equal sample standard deviation, while thicker line is equal to the averaged standard deviation of the regional model. The shaded area corresponds to better performances of the sample estimate compared to the regional one.

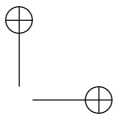
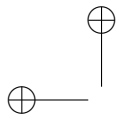
variance in order to weight properly these values in the regressions. With this approach, it is also possible to use the few available data to directly estimate the lower-order L -moments and to adopt the regional model for the other cases.

The regionalization of the L -moments without the creation of (homogeneous) regions, allows one to create a unique relationship for the whole dataset and provides regional predictions of index-flood of high quality. On the other hand, for higher-order L -moments, the regression is not able to completely describe the sample variability. Although the subdivision in regions is perhaps able to produce smaller prediction errors, our approach easily handles the uncertainty of the regression model and avoids the subjectivity of procedures that create regions and estimate their homogeneity. In this sense the model provides a “global” optimization rather than a “local” one.

Finally, the work also considers an approach that includes the non-systematic measurements of flood events. In literature, non-systematic data are commonly referred to as historical flood, occurred before the beginning of the gauged period. However, in the Italian context, often we found time series records with large gaps and few large events measured during this “ungauged” period. Therefore, this information can be considered as non-systematic data and used as a small set of valuable additional measurements.

All these characteristics, make the procedure particularly suitable in data-scarce situations.





Chapter 3.

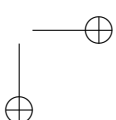
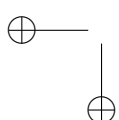
Along-stream estimation approach

Contents

3.1. Introduction	47
3.2. Extension of the regional procedure	50
3.2.1. Regional covariance and correlation	50
3.2.2. Formulae for log-transformed data	52
3.3. Along-Stream information propagation method 53	
3.3.1. Methods and hypotheses	53
3.3.2. Example about functions and assumptions	58
3.3.3. Organization of nested basins	59
3.4. Model reliability: simplified approach	61
3.4.1. Uncertainty of the propagated estimate	61
3.4.2. Assessment of the variance parameter	63
3.4.3. Validity of the simplified approach	65
3.5. Model reliability: analytical approach	71
3.6. Final remarks	73

3.1. Introduction

Regional models aim at transferring information from a set of gauged sites to the ungauged basin of interest. Although different types of models have been developed in literature, their common attitude is to approach the lack of hydrologic information moving to the so called descriptors space. The latter is a set of catchment characteristics, usually topographic, morphological, pedological or climatic indexes, that are computable for every basin without resorting to any hydrologic data. Then, suitable relationships are built to relate these characteristics to the desired hydrological variable.



Differently from the regional approaches, the basic concept underpinning the model developed in this chapter is the transfer of hydrological information to an ungauged site located upstream or downstream the gauging station. The information we are interested in, i.e. the information we transfer along the stream network, are those used to reconstruct the flood frequency curve, such as the L -moments mentioned in the chapter 2. Note that this information is not related to the discharge value at a particular time, that could be investigated, for instance, by means of a rainfall-runoff model, but rather to a characteristic discharge, as could be the average annual runoff. In addition to the hydrologic information, this approach is based on the knowledge of the structure of the drainage network, in order to properly identify how points are directly connected. An equivalent way to describe the same concept is by saying that the two basins are nested. The model is therefore named Along-Stream (AS) approach, and it involves at least one variable calculated in a gauged (or donor) basin, that is used to propagate the information towards the ungauged (target) site where the variable of interest is reconstructed. In order to avoid a misleading notation, hereafter we will refer to the donor station using the subscript d , while subscript t will be used for the target site.

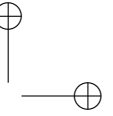
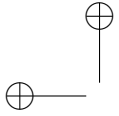
The issue of prediction or interpolation of hydrological variables along the river network is not frequently discussed in the literature, although some notable examples are present. Gottschalk [1993a,b] introduced the problem of correlation and covariance of runoff and its interpolation along the river network, adapting the theory of stochastic processes to the hierarchical structure of nested catchments. This approach has been extended by Gottschalk et al. [2006], and the same concepts are used by Skoien et al. [2006] in the development of a kriging procedure that accounts for the river structure, named topological kriging or top-kriging. Although the final aim is the same, the procedure developed here follows a different approach.

Kjeldsen and Jones [2007], instead, studied the problem of interpolation

of runoff statistics considering a local correction of regional estimation. This approach is more similar to the one developed in this chapter because it relies on the same information transfer scheme; however, a different implementation procedure is proposed.

The AS procedure is developed and applied here to the same statistics defined in section 2.2.1, i.e. the index-flood, the L_{CV} and the L_{CA} that summarize the essential statistics for what concerns the estimation of flood quantiles. The AS model will then be available in addition the regional procedure discussed in chapter 2 to predict the same variables. Then, the results from the two different approaches can be combined in order to obtain more reliable final estimates in ungauged sites. In general, when two or more models are available for the same goal, one can consider one of the following scenarios:

- Model competition: the results of different models (in this work AS and regional prediction) can be evaluated separately and then compared, in order to identify which model is more efficient in the reconstruction of the variable of interest. In this case study, AS and regional predictions are expected to have a different reliability depending on the location of the target site, and, in particular, to its distance from the donor site. Under this perspective, the aim of the AS approach is to identify an alternative procedure that is more appropriate for the analysis at some ungauged basins.
- Model cooperation: the output of one model is used to initialize the other model. In this work, for instance, the regional estimate can be used as an additional parameter for the along-stream estimation function, and thus to contribute to the final AS prediction. This viewpoint can be interpreted as follows: the AS approach can be used to correct locally the regional model estimate accounting for the specific information present in a close donor site.



- Model combination: given different estimates of the same variable one combines them through suitable relations aiming at minimizing the variance of the resulting considered estimator.

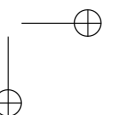
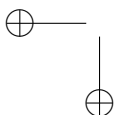
Of course, also other kinds of models, different from the regional one (e.g. the rational formula), can be adopted to compete/cooperate or to be combined with the AS model.

3.2. Extension of the regional procedure

3.2.1. Regional covariance and correlation

The regional procedure developed in chapter 2 for L -moments estimation in ungauged basins, can be profitably used in the along-stream model, as already suggested, to improve the reliability of the final estimates. For this purpose, covariances and correlations of concurrent regional predictions are particularly important, as the along-stream modelling approach investigates hydrological variables at close locations. This is also important when the final aim is to map an hydrological variable along the river network.

Despite this, the literature on regression models usually treats the prediction of one variable at a time, while correlation between concurrent estimates are rarely investigated [e.g. Hahn, 1972, Seber and Wild, 1989]. The covariance between two concurrent regression predictions \hat{y}_1 and \hat{y}_2 , obtained from two different sets of descriptors (\mathbf{x}_1 and \mathbf{x}_2) and the same set of regression coefficients $\hat{\beta}$, is defined as



$$\begin{aligned}
\text{cov} [\hat{y}_1, \hat{y}_2] &= \text{E} [(\hat{y}_1 - y_1)(\hat{y}_2 - y_2)^T] \\
&= \text{E} [(\hat{y}_1 - \text{E}[\hat{y}_1])(\hat{y}_2 - \text{E}[\hat{y}_2])^T] \\
&= \text{E} \left[\left(\mathbf{x}_1 \hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_1 - \text{E}[\mathbf{x}_1 \hat{\boldsymbol{\beta}}] \right) \left(\mathbf{x}_2 \hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_2 - \text{E}[\mathbf{x}_2 \hat{\boldsymbol{\beta}}] \right)^T \right] \\
&= \text{E} \left[\left(\mathbf{x}_1 (\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}]) + \boldsymbol{\varepsilon}_1 \right) \left(\mathbf{x}_2 (\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}]) + \boldsymbol{\varepsilon}_2 \right)^T \right] \\
&= \text{E} \left[\mathbf{x}_1 (\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}]) (\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}])^T \mathbf{x}_2^T \right] + \text{E} \left[\boldsymbol{\varepsilon}_1 (\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}])^T \mathbf{x}_2^T \right] \\
&\quad + \text{E} \left[\mathbf{x}_1 (\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}]) \boldsymbol{\varepsilon}_2^T \right] + \text{E} [\boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_2^T] \\
&= \mathbf{x}_1 \text{cov} [\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}] \mathbf{x}_2^T.
\end{aligned} \tag{3.1}$$

Since the error $\boldsymbol{\varepsilon}$ has zero mean, the two mixed terms in the form $\mathbf{x}(\hat{\boldsymbol{\beta}} - \text{E}[\hat{\boldsymbol{\beta}}])\boldsymbol{\varepsilon}$ are null, as well as the expectation of the two concurrent errors $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ that are mutually independent.

In this context, we consider the regional model with an error structure defined as in Stedinger and Tasker [1985], and the covariance matrix of the regression coefficients $\hat{\boldsymbol{\beta}}$ is provided by equation (2.23). This term contains the information about the structure of the model, and its substitution into equation (3.1) leads to

$$\sigma_{\hat{y}_1 \hat{y}_2} = \text{cov} [\hat{y}_1, \hat{y}_2] = \mathbf{x}_1 \left(\mathbf{X}^T \hat{\boldsymbol{\Lambda}}^{-1} \mathbf{X} \right)^{-1} \mathbf{x}_2^T, \tag{3.2}$$

where \mathbf{X} is the descriptors matrix of the calibration set, and $\hat{\boldsymbol{\Lambda}}$ describes the error structure of the regional model (i.e. the variance and covariances of the samples estimates and the model error, see equation (2.19)). The correlation coefficient ρ between \hat{y}_1 and \hat{y}_2 is determined in the standard way as the ratio between the covariance of a pair of predictions and the product of their standard deviations, i.e.

$$\rho_{1,2} = \frac{\text{cov} [\hat{y}_1, \hat{y}_2]}{\sqrt{\text{var} [\hat{y}_1] \text{var} [\hat{y}_2]}} \tag{3.3}$$

where $\text{var} [\hat{y}]$, calculated according to equation (2.25), has the form

$$\sigma_{\hat{y}}^2 = \text{var} [\hat{y}] = \hat{\sigma}_{\delta}^2 + \mathbf{x} \left(\mathbf{X}^T \hat{\mathbf{\Lambda}}^{-1} \mathbf{X} \right)^{-1} \mathbf{x}^T.$$

3.2.2. Formulae for log-transformed data

When the variable of interest used in the regression is preliminarily transformed, as is usual with the index-flood, covariance and correlation of concurrent estimation computed with equations (3.2) and (3.3) need to be back-transformed to the original space. In case of logarithmic transformation, the variable \hat{y} predicted by the (linear) regression model is normally distributed, so the related back-transformed variable μ becomes lognormally distributed. Two concurrent predictions, correlated through the coefficient of equation (3.3), can be described through a bivariate normal distribution

$$\{\hat{y}_1, \hat{y}_2\} \sim \mathcal{N} (y_1, y_2, \sigma_{y_1}^2, \sigma_{y_2}^2, \rho_y)$$

with five parameters. Analogously to the one-dimensional case, the parameters of the bivariate normal distribution are back-transformed, generating a bivariate lognormal distribution

$$\{\hat{\mu}_1, \hat{\mu}_2\} \sim \text{log}\mathcal{N} (\mu_1, \mu_2, \sigma_{\mu_1}^2, \sigma_{\mu_2}^2, \rho_{\mu})$$

where the two means and the two variances are calculated according to equation (2.27) and (2.28). The correlation is deduced through the joint covariance [e.g. Johnson and Kotz, 1986]

$$\text{cov} [\hat{\mu}_1, \hat{\mu}_2] = (\exp [\text{cov} [\hat{y}_1, \hat{y}_2]] - 1) \cdot \exp \left[\hat{y}_1 + \hat{y}_2 + \frac{1}{2} (\text{var} [\hat{y}_1] + \text{var} [\hat{y}_2]) \right], \quad (3.4)$$

and is calculated as

$$\rho_{\mu} = \text{cor} [\hat{\mu}_1, \hat{\mu}_2] = \frac{(\exp [\text{cov} [\hat{y}_1, \hat{y}_2]] - 1)}{\sqrt{(\exp [\sigma_{y_1}^2] - 1) (\exp [\sigma_{y_2}^2] - 1)}}. \quad (3.5)$$

An example of the correlation between concurrent estimation is shown in figure 3.1, where the correlation coefficient of Q_{ind} , L_{CV} and L_{CA} between the regional estimate at Tavagnasco station (code 15) and all the points of the upstream drainage network is mapped. Regional models adopted for this application are referred to in table 2.I, as $\ln Q_{ind2}$, L_{CV2} and L_{CA2} . As one can see from these maps, locations very close to the reference station of Tavagnasco have correlation coefficients of the order of 0.10-0.15 for Q_{ind} and L_{CV} , and a about 0.3 for L_{CA} . This apparent contradiction can be explained considering equation (3.3) in which the numerator, the covariance of concurrent predictions, does not contain the model variance term σ_δ that is, instead, present in the variance of a single prediction. In the regional models developed in this work, the model variance dominates the total prediction variance (about 90%, from table 2.II); thus, even when the covariance is maximized the correlation does not exceed the value of 0.10-0.15 (for the index-flood).

3.3. Along-Stream information propagation method

3.3.1. Methods and hypotheses

After setting the problem by the viewpoint of space correlation of the regional estimates, we now discuss the basic hypotheses and the methodologies developed in the along-stream approach, that will be used to estimate a generic hydrological variable P by propagating the information from a donor site d to the target site t . This approach is based on a few preliminary hypotheses. 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

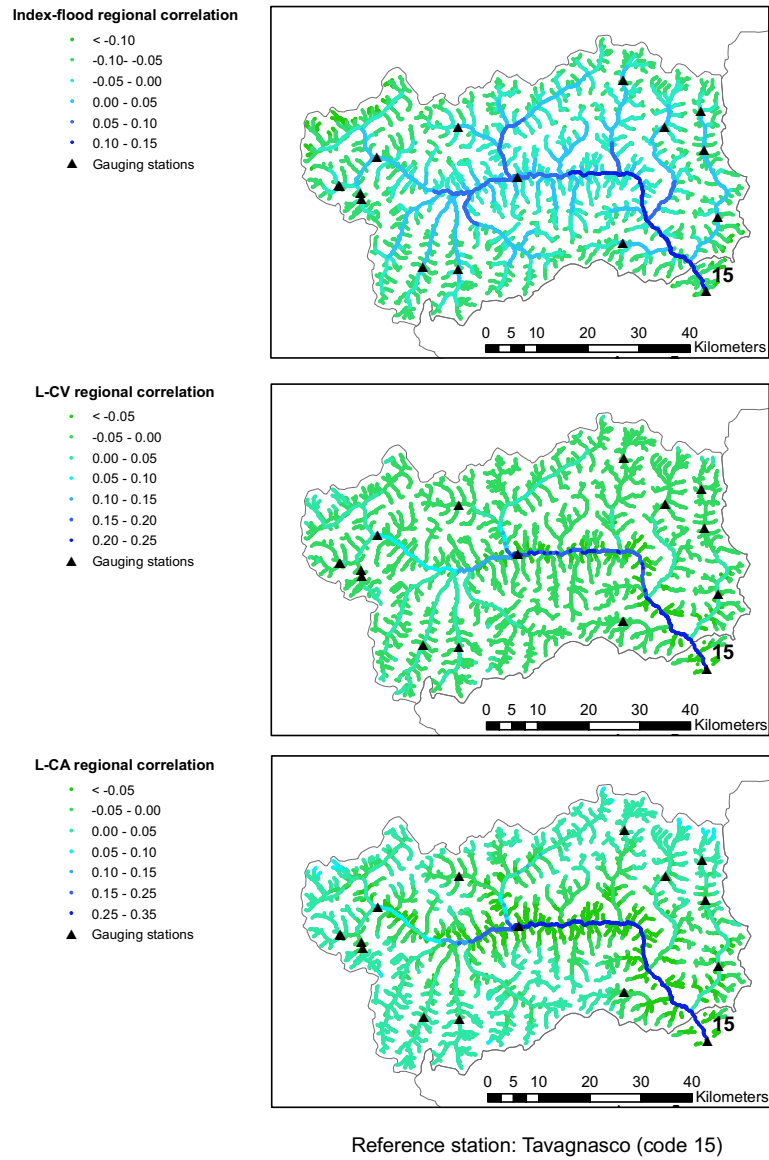


Figure 3.1.: Map of the correlation coefficient between the regional prediction at Tavagnasco (station code 15) and the regional prediction at each of the upstream-points of the drainage network, for Q_{ind} , L_{CV} and L_{CA} .

used in the information transfer, i.e.

$$P_t = f(S_d, \theta)$$

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

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

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

The proximity and transferability hypotheses are represented through the sketch in figure 3.2 (panel a) where the arrows show possible directions for the information transfer. In general, the function used to transfer information is not known, but can be approximated by any function that satisfies the hypotheses made. This function must be a good approximation of 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 panel b in figure 3.2). The validity domain hypothesis is very important to assess the reliability of the AS method, and will be treated in an intuitive way. 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 is not necessarily the geographic distance or the length of the drainage path. Moreover, given a particular function for the information transfer, and its corresponding domain of validity, the variance of the AS prediction is supposed to increase moving away from the donor site, but still within the validity domain. Out of

that, no AS predictions are reliable, so it is no longer necessary to compute their variance. A sketch representing this aspect is shown in figure 3.2 panel (c).

In this work, the application of the along-stream modelling approach involves also the regional model, and can be thought as an approach based on both the ideas of cooperation and competition with the regional model. In particular, the regional model tries to catch the “global” variability of hydrological variables, without considering the “local” structure of the river that can be accounted, on the other hand, by the AS model. The AS estimates is then calculated on the basis of the regional ones (cooperation of models). In contrast, reliability of the AS predictions decrease with increasing distances between the donor and the target basins; the problem is thus to identify a procedure that allows one to decide if the AS estimates can be considered reliable or if the regional one should be preferred (competition between models). In this application, the regional models considered for the L -moments estimations, previously developed in chapter 2, are summarized in tables 2.I, 2.IV and 2.II, where are referred as to $\ln Q_{ind2}$, $LCV2$ and $LCA2$ for index-flood, L_{CV} and L_{CA} respectively.

The first step to implement the along-stream estimation procedure is to define a suitable formula to compute the variable P at the target site t , according to all the hypotheses made. Here we adopt the formula used by Kjeldsen and Jones [2007], although the methods hereafter developed follow a different approach. Let \mathcal{T} be the function used for the along-stream information transfer, that reads

$$\mathcal{T}_{t,d} = \frac{R_t}{R_d} \cdot S_d. \quad (3.6)$$

where the symbol R refers to the regional estimates and S is the at-site variable. The propagated estimate can be simply written as:

$$P_t = [\mathcal{T}_{t,d}]_{D \leq D'} \cdot \quad (3.7)$$

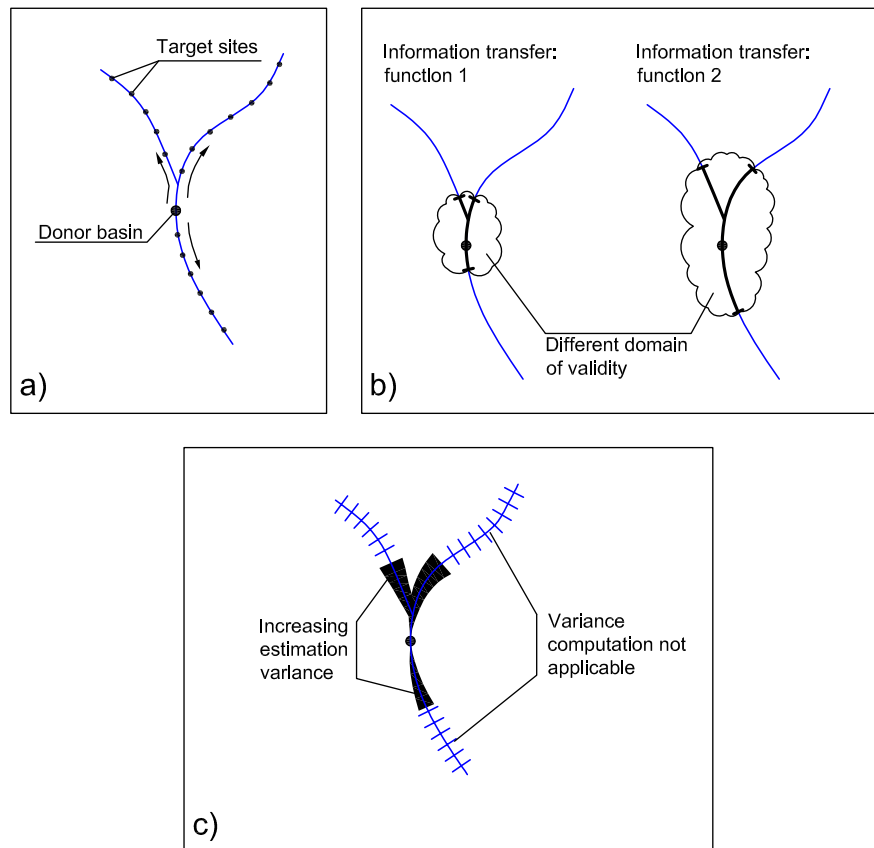


Figure 3.2.: A sketch of the along-stream propagation of information method. An hydrological variable calculated at the donor (gauged) site is used to predict the value of the same variable in the target locations located upstream or downstream (panel (a)). Different functions to achieve this aim can be adopted (panel (b)); however, each function has a particular domain of validity around the donor station. The variance of the new predictions is supposed to increase moving away from the donor station, within the domain of validity (panel (c)). This is no longer applicable out of the validity domain.

where D is the generalized distance between t and d and D' is the threshold distance beyond which the function is no longer effective. The symbol $[\cdot]_{D \leq D'}$ underlines the fact that the transfer formula is valid only within its validity domain. Note that all the symbols P , R and S represent a generic hydrological variable (index-flood, L_{CV} and L_{CA} in this particular context). Equation (3.6) can be interpreted as follows: the regional estimate R_t in t is corrected by a factor equal to the relative error that the regional model produces in d (i.e. S_d/R_d). In practice, the regional model is supposed to have the same error magnitude in evaluating two close locations. For $D \rightarrow 0$ it is straightforward to verify that $P_t \rightarrow S_d$.

3.3.2. Example about functions and assumptions

An example to intuitively describe the assumptions made about the domain of validity of a function is now reported. Assume that there are two different functions available for the information transfer along the stream network (similarly to the representation in figure 3.2, panel b). We define the first function as

$$P_t^{(1)} = [S_d]_{D \leq D'^{(1)}}, \quad (3.8)$$

where D is the generalized distance between t and d and $D'^{(1)}$ is the threshold distance beyond which function 1 is no longer effective. The second function is, as in equation (3.7):

$$P_t^{(2)} = [\mathcal{T}_{t,d}]_{D \leq D'^{(2)}}. \quad (3.9)$$

The first function simply states that the propagated prediction $P_t^{(1)}$ is equal to the at-site variable calculated in d . Obviously, equation (3.8) can be considered valid only in a very small neighborhood of d , i.e. the threshold $D'^{(1)}$ is supposed to be very low, and thus $D'^{(1)} \leq D'^{(2)}$.

Depending on the distance D there are three different possibilities:

- $D \leq D'^{(1)} \leq D'^{(2)}$: both the AS models are valid, the most appropriate

can be selected on the basis of the prediction variance;

- $D^{(1)} \leq D \leq D^{(2)}$: only model 2 can be used to propagate the information along the stream network;
- $D > D^{(2)}$ neither model can be used.

3.3.3. Organization of nested basins

In the next sections, we will investigate the suitability of the AS approach considering, as a case study, a set of 70 basins located in northwestern Italy; the complete database is mainly constituted by the catchments already used for the regional analysis of chapter 2; however the data are organized in a different way. In this context, in fact, it is more appropriate to work in terms of pairs of basins $\{t, d\}$, rather than with a single catchment at a time. Figure 3.3 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, as well as basins with no connections. All the connected (nested) catchments have been considered as a possible pair of donor-target site, characterized by a generalized distance $d_{t,d}$ among them.

Considering all the possible connections of two stations along the same drainage path (nested basins), there are a total of 71 connections (e.g.: from figure 3.3, basins 1 is nested to basin 15 even if there is the intermediate basin 13). All the basins under analysis are actually gauged basins; however, the connections are considered “in both directions”, e.g.: if basin 9 is upstream basin 10, we first consider basin 9 as the donor site and basin 10 as the target (ungauged) site, then the procedure is repeated using basin 10 as the donor station and basin 9 as the target (ungauged) site. In this way, the overall number of usable connections $\{t, d\}$ becomes 142.

The distance between two catchments can be defined in different ways, although it is preferable to avoid both the geographical distance and the

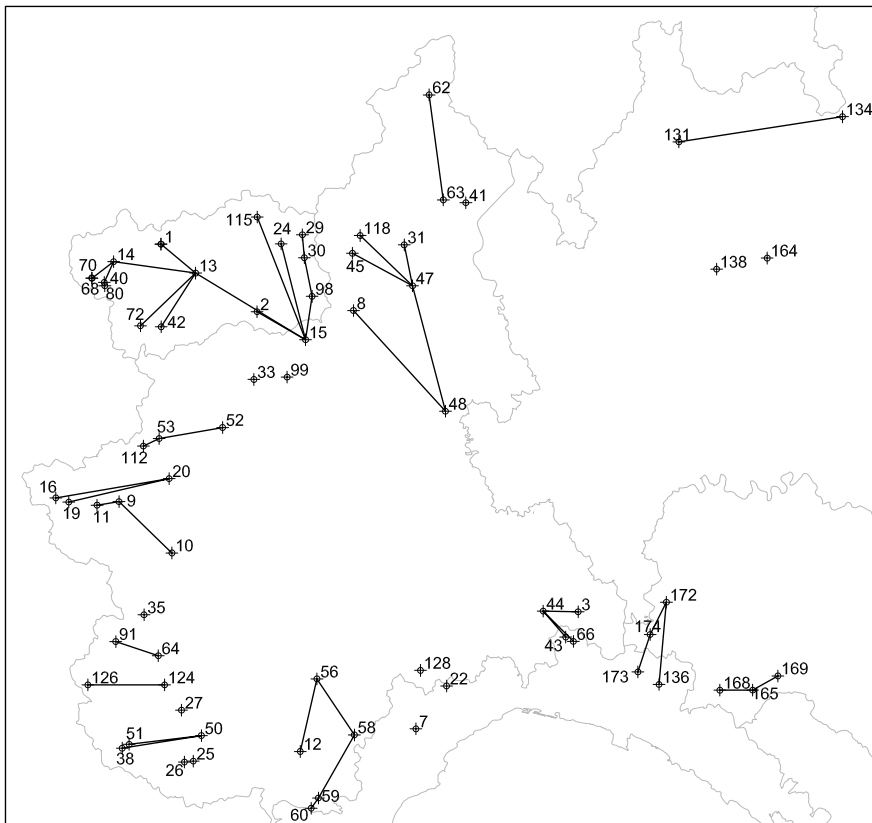


Figure 3.3.: Gauging stations lying on the same drainage path, either upstream or downstream, that are directly connected (nested basins) are schematically linked with a line. Some catchments have multiple connections, others are isolated.

length of the drainage path linking the two points. In fact these definitions are not representative of the abrupt change in basin characteristics that is expected between two points located just upstream and just downstream a tributary. We propose a definition of distance based on the basin area A ,

$$D = \log(A_{\max}/A_{\min}) \quad (3.10)$$

with $A_{\max} = \max[A_t, A_d]$ and $A_{\min} = \min[A_t, A_d]$. Under the proximity hypothesis (but not in general), two basins with the same area have null distance (they are the same basin), so their estimates must coincide (congruence hypothesis). An alternative simple definition of distance, that involves also the basin mean elevation H , reads

$$D = \log(A_{\max}/A_{\min} \cdot H_{\max}/H_{\min}). \quad (3.11)$$

This definition is supposed to work well when the mean basin elevation and the basin area do not contain redundant information, for instance when the dataset is composed of basin from both mountainous and plain areas. In our case study we use the definition given in equation (3.10).

3.4. Model reliability: simplified approach

3.4.1. Uncertainty of the propagated estimate

The basics of the AS method can be summarized in two steps: (i) choice of a suitable formula for the information transfer, and (ii) definition of the threshold distance D' that is strictly related to the formula adopted in (i). In section 3.3, a practical formula (equation (3.7)) is adopted for this case study, without providing a quantitative assessment of D' . This section investigates the suitability of a simplified approach for D' quantification and an overall evaluation of the performance of the along-stream estimation approach.

The AS procedure, applied to the index-flood by means of equation (3.7), and based on the 142 pairs of catchments, yields the P predictions shown

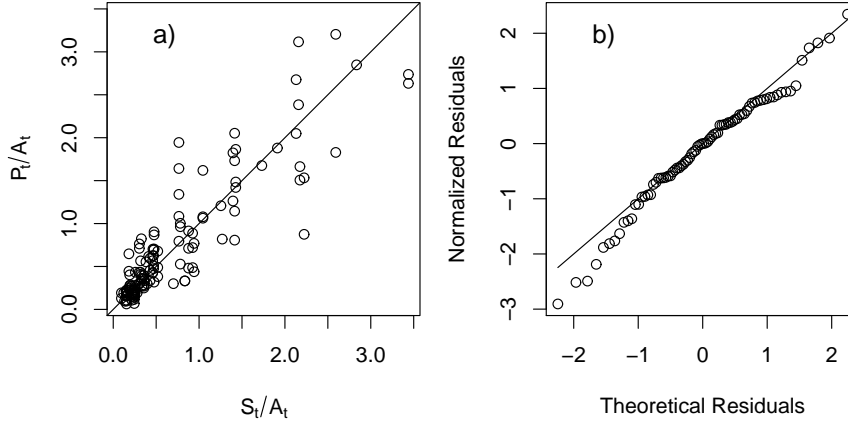


Figure 3.4.: Comparison between local and sample estimates relative to the whole available dataset (panel a). Multiple P_t values relative to the same S_t represents multiple sources for local estimation. Panel b shows the normality plot for the residuals $P_t - S_t$. Values are normalized by the catchment area in order to provide a clearer representation.

in figure 3.4, panel a, compared with the correspondent at-site estimates. Basins with multiple connections can be easily identified because there are multiple P_t estimates corresponding to the same S_t value. Panel b of the same figure reports the normality plot relative to the residuals $P_t - S_t$, which will be used afterward to evaluate the variance of P_t .

At this stage, it is necessary to define a suitable formula to calculate the uncertainty of P_t . In the simplified approach, a model for P_t uncertainty is

$$CV_{P_t} = (1 + \alpha \cdot D) \cdot CV_{S_d} \quad (3.12)$$

where CV is the coefficient of variation, i.e. the ratio between standard deviation and mean of the variable. Considering the definition of P_t given in equation (3.7), and the definition of CV as the ration between standard deviation and mean, we obtain

$$\sigma_{P_t} \cdot \frac{R_d}{R_t \cdot S_d} = (1 + \alpha \cdot D) \cdot \frac{\sigma_{S_d}}{S_d} \quad (3.13)$$

and thus

$$\sigma_{P_t} = (1 + \alpha \cdot D) \cdot \sigma_{S_d} \cdot \frac{R_t}{R_d}. \quad (3.14)$$

This model for predicting σ_{P_t} can be interpreted as follows: the standard deviation of P_t is basically the standard deviation of the at-site estimate in the gauged site, augmented proportionally to a factor α that accounts for both the non-correctness of the AS transfer function and for the uncertainty of all the variables involved in equation (3.7). Furthermore, for $D \rightarrow 0$ it is straightforward to verify that $\sigma_{P_t} \rightarrow \sigma_{S_d}$, confirming the congruence hypothesis.

3.4.2. Assessment of the variance parameter

The evaluation of the uncertainty of the AS estimate using equation (3.14) requires to preliminarily estimate the parameter α , calibrated on the basis of the available dataset rearranged to account for the donor-target correspondences. For each pair of basins, the residual between P_t and its corresponding at-site value S_t is

$$\delta_t = P_t - S_t \quad (3.15)$$

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

$$\delta_t \sim \mathcal{N}(0, \sigma_{P_t}^2 + \sigma_{S_t}^2). \quad (3.16)$$

Substituting equation (3.14) in equation (3.16), we obtain the final expression for the residual variance, parameterized by α , that reads

$$\sigma_{\delta}^2 = (1 + \alpha \cdot d)^2 \cdot \sigma_{S_d}^2 \cdot \left(\frac{R_t}{R_d}\right)^2 + \sigma_{S_t}^2. \quad (3.17)$$

The coefficient α can be estimated by means of a maximum likelihood approach. The likelihood function \mathcal{L} of the residuals, that are supposed to follow a normal distribution with equation (3.16), is

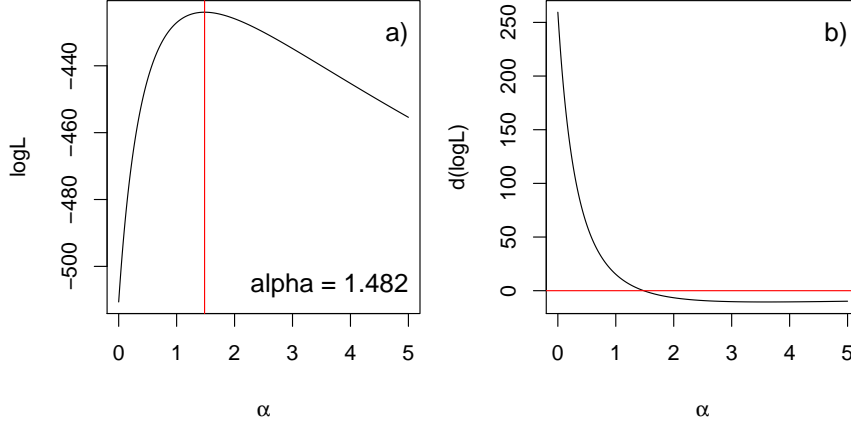


Figure 3.5.: The maximum likelihood estimator of α is identified on the log-likelihood plot (panel (a)) considering only the subset of basin pairs with $D \leq \log(10)$. Panel (b) shows where the derivative of the log-likelihood function equals zero.

$$\mathcal{L}(\delta) = \prod \frac{1}{\sqrt{2\pi\sigma_\delta^2}} \exp \left[-\frac{1}{2} \left(\frac{\delta - \mu_\delta}{\sigma_\delta} \right)^2 \right] \quad (3.18)$$

that can be handled more easily after a logarithmic transformation:

$$\log \mathcal{L}(\delta) = -\frac{1}{2} \sum \left[2\pi\sigma_\delta^2 + \frac{\delta^2}{2\sigma_\delta^2} \right] \quad (3.19)$$

The maximum likelihood estimator of α can be numerically computed by a maximization of equation (3.19) or putting equal to zero its first derivative

$$\frac{d \log \mathcal{L}(\delta_{t,d})}{d\alpha} = \frac{\delta^2 \sigma_{S_d}^2 (R_t/R_d)^2 (1 + \alpha D) D}{\left[\delta^2 \sigma_{S_d}^2 (R_t/R_d)^2 (1 + \alpha D)^2 D + \sigma_{S_t}^2 \right]^2} - \frac{\sigma_{S_d}^2 (R_t/R_d)^2 (1 + \alpha D) D}{\delta^2 \sigma_{S_d}^2 (R_t/R_d)^2 (1 + \alpha D)^2 D + \sigma_{S_t}^2}. \quad (3.20)$$

An example is reported in figure 3.5, that shows the log-likelihood function and its first derivative. A more detailed discussion about the choice of this value is provided in the following sections.

3.4.3. Validity of the simplified approach

As mentioned above, the aim of the whole procedure is to use both the regional and the AS approaches to improve the final prediction in ungauged sites. From a practical point of view, it is necessary to define the operational (O) prediction as the estimate obtained from either the AS or the regional procedure, by selecting the appropriate model with the following rules:

	$\sigma_{P_t} \leq \sigma_{R_t}$	$\sigma_{P_t} > \sigma_{R_t}$
$D \leq D'$	along-stream	regional
$D > D'$	regional	regional

The correct value of D' is not known a priori, but can be evaluated through an iterative procedure, based on the following steps:

- a tentative value of D' is empirically defined;
- the AS estimate P_t is evaluated, as well as the regional one, R_t ;
- the residuals of the AS estimates are computed and the parameter α is evaluated in the max-likelihood framework applied only to the basins pairs within D' ;
- based on α , the variance of the AS prediction is computed with equation (3.14) and it is compared against the variance of the regional prediction at the same location;
- the operational estimate is constructed choosing the model with the lower uncertainty;
- the errors obtained by the operational estimate are compared with those of the regional model that is considered as the reference model;
- the procedure is repeated, changing the tentative value of D' .

The iterative procedure has been applied to the index-flood predictions using, as distance measure between basins, equation (3.10). The main results

are summarized in table 3.I where a series of tentative threshold distances are used; the value of α , computed only on the restricted dataset, is also reported. The light-grey column of the same table is relative to a threshold distance that includes all basins, and therefore can be used as a reference condition, since it correspond to the case of unbounded validity domain. Differently, the heavy-grey column refers to the optimal distance threshold identified for this case study, and relative to the index-flood.

Table 3.I shows also the percentage of basin pairs for which the AS model is applicable ($D \leq D'$) and the errors related to the operational (O), along-stream (AS) and regional (R) models. The mean error, named ME in the table, is computed averaging the errors

$$E_{\{t,d\}} = \frac{(\text{prediction})_d - S_t}{\sigma_{S_t}} \quad (3.21)$$

obtained for each pair $\{t, d\}$, where “prediction” indicates one of the three possible models. Some important remarks can be deduced from the errors reported in table 3.I. Firstly, the AS-ME presents a clear trend, increasing with an increasing threshold distance. This is an evidence that the use of a restricted domain of validity improves the effectiveness of the along-stream prediction, reducing the averaged error. Furthermore, the application of the AS estimation to the whole dataset produces a mean error greater than those relative to the regional model: in this case the AS model is no longer appropriate. Instead, the error R-ME produced by regional models is, as expected, less influenced by the choice of a particular threshold distance. A further consideration arises looking at the operational error O-ME that is always lower than both AS-ME and R-ME; this is a confirmation that the operational model is able to correctly select the AS model rather than the regional one. Finally, the use of a low threshold distance lead to better results, but, in this way, the AS approach is limited to a few basins. For instance, very good performances can be achieved with $D' = 0.81$, but, in this case, only the 11.3% of the basins can benefit of the along-stream model.

On the other hand, large domains of validity increase the errors and decrease the effectiveness of the operational estimator. For these reasons, we selected an “optimal” distance as a compromise between these two opposite factors. Such a threshold distance correspond to basins that have a ratio between their areas of 10; i.e. for pairs of basins whose areas differ of at most one order of magnitude, the AS approach is appropriate when working with the index flood.

The effect of the optimal threshold distance is shown in figure 3.6 (panel (a)) where the normalized errors of the operational model are compared against those of the regional (reference) model. The points on the graph can be divided in four different classes:

- filled-circles on the bisector represents the basins out of the validity domain, where only the regional model is applicable;
- empty-circles on the bisector are basins within D' for which the regional model have been selected as operational model;
- empty-circles below the bisector are basins within D' for which the AS models have been selected and the operational estimates improve the regional one;
- empty-circles above the bisector are basins within D' for which the AS models have been selected and the operational errors are greater than the regional ones.

It is evident that, in most of the points in which the AS approach is suitable to be used, the operational estimates are better than the corresponding regional ones. Only for few basins there is, although moderately, an increase in the mean operational error.

These results are positive when compared to those of panel (b) of the same figure, where no threshold distance has been applied. Although the mean operational error still suggests to use the AS model, the points dispersion

Table 3.1.: Summary of the results obtained with the operational (O), the along-stream (AS) and the regional (R) predictions, for different distance thresholds. ME indicates the mean error calculated as in equation (3.21). The light-grey column is relative to a threshold distance that include all the basins, that can be used as a reference condition with unbounded validity domain. The heavy-grey column shows the optimal D' value adopted for this case study. α is computed only on the basins pairs with $D \leq D'$.

	Tentative										Optimal
D'	0.81	1.12	1.39	1.83	2.15	2.55	3.11	3.6	4.18	5.03	2.3
α	2.28	1.12	1.17	1.15	1.66	1.51	1.55	1.59	1.68	1.54	1.48
[% basins] $_{D \leq D'}$	11.3	21.1	31	40.9	50.7	60.6	70.4	78.9	90.1	100	56.3
[O-ME] $_{D \leq D'}$	1.72	1.69	1.82	1.98	2.52	2.47	2.68	2.82	3.1	3.09	2.47
[AS-ME] $_{D \leq D'}$	1.77	1.79	2.35	2.45	2.95	2.97	3.21	3.45	3.86	3.8	2.84
[R-ME] $_{D \leq D'}$	2.35	2.02	2.16	2.32	2.88	2.96	3.02	3.07	3.18	3.23	3.02

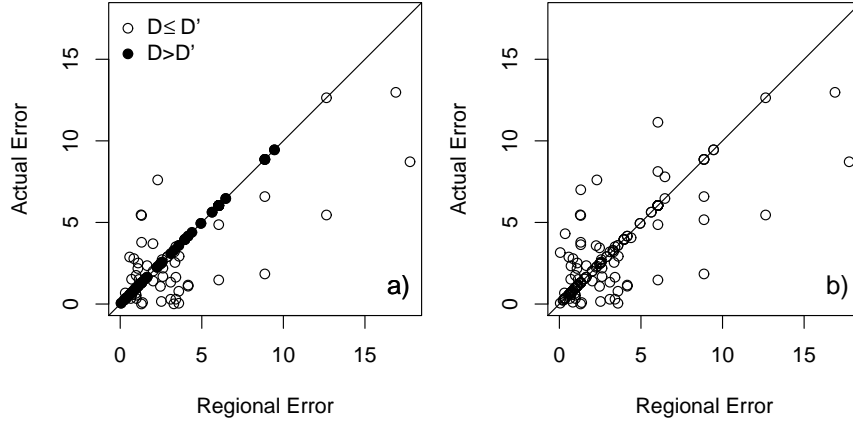


Figure 3.6.: Errors due to regional estimates of Q_{ind} compared with the errors produced by operational model. The figure shows, with open circles, the errors obtained for basins pairs closer than the threshold distance while filled circles are relative to the more distant catchments (whose estimated is fixed equal to the regional one). All the points below the solid line represents basins where the Q_{ind} estimates are improved by the use of the along-stream information transfer procedure. Panel a is relative to a threshold distance $D' = \log(10)$, while b represents the same but with no limitation on distances.

highlights the fact that the variance of the AS prediction is no longer appropriate to describe the reliability of the AS model. This is again to say that, for basins beyond the threshold distance the regional model is the most appropriate.

The same procedure has been applied to the L_{CV} and L_{CA} without reaching appreciable results. Figure 3.7 and figure 3.8 clearly show that the AS model does not produce reliable results and, when applicable, produces a deterioration of the regional estimates. These results can be interpreted as the inability of a simple AS model (as equation (3.7)) to catch the behavior of the L_{CV} or L_{CA} . However, this problem is also due to the high uncertainty of the higher-order L -moments that are based on short data records. This uncertainty makes impossible a correct estimation of the parameter α , and thus the bounds of the validity domain.

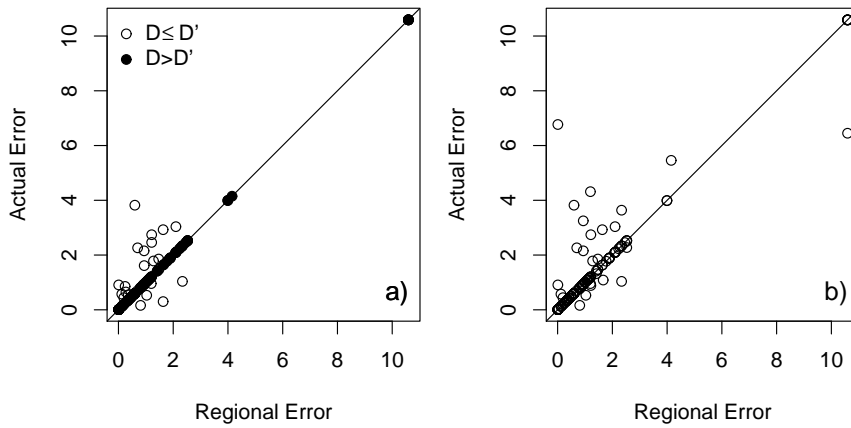


Figure 3.7.: Operational versus regional errors for L_{CV} with optimal threshold distance (panel (a)) and without threshold distance (panel (b)).

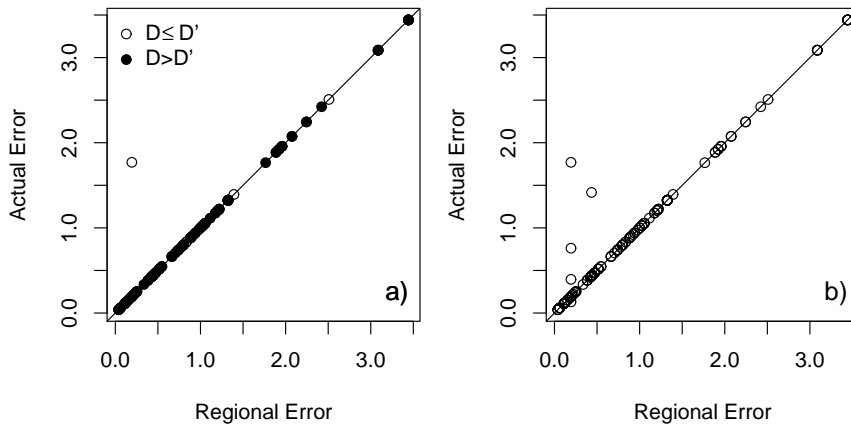


Figure 3.8.: Operational versus regional errors for L_{CA} with optimal threshold distance (panel (a)) and without threshold distance (panel (b)).

3.5. Model reliability: analytical approach

The simplified approach to evaluate the reliability of the along-stream model considers a simple formulation for the quantification of the variance of P_t . This expression contains a parameter to be estimated from the available data that accounts for both the sampling and the structural errors present in the model. In general, however, the presence of faraway couples of donor-target sites and/or high uncertainties in sample estimates, can prevent a proper estimate of this parameter. To try to overcome this inconvenience, and allow a simple estimation of the AS prediction variance for short distances, we develop a more detailed analytical framework.

The true value of the hydrological variable P_T at a point is not known, but can be represented as the sum of a deterministic function \mathcal{T} and a stochastic error ε ,

$$P_T = \mathcal{T} + \varepsilon, \quad (3.22)$$

where ε is supposed to be normally distributed with zero mean and variance σ_ε^2 and \mathcal{T} is a derivable function. In this analysis we use the same function for the along-stream information propagation \mathcal{T} proposed in section 3.4 (equation (3.6)).

Since ε has zero mean, it does not influence the AS prediction, that still reads $P_t = [\mathcal{T}_{t,d}]_{D \leq D'}$, while its variance can be written in the form

$$\sigma_P^2 = \text{var}[P_T] = \text{var}[\mathcal{T}] + \text{var}[\varepsilon], \quad (3.23)$$

where \mathcal{T} and ε are supposed to be independent, at least for $D \leq D'$.

The variance of \mathcal{T} can be calculated, for instance, using the propagation of variance of a known function. With \mathcal{T} defined as in equation (3.6) we obtain

$$\begin{aligned} \text{var}[\mathcal{T}] &= (\mathcal{T}'_{R_t})^2 \cdot \text{var}[R_t] + (\mathcal{T}'_{R_d})^2 \cdot \text{var}[R_d] + (\mathcal{T}'_{S_d})^2 \cdot \text{var}[S_d] \\ &\quad + 2(\mathcal{T}'_{R_t} \mathcal{T}'_{R_d}) \cdot \text{cov}[R_t, R_d] + 2(\mathcal{T}'_{R_t} \mathcal{T}'_{S_d}) \cdot \text{cov}[R_t, S_d] \\ &\quad + 2(\mathcal{T}'_{R_d} \mathcal{T}'_{S_d}) \cdot \text{cov}[R_d, S_d], \end{aligned} \quad (3.24)$$

where the symbol \mathcal{T}'_x denotes the first derivative of the \mathcal{T} function with respect to x . The covariance terms involving the regional and the empirical estimators together are null because R and S are independent. Differently, the remaining covariance term, $\text{cov}[R_i, R_j]$, involves two concurrent regional estimates that are correlated because R_i and R_j stem from the same regression model. This value, discussed in detail in section 3.2, can be calculated by equation (3.2) for native linear regression models, or by equation (3.4) for log-linearized regressions. In this particular case, since $\mathcal{T}'_{R_i} = \frac{S_j}{R_j}$, $\mathcal{T}'_{R_j} = -\frac{R_i \cdot S_j}{R_j^2}$ and $\mathcal{T}'_{S_j} = \frac{R_i}{R_j}$, equation (3.24) reduces to the following expression:

$$\begin{aligned} \text{var}[\mathcal{T}] &= \left(\frac{S_j}{R_j}\right)^2 \cdot \text{var}[R_i] + \left(-\frac{R_i \cdot S_j}{R_j^2}\right)^2 \cdot \text{var}[R_j] \\ &+ \left(\frac{R_i}{R_j}\right)^2 \cdot \text{var}[S_j] - 2 \left(\frac{R_i \cdot S_j^2}{R_j^3}\right) \cdot \text{cov}[R_i, R_j]. \end{aligned} \quad (3.25)$$

The effect of mutual correlation between concurrent regional estimates is reported, for the set of catchments under study and for the index-flood, in figure 3.9 where the correlation $\text{cor}[R_t, R_d]$ is plotted against the distance between the target and the donor basin. As already discussed in section 3.2, the correlation coefficients is intrinsically limited to a value around 0.10 also for short distances, due to the high model error present in the regional model. Correlation becomes negative for distance between basins of about 4, i.e. for $A_{\max}/A_{\min} \sim 50$.

The analytical framework proposed can be used in two ways:

- working within small distances, so that the residual variance σ_ε^2 can be neglected;
- parameterize σ_ε^2 as a function of D , then assess the reliability of the AS approach analogously to the simplified approach.

Preliminary analyses have shown that the latter approach cannot be easily solved, then a more detailed investigation is left to future works.

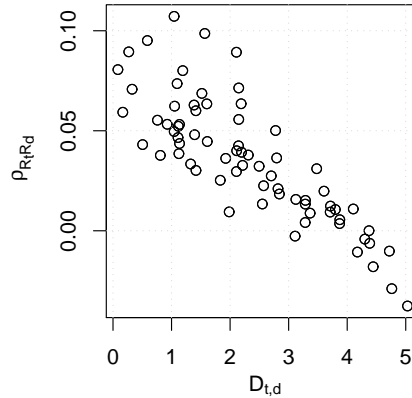


Figure 3.9.: Correlation coefficients of two simultaneous prediction of the index-flood by the regional model at sites i and j versus the distance between catchments. The evident convergence around a maximum value of about 0.10 for null distances is due to the large effect of the model error on the total prediction error.

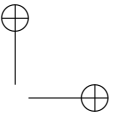
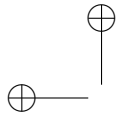
3.6. Final remarks

The along-stream (AS) estimation approach proposed in this chapter hinges on the river structure to perform the information transfer towards ungauged sites. This is rather different from regional procedures because it is based on local relationships instead of global ones, as the estimation can be done only between nested catchments. This conceptual difference between along-stream and regional models allowed us to combine the two methods and develop a general framework for the evaluation of both an hydrological variable and its variance at ungauged locations. The framework can be also extended to other kinds of models, like physically-based or conceptual ones.

The combined use of regional and along-stream procedure has been studied through the definition of a simple formula to locally correct the regional estimates on the basis of neighboring gauged stations. After that, the uncertainty of the propagated variable and the reliability of the method has been analyzed following two different paths: a simplified and an analytic approach.

The simplified approach, based on a parametrization of the prediction variance, proved a good ability in estimating the index flood when donor and target site are within a generalized distance threshold. In particular, the distance is computed as a function of the two basin areas, and the along-stream model achieves the best results when the ratio between their areas is not greater than 10. In these cases, the AS predictions improves significantly the regional values. The same procedure applied to L_{CV} and L_{CA} does not yield as good results probably due to the greater uncertainty of the sample estimates of higher-order L -moments. Despite this, the improvement of the index-flood is still an important achievement.

The second approach is based on the assumption that the prediction variance is made up by two components: one due to the AS formula adopted, and the second one due to the model incorrectness and sampling errors. For closely nested catchments, the latter is negligible, and the variance of the estimate can be assessed. Differently, for distant basins, this term need to be parameterized and evaluated.



Chapter 4.

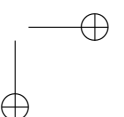
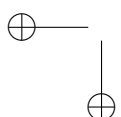
Distance-based regional approach for flow duration curves

Contents

4.1. Introduction	75
4.2. Distance-based method	77
4.2.1. (Dis)similarity between curves	78
4.2.2. Distance matrices, linear regression and Mantel test	81
4.2.3. Cluster analysis	84
4.3. Case study: distance-based method application	88
4.3.1. Hydrological and geomorphologic data	88
4.3.2. Procedure setting	90
4.3.3. Regions definition	92
4.4. Comparison with parametric models	93
4.5. Final remarks	99

4.1. Introduction

The problem of estimating hydrological variables in ungauged basins has been studied in the previous chapters for flood quantiles. In this chapter we deal with a different specific descriptor of the runoff distribution in a basin: the whole *Flow Duration Curve* (FDC). A flow duration curve represents the daily flow distribution approximation in a stream rearranged to show the percentage of time during which a discharge value is equalled or exceeded. Strictly speaking this is not a probability curve, because discharge is correlated between successive time intervals and discharge characteristics



are dependent on the season; hence the probability that discharge in a particular day exceeds a specified value depends on the discharge on preceding days and on the time of the year [Mosley and McKerchar, 1993, p. 8.27]. However, a flow duration curve is often interpreted as the complement of the cumulative distribution function of the daily streamflow values at a site. The FDC also provides a graphical summary of streamflow variability and is often used in hydrologic studies for hydropower, water supply, irrigation planning and design, and water quality management (a review on many applications is provided by Smakhtin [2001]).

The empirical FDC is constructed from observed streamflow time series. These observations can have different time-scale resolution, although mean daily streamflow values are commonly used. The data are ranked in descending order and each ordered value is associated with an exceedance probability F , for example through a plotting position formula. If the FDC is constructed on the basis of the whole available data set, merging together all available years of data, it represents the variability of flow over the entire observation period. This representation is valid when the dataset is sufficiently long. A different approach, introduced by Vogel and Fennessey [1994], is to consider annual FDCs separately, i.e., to consider a different FDC for each year when data are available [e.g., Claps and Fiorentino, 1997, Iacobellis, 2008]. A parametric model able to represent both the total and the annual FDCs for gauged and ungauged sites has been proposed, for instance, by Castellarin et al. [2004b, 2007].

In the present work only total FDCs will be considered, adopting a non-parametric approach for their representation. The FDCs are modelled following the index-value approach, in which the flow duration curve $Q(F)$ is the product of two terms $Q(F) = \mu \cdot q(F)$, where the *index flow* μ is the scale factor and the *dimensionless total flow duration curve* $q(F)$ represents the shape of the FDC. The present work focuses on the regionalization of the dimensionless curve, while the estimation of the index flow will not be

treated. In section 4.2 a distance-based method is described. This method is applied to a case study in section 4.3, where a set of basins located in North-Western Italy and Switzerland is investigated. The method's performances against alternative parametric methods are finally checked in section 4.4.

4.2. Distance-based method

Leaving aside the index-flow estimation, the regional FDC model developed here is based on concepts different from those of chapter 2. Here, the regional procedure still requires the creation of separate regions as in more traditional approaches; however, the curves are grouped according to their shape (dis)similarity. In standard approaches [e.g., Fennessey and Vogel, 1990, Singh et al., 2001, Holmes et al., 2002], this shape is represented in a parametric way. For instance, the coefficient of variation (CV) or the L -CV [Hosking and Wallis, 1997] of the curve can be used for this purpose. In this case, the selected parameter is related to basin descriptors through a linear or a more complex model. A regression analysis is performed with different combinations of descriptors, and those that are strongly related with the parameter are used for its estimation in ungauged sites.

The distance-based approach proposed here considers the dimensionless FDC as a whole, without resorting to statistical descriptors of its shape. This means that a curve is not fitted by an analytical function, which would imply a parametric representation of the FDC. Considering the curve as a whole object is particularly useful in those cases where the highly variability in the curves shapes lead to a difficult, or even unreliable, parameterization. This is the case, for example, of the research in the ecological field [e.g. Legendre and Legendre, 1998] as well as in the patterns recognition procedures [Pekalska and Duin, 2005], where the distance-based methods are widely used.

The multiregression approach can still be used to study the (dis)similarity between pairs of basins. The procedure is synthetically described below as a sequence of logical steps, while details are provided in the following

subsections:

1. for each couple of stations, a dissimilarity index between dimensionless curves is calculated using a predefined metric (section 4.2.1);
2. for each considered basin descriptor (e.g., area, mean elevation, mean slope, drainage path length, etc), the absolute value of the difference between its measure in two basins is used as the descriptor distance;
3. the distances between couples of FDCs (and between basin descriptors) are organized in distance matrices (section 4.2.2);
4. a multiregression approach is applied using the FDC distance matrix as the dependent variable, and the descriptor distance matrices as the independent variables; this serve to select the relevant basin descriptors (those associated to the best regression model) (section 4.2.2);
5. in the resulting descriptors' space, stations with similar descriptor values (small distances between descriptors) are grouped together into regions through a cluster analysis (section 4.2.3);
6. the regional dimensionless flow duration curve is estimated by taking the average of all the curves belonging to the cluster, as in the “graphical approaches” reviewed by Castellarin et al. [2004a] and references therein.

Critical points of this procedure, discussed more in detail in the following, are the choice of a suitable distance measure for the dimensionless flow duration curves, the identification of the best regression model between distance matrices, and the choice of the method of cluster analysis for the formation of the regions.

4.2.1. (Dis)similarity between curves

Let Q_s^* be the sequence of N_s daily discharges in the gauged station s , containing all the recorded values. Based on these data the scale factor μ_s is

first computed as the average of the whole sequence. Then, the dimensionless sequence $q_s^* = Q_s^*/\mu_s$ is rearranged in descending order and each value $q_{i,s}$, with $i = 1, 2, \dots, N_s$, is associated to its exceedance probability (i.e., through the Weibull plotting position)

$$\left\{ \frac{1}{N_s + 1}, \frac{2}{N_s + 1}, \dots, \frac{N_s}{N_s + 1} \right\}. \quad (4.1)$$

The distance-based procedure proposed here is based on the comparison between couples of curves: for this purpose it is convenient the two curves have the same number of elements. Since total FDCs have generally different lengths, depending on the number of years they cover, we resample them to make the curves comparable. For this purpose, we resample the FDCs at the frequency values

$$\left\{ \frac{1}{365 + 1}, \frac{2}{365 + 1}, \dots, \frac{365}{365 + 1} \right\}, \quad (4.2)$$

obtaining a new representation of the FDC in the station s :

$$\{q_{1,s}, q_{2,s}, \dots, q_{365,s}\}. \quad (4.3)$$

Other sampling rates can be used to better sample particular parts of the curves. In this work we have also considered an alternative sampling method that produces 365 equally spaced values in the z -space, where z is the normal reduced variate (with zero mean and unit variance). Back-transforming these values to the frequency space, the 365 values are no more equally spaced but more concentrated around higher and lower frequencies. Figure 4.1 sketches two curves with different number of elements resampled with a constant and a z spacing in the frequency axis.

In this approach a measure of similarity between curves (hereafter termed *distance*) is required. Given two FDCs, relative to two gauging stations

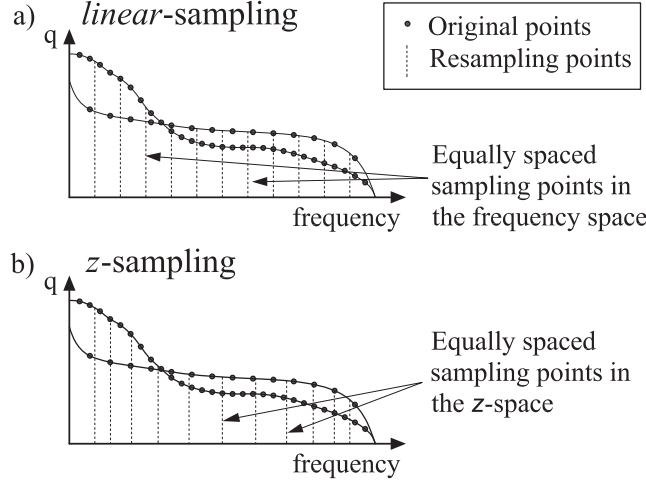


Figure 4.1.: Comparison of dimensionless flow duration curves. Sampling points with constant spacing in frequency representation (a), and with a denser presence on the FDC tails due to normal transformation (b).

s_1 and s_2 , constituted by 365 elements each: $\{q_{1,s_1}, q_{2,s_1}, \dots, q_{365,s_1}\}$ and $\{q_{1,s_2}, q_{2,s_2}, \dots, q_{365,s_2}\}$, a simple measure of their dissimilarity can be defined as the “distance” calculated by the norm of order one,

$$\delta_{s_1,s_2} = \sum_{i=1}^{365} |q_{i,s_1} - q_{i,s_2}|. \quad (4.4)$$

The value δ_{s_1,s_2} can be interpreted also as an approximation of the area between the curves. The computation of the distance according to equation (4.4) is exemplified in figure 4.2 for two generic FDCs.

If n is the number of sites where data are available, the distance measures for each FDC pair are organized in a $n \times n$ distance matrix like:

$$\Delta = \begin{pmatrix} 0 & \delta_{1,2} & \dots & \delta_{1,n} \\ \delta_{2,1} & 0 & & \vdots \\ \vdots & & \ddots & \\ \delta_{n,1} & \dots & & 0 \end{pmatrix} \quad (4.5)$$

where the elements δ_{s_1, s_2} are distances between curves (calculated with equation (4.4)). Analogously, matrices like (4.5) can contain distances between catchment descriptors (if d_1 is the value of the descriptor for basin 1 and d_2 for basin 2, then $\delta_{1,2} = |d_1 - d_2|$). Since the matrices are symmetric and with null diagonal values, after removing the redundant values, only $n(n - 1)/2$ values per matrix are informative.

The distance measure of equation (4.4) not only depends on the resampling method but also on the “measurement space” considered for the representation of flows. For example, if the flows are transformed to provide a more convenient representation of the FDC, the distances δ_{s_1, s_2} are affected by the transformation. Three main representations of the FDC are considered in this work: (a) flow data plotted versus their corresponding plotting position, (b) log-transformed flows versus their corresponding plotting position and (c) log-normal probability plot (log-transformed flows versus normal reduced variate). There are no particular reasons to prefer a priori one of these representations, therefore all of them are considered in the case study and will be respectively referred as “linear representation”, “logarithmic representation” and “log-normal representation” (see figure 4.2). Three parametric functions will be used in a traditional regional FDC estimation exercise in section 4.4, for comparison to the distance-based procedure developed here.

4.2.2. Distance matrices, linear regression and Mantel test

In this section we show how to identify the catchment descriptors that, thanks to their relations with the FDCs, should be used for the formation of cluster regions. A different distance matrix, hereafter termed Δ_{X_i} , is determined for each descriptor, while the distance matrix for the dimensionless FDCs is called Δ_Y . The relation between the distance matrix Δ_Y and the various Δ_{X_i} is assessed using a multi-regressive approach. Note that the multi-regressive approach based on distance matrices is not used to estimate

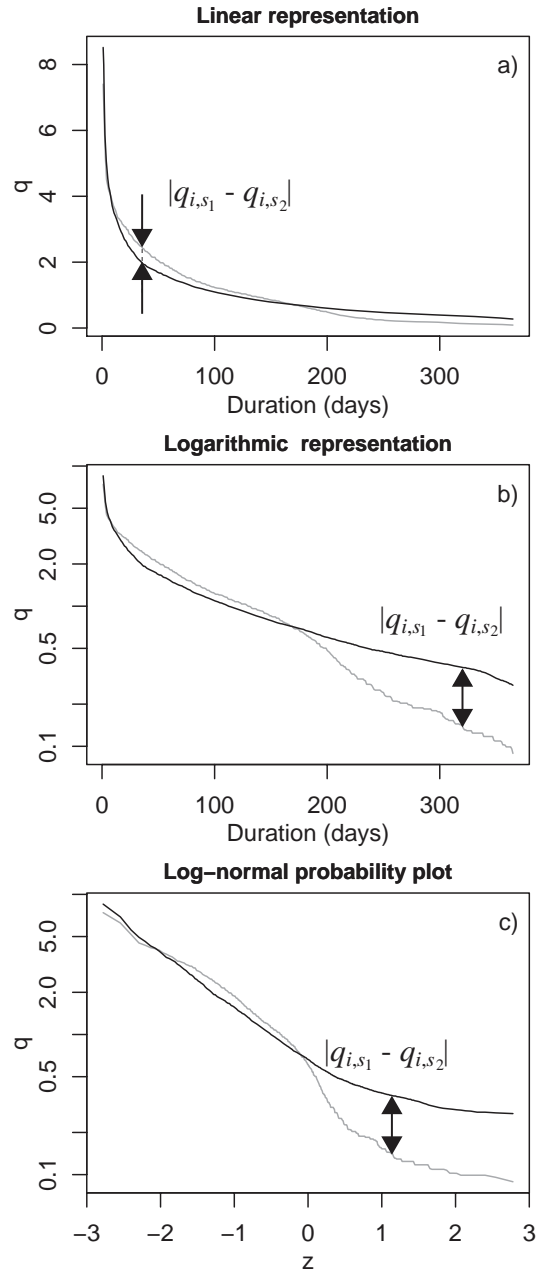


Figure 4.2.: Distance between two FDCs calculated following equation (4.4). The three panels show a pair of FDCs in three different representation spaces: panel (a) is the linear representation (flow values versus exceedance frequency); panel (b) is the logarithmic representation in which discharges are log-transformed; panel (c) represents the log-normal probability plot in which the abscissa is the normal reduced variate z .

FDC coefficients, but to identify the descriptors to be used in the following step for region creation. We start considering a simple linear model:

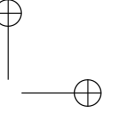
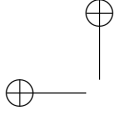
$$\Delta_Y = \beta_0 + \beta_1 \Delta_{X_1} + \dots + \beta_p \Delta_{X_p} + \varepsilon \quad (4.6)$$

with p as the number of descriptors involved, β_i as the regression coefficients and ε the residual matrix. The best possible regression is selected through the *adjusted coefficient of determination*

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (4.7)$$

where R^2 is the standard coefficient of determination [e.g., Kottegoda and Rosso, 1997], p the number of descriptors and n the number of basins considered. The regression coefficients and R^2 can be computed in a standard way [Legendre et al., 1994], that is to say that it does not matter if the elements are organized in a distance matrix. However, in the formulation of the adjusted coefficient of determination it is better to use the value n (the number of basins) instead of $n(n - 1)/2$ that is the number of points involved in the regression (namely the number of distance values). This is due to the fact that the values inside the matrices are not mutually independent. Dependency has another significant impact on the method. In particular, the validity of the tests used to assess the significance of the independent variables (e.g., the Student t test) is affected. A different significance test, as the *Mantel* test [Mantel and Valand, 1970], is then needed, which accounts for the non-independence of the elements in the distance matrices.

The Mantel test was originally proposed by Mantel and Valand [1970] for analysis of correlation between distance matrices, and since then it has been widely improved and used with many different kinds of data. In fact, distance matrices have been frequently used in the biological and ecological sciences

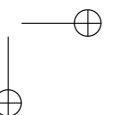
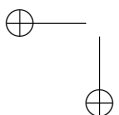


[e.g., Legendre, 1993, Lichstein, 2007]. The *simple Mantel test* [Mantel and Valand, 1970] is used to evaluate the significance of the linear correlation between two distance matrices. This test is performed computing a statistic (usually the Pearson correlation coefficient) between all the pairwise elements of the two matrices. Its significance is tested by repeatedly permuting the objects in one of the matrices, and recomputing the correlation coefficient each time; permutations are performed simultaneously exchanging the rows and the columns of the matrices (e.g., if rows of indexes 2 and 10 are exchanged, also columns of indexes 2 and 10 have to be exchanged [see Legendre et al., 1994]). The significance of the statistic is assessed by comparing its original value to the distribution of values obtained from the permutations, which are considered as many realizations of the null hypothesis of no correlation.

The simple Mantel test can be extended to multiple predictor variables to be applied in multiple linear regression models as (4.6). The extension has been introduced by Smouse et al. [1986], discussed and improved by Legendre et al. [1994] and recently applied in the ecological field by Lichstein [2007]. Following the procedure of Lichstein [2007] each matrix, after removing redundant values, is unfolded into a vector of distances, and regression is performed in the classical way. Then, a null distribution is constructed permuting the elements only in the dependent variable distance matrix Δ_Y . Similarly to what described for the simple Mantel test, the rows and the columns of the matrix Δ_Y are permuted simultaneously and each regression coefficient is tested individually.

4.2.3. Cluster analysis

The proposed procedure serves for the estimation of a FDC in an ungauged basin on the basis of curves relative to other basins. Given a large group of candidate “donor” basins, we want to extract a subset of basins that have geomorphologic and climatic characteristics similar to those of the target site. The FDCs collected in these sites will be used for the estimation of the

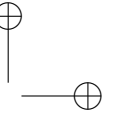
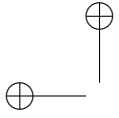


unknown curve. There are different regionalization techniques to choose the subset of basins, for example leading to the formation of fixed regions through cluster analysis [Hosking and Wallis, 1997, Viglione et al., 2007], or based on the method of the region of influence [ROI, Burn, 1990]. In this work the first approach is adopted, selecting fixed regions by splitting the descriptors space in non-overlapping areas by means of a cluster analysis. However, the generalization of the method to the ROI technique is straightforward. The definition of the descriptors space depends on the outcome of the multi-regressive procedure described in section 4.2.2, that allows one to identify a group of significant geomorphoclimatic parameters.

The cluster analysis method used here is a mixed method in which the Ward hierarchical algorithm [Ward, 1963] is followed by a reallocation procedure that minimizes the dispersion within each cluster. The Ward algorithm is agglomerative; it starts with a configuration in which each element is a cluster itself, and progressively merges clusters in a way to produce the minimum information loss, measured as the sum of squared deviation of each element from its cluster centroid. We use the Ward algorithm because it is able to generate compact clusters with an evenly distributed number of elements. A disadvantage is that it does not allow elements reallocation, so that the final configuration could not be the optimal one. To avoid this inconvenience, a reallocation procedure is applied in concurrence with the agglomerative clustering. For instance, if the Ward clustering yields a final configuration with k clusters we compute the statistic

$$W = \sum_{i=1}^k \left(\sum_{j=1}^{n_i} D_{i,j}^2 \right), \quad (4.8)$$

where $D_{i,j}$ is the Euclidean distance between the j -th element of the i -th cluster and the cluster centroid, and n_i is the number of elements contained in the i -th cluster. An element is moved to another cluster if the new con-



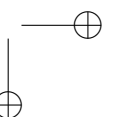
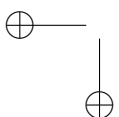
86 Distance-based regional approach for flow duration curves

figuration provides a lower value of W . The procedure ends when W stops reducing after the reallocations, so that every element of a cluster is closer to its center of mass than to the centroid of the nearby cluster.

The reallocation procedure leads to an optimal configuration with k regions. A controversial point of the procedure is the choice of the optimal number of clusters. Usually, in regional analyses, the aim is to get the smallest possible number of homogeneous regions, so that each of them has a large enough number of elements. In this work, the selection of the ideal number of clusters is done investigating different k values and evaluating, for each configuration, a quality index. This index is computed by estimating (in cross-validation mode) the curves for all sites by using the regional model (with a given k), and computing the distance as in equation (4.4) where, in this case, s_1 is the measured curve and s_2 is the estimated one. This distance is adopted as an error measure and the overall mean error is used as a quality index to select the number of clusters. This method does not ensure that the clusters are homogeneous, because no homogeneity test is explicitly used.

After having subdivided the descriptors space in regions, one can proceed to the estimation of the flow duration curve in ungauged sites. For one such site one must first determine the values of the descriptors selected in the procedure of section 4.2.2. The descriptors at the ungauged site are entered as coordinates in the descriptors' space and the site is assigned to the cluster whose centroid is the closest to the basin descriptors. The curves of all basins belonging to the selected cluster will be used to build the regional curve. This latter curve is simply estimated point by point as the average of the values of q relative to each duration for the curves belonging to the selected region, as in the graphical approach described in Castellarin et al. [2004a].

The descriptors used in the cluster analysis are preliminary standardized (i.e., converted into variables with zero mean and unit variance). Standard-



ization of raw descriptors values avoids an unwanted weighting effect due to the different measurement units. If the descriptors are assumed to have different importance in the cluster creation, a procedure can be adopted to give different weights to each descriptor. Regression coefficients of equation (4.6) can be used to compare the relative effect of each descriptor distance matrix, if the distance matrices have been previously standardized: the greater the coefficient, the greater the relative effect of its descriptor distance matrix on the curve distance matrix, so that the coefficients can be used as weights. This weighted clustering procedure will be tested in the following sections by comparing it to the standard unweighted clustering.

After the regional curves have been determined, it is necessary to evaluate if they can be considered significantly different from each other, because otherwise the regions should be merged. To assess if two regional curves are significantly different, we use a procedure based on the distances between curves. First, a *reference distance* is computed as the median (or the mean) of the distances between each empirical curve and the regional one. Then, the distance matrix of the regional curves is computed and all its elements are compared against the reference distance: two regional curves are considered significantly different if their distance is greater than the reference distance, otherwise the two clusters are merged together. This procedure is repeated until all the regional FDCs are significantly different.

Note that the reference distance and the distance matrix of the regional curves depend on the representation space on which the distances are calculated, hence different results are expected using different representation spaces.

4.3. Case study: distance-based method application

4.3.1. Hydrological and geomorphologic data

The application of the distance-based procedure for regional estimation of FDC has been carried out in the R statistical environment [R Development Core Team, 2007], integrated for Mantel test and cluster analysis with the *nsRFA* package [Viglione, 2007a].

Available data include 95 river basins located in northwestern Italy (36 basins of Piemonte and Valle d'Aosta regions) and in Switzerland (59 basins); the geographical location of the gauging stations is shown in figure 4.3. Italian flow data derive from the publications of the former Italian Hydrographic Service and include series lengths ranging between 7 years and 41 years. Hydrological and geomorphological variables relative to Italian basins are included in the widest CUBIST database [CUBIST Team, 2007] that contain such data for more than 500 basins in Italy. The catchment area of Northwestern Italy basins ranges between 22 and 7983 km², and their average elevation ranges from 494 to 2694 m a.s.l. Switzerland data are included in the Reference Hydrometric Network (SHRN) provided by the BAFU (Bundesamtes für UmweltSwiss) and include daily streamflow series with a minimum length of 18 years and a maximum length of 99 years. The catchment area of Switzerland basins ranges between 7 and 616 km², while their average elevation varies from 475 to 2847 m a.s.l. Geomorphological characteristics of each basin has been obtained from a digital terrain model (about 90m cell grid) provided by NASA [2000] with automatic procedures originally developed by Rigon and Zanotti [2002] under a GRASS GIS environment. For the complete list of basins considered, whose codes are referred in figure 4.3, and their geomorphologic variables see appendix B and A.

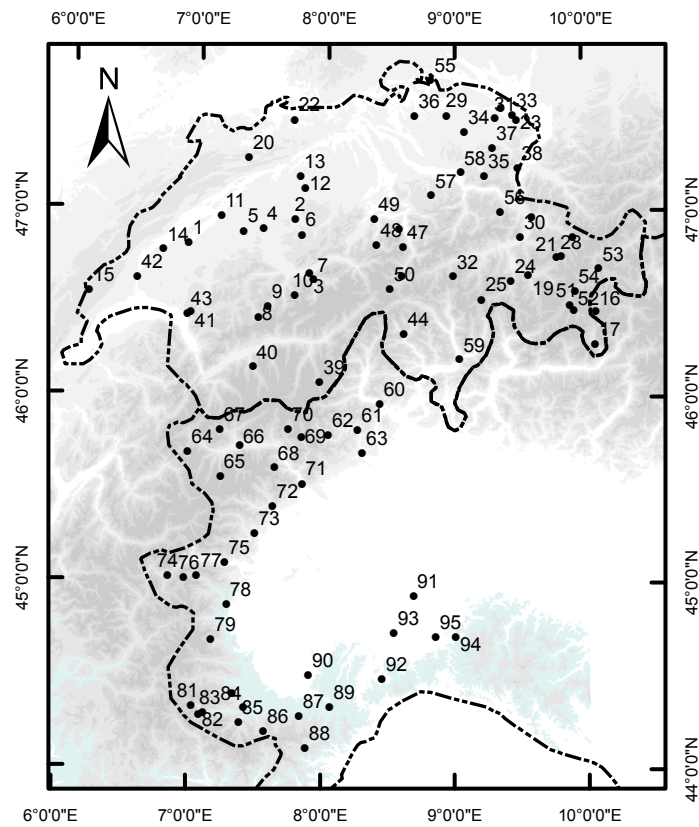


Figure 4.3.: Geographical location of the gauging stations of the 95 catchments considered in the study. Basins 1 to 59 belong to Switzerland, while the remaining ones are located in the Northwestern part of Italy, in Piemonte and Valle d'Aosta regions.

4.3.2. Procedure setting

Several linear regression models between distance matrices have been investigated using relation (4.6). They are built using different combination of:

- Curve distance matrices Δ_Y : the three representations described in section 4.2.1 and figure 4.2 (linear, logarithmic and log-normal plot) are considered;
- Descriptors distance matrices Δ_X : all possible combination from one to five descriptors have been taken into account.

Regression models are ordered in terms of R_{adj}^2 values and tested for significance with the multiple Mantel test, with a significance level of 0.05. Furthermore, a test against multicollinearity has been performed in order to exclude variables with redundant information [Montgomery et al., 2001].

For the linear representation, best results are obtained with four and three descriptors. Lower R_{adj}^2 values arise from simpler models with only two descriptors. In the logarithmic space, the best model is again characterized by four descriptors, but in this case simpler models with two parameters have comparable R_{adj}^2 . In the log-normal space none of the solutions accepted after testing are based on more than two descriptors. We decided to adopt models with two parameters because of their higher robustness (see table 4.I). The R_{adj}^2 values obtained with regression models with distance matrices are very low, although the descriptors result to be statistically significant. In this regard it is important to remind that regressions are only used for the selection of the suitable descriptors and not for direct estimation.

Table 4.I shows the three best models for each representation with two descriptors, where all the models have been tested for significance of regression coefficients with the Mantel test with a level of significance of 0.05. It appears that, considering together the three representations of different curve distance matrices, the most significant descriptors are always the same: the

Table 4.I.: Regression models with two descriptors that well describe the relationship between curve distance matrix and descriptors distance matrices. All the models pass the Mantel test (significance of regression coefficients) with a level of significance of 0.05 and the VIF test (multicollinearity) with threshold equal to 5. The curve distance matrix is calculated in three different representation spaces: the linear, the logarithmic and the log-normal one.

Best relation	Representation space		
	Linear	Logarithmic	Log-normal
1st	$H + MHL$	$H_{min} + MHL$	$H_{min} + MHL$
2nd	$H_{min} + MHL$	$H_{min} + P_m$	$H + MHL$
3rd	$H + p_m$	$P_m + MHL$	$P_m + MHL$

Table 4.II.: Brief description and range of variation of the descriptors used by the distance-based models (see table 4.I).

Descriptor	Definition	Min	Mean	Max
H	mean elevation of the drainage basin above sea level (m)	475	1665	2847
H_{min}	minimum elevation of the drainage basin above sea level (m)	82	839	1974
MHL	mean hillslope length (m)	584.1	759.5	973.6
p_m	average of the slope values associated to each pixel in the DEM of the drainage basin (%)	4	39.9	61.6
P_m	mean large-scale slope (%)	0.8	15.7	50.1

minimum basin elevation (H_{min}), the mean elevation (H), the mean hillslope length (MHL), the mean basin slope (p_m) and the modified basin slope (P_m). A summary of the range of these descriptors is reported in table 4.II. This suggests to adopt the same set of descriptors with all the three representation spaces; H_{min} and MHL has been selected. The adoption of these two descriptors is coherent with the typology of investigated basins. In fact, since we are considering mainly mountain basins, the elevation descriptor is expected to be relevant because of its strong relation to snow-accumulation and snowmelt mechanisms; similarly, the hillslope mean length provides a synthetic description of runoff routing mechanisms.

4.3.3. Regions definition

The second step, after the choice of the suitable descriptors, is to pool the catchments together with the cluster analysis, as described in section 4.2.3. The procedure is applied to both the weighted and the unweighed cluster configurations. For all the three representation spaces, the unweighed procedure often demonstrates better performances, while the weighted procedure leads to marginal, if any, improvements that do not justify its use. Following the criteria mentioned in section 4.2.3 and considering all the three representation spaces, the suggested number of clusters obtained for Italian and Switzerland data is four.

This configuration is then checked, to assess if the regional FDCs are significantly different, using the procedure described in section 2.3 for all the three representation spaces. The FDCs of the original four clusters cannot be considered significantly different from each other, neither in the linear space, nor in the other two logarithmic spaces. Thus, for each representation space, the two most similar clusters are merged together. The new configurations with three clusters can be accepted in the linear space only. Applying again the procedure for the logarithmic and log-normal space we obtain two configurations consisting of two clusters each. To select one among these different configurations of clusters, we perform the following cross-evaluation: for each set of clusters (e.g., the one obtained in the linear space), we check if the difference between the regional FDCs is significant in the other representation spaces (i.e., also in the logarithmic and log-normal spaces). Based on this cross-evaluation, we choose the configuration with 2 clusters obtained in the logarithmic space, which is represented in Figures 4.4 and 4.5. Hence, this latter configuration will be used as the result of the distance-based model.

The final regions obtained are shown in figure 4.4. Curves belonging to each cluster are grouped together and the regional curves are derived as the average of all curves belonging to the region. Figure 4.5 shows the regional curves (black lines) obtained from curves belonging to the cluster (grey lines)

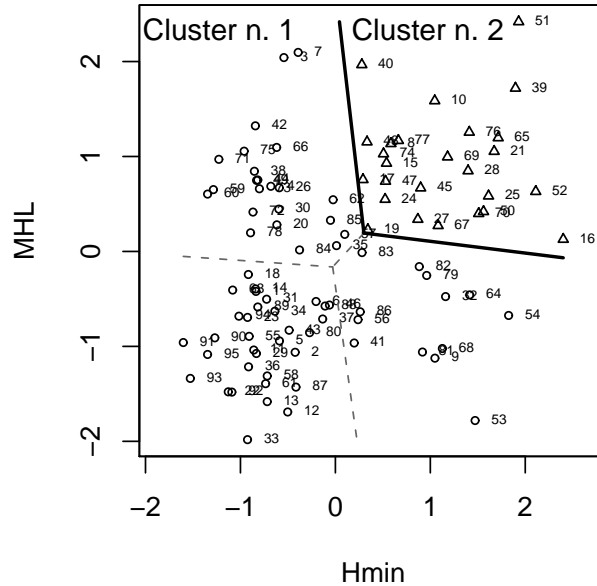


Figure 4.4.: Disjoint regions in the space of catchment descriptors: Hmin is the minimum basin elevation and MHL is the mean hillslope length. The dashed lines represent the boundaries between the 4 clusters obtained before merging the clusters whose FDCs cannot be considered significantly different. The final 2 disjoint regions are separated by the solid line.

in the log-normal space. Although every curve bundle appears to be quite wide, regional curves are able to represent two characteristic behaviors. In fact, we can observe an almost straight curve and a “S” shaped curve. A quantitative representation of model quality and estimation errors is reported in the following section, where a comparison against some parametric methods is performed.

4.4. Comparison with parametric models

The distance-based regional procedure developed in this work is tested against some standard parametric regional models. In general, the choice of the reference model is not trivial and more than one function can be used to describe

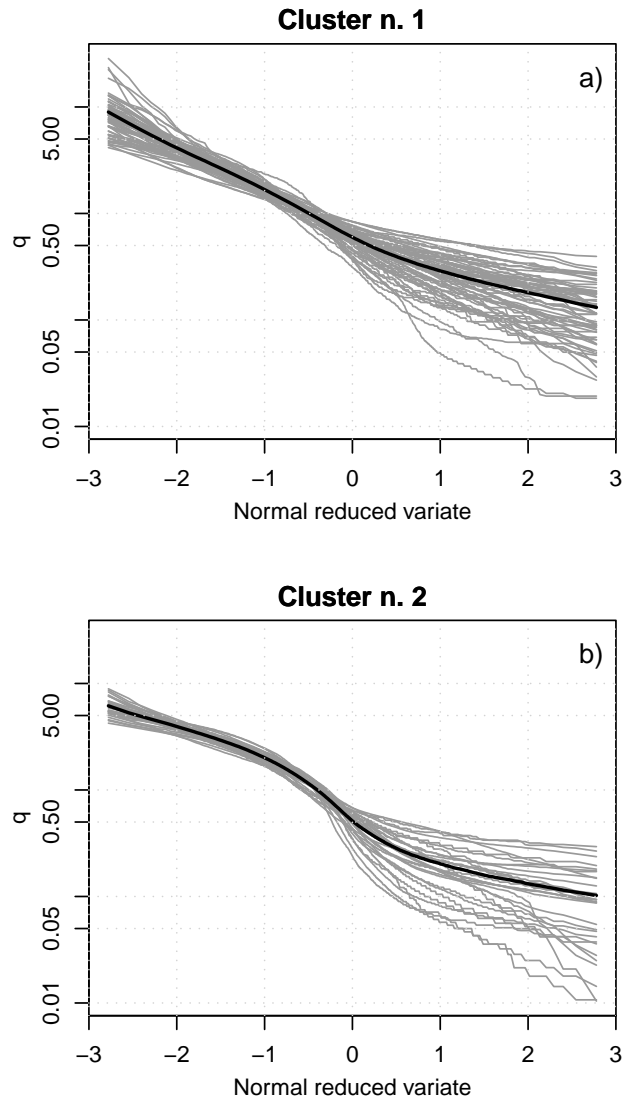


Figure 4.5.: Flow duration curves grouped by cluster (in grey) and corresponding regional curves (in black).

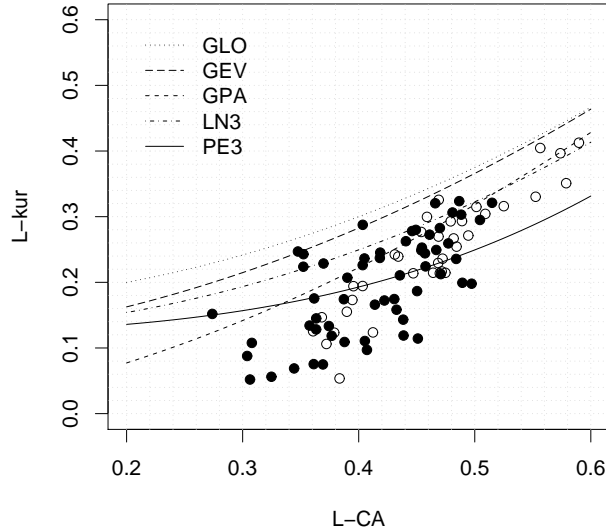


Figure 4.6.: L-moments ratio diagram for the dimensionless FDCs of the 95 basins (filled circles for Switzerland data and white circles for Italian data). The lines indicate different theoretical three-parameter distributions: generalized logistic (GLO), generalized extreme-value (GEV), generalized Pareto (GPA), lognormal (LN3), Pearson type III (PE3).

the FDCs. For this purpose, a useful tool is the L -moments ratio diagram of figure 4.6 [Hosking and Wallis, 1997] where one plots the L_{CA} (coefficient of L -skewness) of each dimensionless FDC versus its corresponding L_{kur} (coefficient of L -kurtosis). The lines represent the domain of the distributions over the $L_{CA} - L_{kur}$ space and can help one to identify the distribution to be used. This approach has been followed, for example, by Castellarin et al. [2007].

In this work, the analysis is performed over a database of 95 basins that have very different characteristics in terms of L_{CA} and L_{kur} , as figure 4.6 shows. The scattering of the points make the choice of the distribution rather difficult. For this reason, different parametric models are used for the comparison with the distance-based procedure.

Each parameter θ of a parametric model is related to the catchments' descriptors d by a linear model of the form

$$\theta = a_0 + a_1 \cdot d_1 + a_2 \cdot d_2 + \dots + a_n \cdot d_n + \varepsilon. \quad (4.9)$$

The first step is to identify a suitable regional model to estimate the generic parameter for an ungauged station, where θ is previously estimated at each station s using a suitable technique. The resulting parameters θ_s are then related to descriptor data (raw data, not distances) for all the catchments (not classified in regions) to identify a regional model (regression) able to describe them. Many linear models of the form of equation (4.9) are considered and validated with a t-Student test followed by a multicollinearity (VIF) test and subsequently ordered by their values of R_{adj}^2 [e.g., Montgomery et al., 2001].

The models considered here are the two-parameter log-normal distribution (LN2), the three-parameter Pearson type III (PE3) and the generalized Pareto (GPA) distributions. The log-normal model is represented by the relation

$$\log(q) = \theta_1 + \theta_2 \cdot z \quad (4.10)$$

where z is the quantile of a normal distribution with zero mean and unit variance corresponding to each flow's plotting position values. In the log-normal probability representation, equation (4.10) is a straight line whose coefficients θ_1 and θ_2 can be estimated with a least squares linear regression.

The GPA probability density function is defined as

$$f(q) = \theta_2^{-1} \exp[-(1 - \theta_3)y], \quad (4.11)$$

with $y = -\theta_3^{-1} \log[1 - \theta_3(q - \theta_1)/\theta_2]$ if $\theta_3 \neq 0$ and $y = (q - \theta_1)/\theta_2$ if $\theta_3 = 0$, where θ_1 , θ_2 and θ_3 are the location, scale and shape parameter, respectively; the PE3 probability density function is defined as

$$f(q) = \frac{(q - \theta_1)^{\theta_2 - 1} \exp[-(q - \theta_1)/\theta_3]}{\theta_3^{\theta_2} \Gamma(\theta_2)}, \quad (4.12)$$

where θ_1 , θ_2 and θ_3 are the location, scale and shape parameter, respectively, and $\Gamma(\cdot)$ is the gamma function. For details about these distributions and for parameters estimation refer to Hosking and Wallis [1997] and Viglione [2007a]. The regional estimation of the models' parameters use the descriptors listed in table 4.III

The distance-based model and the parametric ones are all tested using a cross-validation approach in which one station is considered ungauged and its data are removed from the database. The models are then recalibrated using only the remaining data, and the unknown curve is estimated. After this procedure is repeated for all basins, one can compute, for each basin, the error measure $\delta_{\text{MOD,EMP}}$ as the distance between the estimated FDC and its empirical counterpart.

The non-parametric FDC representation method performs better than the parametric models for most of the analyzed basins, independently of the representation space considered. Figure 4.7 shows a comparison between the errors $\delta_{\text{MOD,EMP}}$ calculated with the parametric and the distance-based approaches. Each parametric model is able to well describe only a subset of the studied basins (see figure 4.6), which is probably the reason why they demonstrate similar and non excellent performances when applied to the whole dataset.

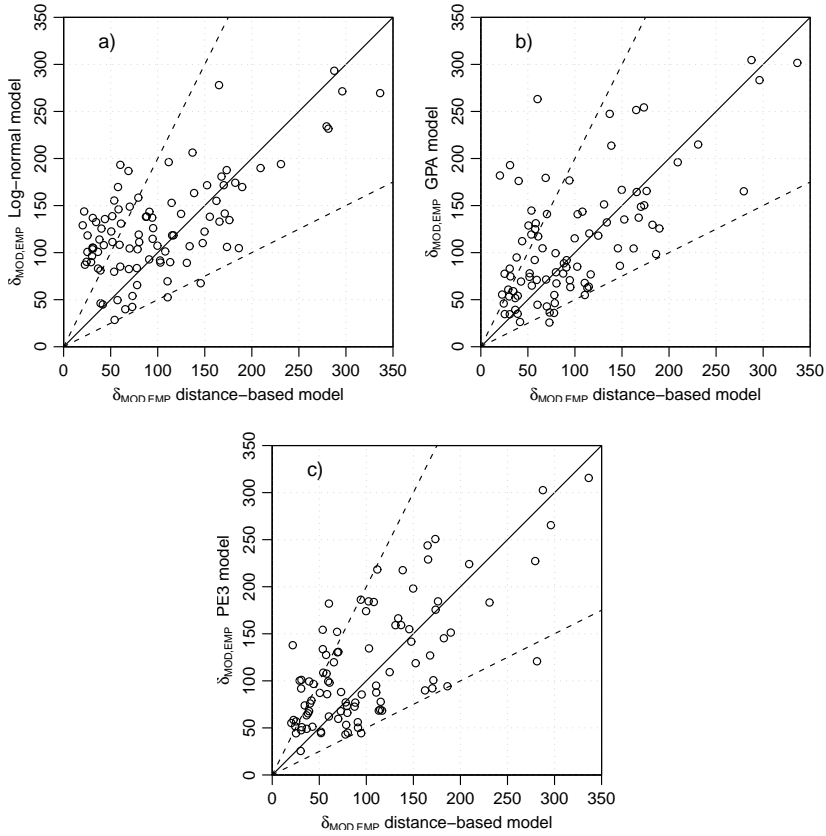


Figure 4.7.: Quality of estimated dimensionless FDCs by the distance-based method compared with the log-normal model (a), the generalized Pareto model (b) and the Pearson type III model (c). The distance between the empirical curve and the estimated one $\delta_{MOD,EMP}$ is reported in the scatter plot for each considered basin. The solid line represents the ratio 1:1 between the errors, while dashed lines delimit the areas where errors for the distance-based model are twice the parametric ones, and viceversa. Points above the solid line represent curves better estimated by the distance-based method; points above the upper dashed line represent curves much better estimated by the distance-based method.

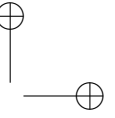
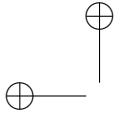
Table 4.III.: Descriptors used to estimate the parametric model's parameters with level of significance (Student test) and Variance Inflation Factor test. See appendix B for more detailed references about descriptors.

Model	Parameter	Descriptors	Student	VIF	R_{adj}^2
Lognormal	θ_1	MA, C_c	< 0.05	< 5	0.12
	θ_2	$X_c, SLDP, p_m, MHL$	< 0.05	< 5	0.17
GPA	θ_1	$X_{max}, SLDP, P_m, MHL$	< 0.02	< 5	0.28
	θ_2	$Y_{min}, IPS25, \cos(O_{ov})$	< 0.05	< 5	0.54
	θ_3	$X_c, Y_c, IPS50$	< 0.05	< 5	0.39
PE3	θ_1	$X_{max}, SLDP, p_m, C_c, MHL$	< 0.02	< 5	0.39
	θ_2	$X_c, Y_{min}, IPS100, C_c$	< 0.05	< 5	0.31
	θ_3	$Y_{min}, IPS50$	< 0.02	< 5	0.28

4.5. Final remarks

The procedure for dimensionless flow duration curves estimation in ungauged basins developed in this chapter hinges on the concept of grouping basins based on distances, that quantitatively represents the dissimilarity between curves and catchment's descriptors. This approach, based on distance matrices, allows one to account for a FDC as a whole object, avoiding the description of the curve by means of a parametric function. Moreover, no assumptions on the shape of the FDCs is made. This is an important feature when one has to manage at the same time curves described by a simple geometry (e.g., almost straight lines in the log-normal probability plot) and curves with more complex behavior (e.g., "S" shaped curves). In fact, complex shapes can be well described by a parametric model only using an high number of parameters, that sometimes can not guarantee a robust parameters estimation.

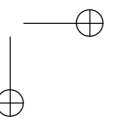
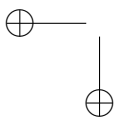
The results obtained by means of the distance-based model (non-parametric representation of the FDC) applied to the present dataset are comparable, and many times better, than the estimation yielded by classical parametric models of the same or greater complexity. These results are obtained on the basis of only two descriptors, while the log-normal model requires six

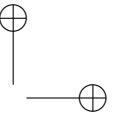
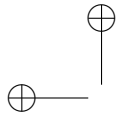


100 Distance-based regional approach for flow duration curves

descriptors for the assessment of two parameters, and the PE3 and GPA models require 8 and 10 descriptors to estimate their three parameters.

The main advantage of the method based on distance matrices is its ability in dealing with curves. For instance, the regionalization method proposed here could be improved considering also “complex” catchment descriptors as the hypsographic curve, or climatic information like the precipitation regime curve.





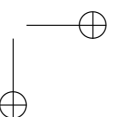
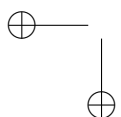
Chapter 5.

Summary and conclusions

The issues addressed in this thesis are focused around the streamflow estimation in the so called ungauged basins, i.e. catchments for which no discharge measurement are available, or where recorded data are not adequate for the analysis. In these cases, the most common approach is to use the information known at some gauged locations to characterize the ungauged basins. This procedure can be done through different statistical procedure.

The first approach has been developed in chapter 2 in which a regional procedure has been studied in order to improve the estimation of flood flows in ungauged basins. The method is thought in particular for areas where many data-scarce stations are present. In fact, short records are usually discarded by traditional regionalization approaches, but they still contain useful information. This is the case of the area presented in the case study, where a lot of gauging stations have been abandoned and re-established only in recent times. The method can be also helpful where new gauging stations have been just installed, because it allows to consider also statistics based on few data.

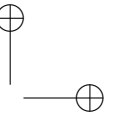
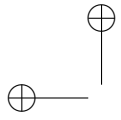
The procedure refers the problem of frequency curve reconstruction in the regionalization of three different L -moments that are subsequently combined. This can be considered as an extension of the index-flood approach, in which the mean is used as a scale factor, while L_{CV} and L_{CA} are combined to build the growth curve. The advantage of this approach where stations with a variable number of data are available is evident: one can calculate the



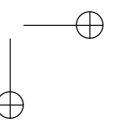
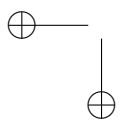
lower-order L -moments on the sample record and regionalise the higher-order ones. The proposed model also allows a clear treatment of the prediction uncertainty. The procedure has been calibrated over a set of 70 catchments and then has been applied in order to map the regional prediction over the whole drainage network (see for example figure 1.2).

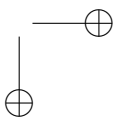
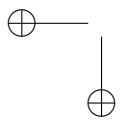
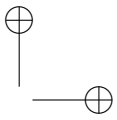
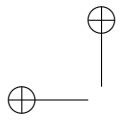
The regional approach is a powerful tool for prediction in ungauged basins, however, it neglects some local-scale information, like the presence of a gauging station close to the ungauged site. The effects of proximity has been investigated in chapter 3 where a procedure to correct the regional prediction is proposed, based on the information retrieved from close gauged stations. This approach is proved to be suitable for the index-flood when the areas of the two catchments have a ratio lower than 10, while no appreciable improvements are found for higher-order L -moments. Anyway, the framework developed in chapter 3 can be easily extended to work also with other kind of models.

Finally, chapter 4 discusses on how to deal with target variables that are not represented by numbers, but that are considered as “whole” curves. Each curve is classified by comparison against other curves so that their dissimilarity can be quantified by a distance measure. Statistical analysis of distances allows grouping the curves in a basin descriptor’s domain and, consequently, allows the whole curve prediction in ungauged basins. The regional model has been applied to the case of flow duration curves in Northwester Italy obtaining a good prediction ability, equal and often better than more traditional approaches. Besides, the distance-based approach has wide improvement possibilities, as that to directly handle “complex” descriptors like curves (e.g. precipitation regime, hypsometric curve, etc), raster maps clipped over basin boundaries (e.g. mean annual precipitation, vegetation cover, etc) and non-numeric/cathegoric data (e.g. soil stratigraphy). Also the tree-structure of river network can be accounted by the distance-based approach, although through a more complicated methodology.



The models developed in this thesis have been applied to the flood frequency curve and the flow duration curve, as hydrological variables; however, the methods proposed are general and can be used also with other kinds of data.





References

- AC Bayliss and DW Reed. The use of historical data in flood frequency estimation. Technical report, Centre for Ecology and Hydrology, 2001.
- DH Burn. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, 26(10):2257–2265, 1990.
- KP Burnham and DR Anderson. *Model Selection and Multi-Model Inference*. Springer, second edition edition, 2002.
- A Castellarin, G Galeati, L Brandimarte, A Montanari, and A Brath. Regional flow-duration curves: reliability for ungauged basins. *Advances in Water Resources*, 27(10):953 – 965, 2004a. ISSN 0309-1708.
- A Castellarin, RM Vogel, and A Brath. A stochastic index flow model of flow duration curves. *Water Resources Research*, 40(3), 2004b. ISSN 0043-1397.
- A Castellarin, G Camorani, and A Brath. Predicting annual and long-term flow-duration curves in ungauged basins. *Advances in Water Resources*, 30(4):937 – 953, 2007. ISSN 0309-1708.
- A Castellarin, DH Burn, and A Brath. Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem? *Journal Of Hydrology*, 360(1-4):67–76, OCT 15 2008. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2008.07.014.
- F Chebana and TBMJ Ouarda. Depth and homogeneity in regional flood frequency analysis. *Water Resources Research*, 44(11), NOV 15 2008. ISSN 0043-1397. doi: 10.1029/2007WR006771.

- K Chokmani and TBMJ Ouarda. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. *Water Resources Research*, 40(12), DEC 28 2004. ISSN 0043-1397. doi: 10.1029/2003WR002983.
- P Claps and M Fiorentino. *Integrated Approach to Environmental Data Management Systems*, volume 2 (31) of *NATO-ASI series*, chapter Probabilistic Flow Duration Curvers for use in Environmental Planning and Management, pages 255–266. Harmancioglu et al., Kluwer, Dordrecht, The Netherlands, 1997.
- CUBIST Team. Cubist project: Characterisation of ungauged basins by integrated use of hydrological techniques. Geophysical Research Abstracts, Vol. 10, EGU2008-A-12048, 2008 SRef-ID: 1607-7962/gra/EGU2008-A-12048 EGU General Assembly 2008, 2007. URL <http://www.cubist.polito.it/>.
- C Cunnane. Methods And Merits Of Regional Flood Frequency-Analysis. *Journal Of Hydrology*, 100(1-3):269–290, JUL 30 1988. ISSN 0022-1694.
- T Dalrymple. *Flood frequency analyses*, volume 1543-A of *Water Supply Paper*. U.S. Geological Survey, Reston, Va., 1960.
- EAH Elmir and AH Seheult. Exact variance structure of sample L-moments. *Journal Of Statistical Planning And Inference*, 124(2):337–359, SEP 1 2004. ISSN 0378-3758. doi: 10.1016/S0378-3758(03)00213-1.
- N Fennessey and RM Vogel. Regional flow-duration curves for ungauged sites in massachusetts. *Journal of Water Resources Planning and Management-ASCE*, 116(4):530 – 549, 1990. ISSN 0733-9496.
- S Gabriele and N Arnell. A hierarchical approach to regional flood frequency-analysis. *Water Resources Research*, 27(6):1281–1289, JUN 1991. ISSN 0043-1397.

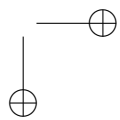
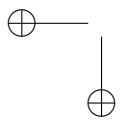
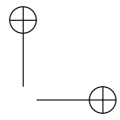
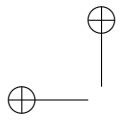
- L Gottschalk. Correlation and covariance of runoff. *Stochastic Hydrology and Hydraulics*, 7(2):85–101, JUN 1993a. ISSN 0931-1955.
- L Gottschalk. Interpolation of runoff applying objective methods. *Stochastic Hydrology and Hydraulics*, 7(4):269–281, DEC 1993b. ISSN 0931-1955.
- L Gottschalk, I Krasovskaia, E Leblois, and E Sauquet. Mapping mean and variance of runoff in a river basin. *Hydrology and Earth System Sciences*, 10(4):469–484, 2006. ISSN 1027-5606.
- VW Griffis and JR Stedinger. The use of GLS regression in regional hydrologic analyses. *Journal Of Hydrology*, 344(1-2):82–95, SEP 30 2007. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2007.06.023.
- GT Hahn. Simultaneous prediction intervals for a regression model. *Technometrics*, 14(1):203–214, February 1972.
- MJ Hall and AW Minns. The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 44(5):693–704, OCT 1999. ISSN 0262-6667.
- RM Hirsch. Probability plotting position formulas for flood records with historical information. *Journal Of Hydrology*, 96(1-4):185–199, DEC 15 1987. ISSN 0022-1694.
- MGR Holmes, AR Young, A Gustard, and R Grew. A region of influence approach to predicting flow duration curves within ungauged catchments. *Hydrology and Earth System Sciences*, 6(4):721 – 731, 2002. ISSN 1027-5606.
- JRM Hosking and JR Wallis. *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, 1997.
- V Iacobellis. Probabilistic model for the estimation of t year flow duration curves. *Water Resources Research*, 44(2):W02413, 2008. ISSN 0043-1397.

- Interagency Advisory Committee on Water Data. *Bulletin 17B - Guidelines for determining flood flow frequency*. US Department of the Interior Geological Survey, 1982.
- NL Johnson and S Kotz. *Distributions in statistics*. Applied probability and statistics. Wiley, New York, 1986.
- TR Kjeldsen and D Jones. Estimation of an index flood using data transfer in the UK. *Hydrological Sciences Journal-Journal des Sciences Hydrologiques*, 52(1):86–98, FEB 2007. ISSN 0262-6667.
- NT Kottegoda and R Rosso. *Statistics, Probability, and Reliability for Civil and Environmental Engineers*. McGraw-Hill Companies, international edition, 1997. ISBN 0-07-035965-2.
- P Lacau and H Chevrier. *Une chapelle de Sesostris 1er a Karnak*. Institut français d'archéologie orientale, 1956.
- F Laio, G Di Baldassarre, and A Montanari. Model selection techniques for the frequency analysis of hydrological extremes. *Water Resources Research*, 45, JUL 18 2009. ISSN 0043-1397. doi: 10.1029/2007WR006666.
- M Lauro. Cartigli e faraoni d'egitto. Internet resource (in Italian), 2009.
- P Legendre. Spatial autocorrelation - trouble or new paradigm. *ECOLOGY*, 74(6):1659 – 1673, 1993. ISSN 0012-9658.
- P Legendre and L Legendre. *Numerical Ecology*. Elsevier Science, Amsterdam, 2nd edition, 1998.
- P Legendre, FJ Lapointe, and P Casgrain. Modeling brain evolution from behavior - a permutational regression approach. *Evolution*, 48(5):1487 – 1499, 1994. ISSN 0014-3820.
- JW Lichstein. Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, 188(2):117 – 131, 2007. ISSN 1385-0237.

- N Mantel and RS Valand. A technique of nonparametric multivariate analysis. *Biometrics*, 27:209 – 220, 1970.
- R Merz and G Bloschl. Flood frequency regionalisation-spatial proximity vs. catchment attributes. *Journal Of Hydrology*, 302(1-4):283–306, FEB 1 2005. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2004.07.018.
- DC Montgomery, EA Peck, and GG Vining. *Introduction to linear regression analysis*. Wiley Series in Probability and Statistics, third edition, 2001.
- MPV Mosley and AI McKerchar. *HandBook of Hydrology*, chapter 8, page 39. McGraw-Hill Companies, international edition, 1993.
- NASA. SRTM, 2000. URL <http://www2.jpl.nasa.gov/srtm/index.html>.
- TBMJ Ouarda, KM Ba, C Diaz-Delgado, A Carsteanu, K Chokmani, H Gingras, E Quentin, E Trujillo, and B Bobee. Intercomparison Of Regional Flood Frequency Estimation Methods At Ungauged Sites For A Mexican Case Study. *Journal Of Hydrology*, 348(1-2):40–58, JAN 1 2008. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2007.09.031.
- E Pekalska and RPW Duin. *The dissimilarity representation for pattern recognition*. World Scientific, Singapore, 2005.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- DS Reis, JR Stedinger, and ES Martins. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. *Water Resources Research*, 41(10), OCT 26 2005. ISSN 0043-1397. doi: 10.1029/2004WR003445.
- R Rigon and F Zanotti. *The Fluid Turtle Library, Users and Programmers Guide*. University of Trento, Italy, 2002.

- GAF Seber and CJ Wild. *Nonlinear Regression*. Series in probability and mathematical statistics. Wiley, New York, 1989.
- RD Singh, SK Mishra, and H Chowdhary. Regional flow-duration models for large number of ungauged himalayan catchments for planning microhydro projects. *Journal of Hydrologic Engineering*, 6(4):310 – 316, 2001. ISSN 1084-0699.
- M Sivapalan, K Takeuchi, SW Franks, VK Gupta, H Karambiri, V Lakshmi, X Liang, JJ McDonnell, EM Mendiondo, PE O’Connell, T Oki, JW Pomeroy, D Schertzer, S Uhlenbrook, and E Zehe. IAHS decade on predictions in ungauged basins (pub), 2003-2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences - Journal - des Sciences Hydrologiques*, 48(6):857–880, 2003.
- JO Skoien, R Merz, and G Bloschl. Top-kriging - geostatistics on stream networks. *Hydrology And Earth System Sciences*, 10(2):277–287, 2006. ISSN 1027-5606.
- VU Smakhtin. Low flow hydrology: a review. *Journal of Hydrology*, 240(3-4):147 – 186, 2001. ISSN 0022-1694.
- PE Smouse, JC Long, and RR Sokal. Multiple-regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, 35(4):627 – 632, 1986. ISSN 0039-7989.
- JR Stedinger and GD Tasker. Regional Hydrologic Analysis .1. Ordinary, Weighted, And Generalized Least-Squares Compared. *Water Resources Research*, 21(9):1421–1432, 1985. ISSN 0043-1397.
- S Uhlenbrook. Catchment hydrology - a science in which all processes are preferential - Invited commentary. *Hydrological Processes*, 20(16):3581–3585, OCT 30 2006. ISSN 0885-6087. doi: 10.1002/hyp.6564.

- US National Research Council. *Estimating Probabilities of Extreme Floods*. Nat. Academy Press, Washington DC (USA), 1988.
- A Viglione. *nsRFA: Non-supervised Regional Frequency Analysis*, 2007a. URL http://www.idrologia.polito.it/~alviglio/index_en.htm. R package version 0.4-5.
- A Viglione. A simple method to estimate variance and covariance of sample L-CV and L-CA, December 2007b.
- A Viglione, F Laio, and P Claps. A comparison of homogeneity tests for regional frequency analysis. *Water Resources Research*, 43(3):W03428, 2007. ISSN 0043-1397.
- RM Vogel and NM Fennessey. Flow-duration curves .2. new interpretation and confidence-intervals. *Journal of Water Resources Planning and Management-ASCE*, 120(4):485 – 504, 1994. ISSN 0733-9496.
- QJ Wang. Unbiased estimation of probability weighted moments and partial probability weighted moments from systematic and historical flood information and their application to estimating the GEV distribution. *Journal Of Hydrology*, 120(1-4):115–124, DEC 1990. ISSN 0022-1694.
- JH Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.



Appendix A.

Hydrological data summary

A.1. Hydrological data for flood flows modelling

The analysis of the flood frequency curves reported in chapter 2 and 3 is based on a set of maximum annual discharge data available for some Italian basins. The following tables provide a short summary of data consistency and sample L -moments of the basins involved in the analysis. Refer to section 2.2.1 for symbols definition.

Code	River	Station	n	non-syst.	l	m	Threshold
1	Artanavaz	St.Oyen	14	0	0	14	Inf
2	Ayasse	Champorcher	29	0	0	29	Inf
3	Borbera	Baracche	21	1	5	37	613
7	Bormida di	Ferrania	22	1	5	34	310
	Mallare						
8	Cervo	Passobreve	13	0	0	13	Inf
9	Chisone	Fenestrelle	18	1	4	22	68
10	Chisone	S.Martino	21	5	9	56	358
11	Chisone	Soucheres	21	1	0	40	94
		Basses					
12	Corsaglia	Presa	25	0	0	25	Inf
		C.Molline					
13	Dora Baltea	Aosta	25	0	0	25	Inf
14	Dora Baltea	Ponte di	14	0	0	14	Inf
		Mombardone					
15	Dora Baltea	Tavagnasco	72	1	15	74	1000
16	Dora di Bar-	Beaulard	12	0	0	12	Inf
	donecchia						
19	Dora Riparia	Oulx	30	0	0	30	Inf
20	Dora Riparia	S.Antonino di	59	1	1	75	350
		Susa					
22	Erro	Sassello	21	0	0	21	Inf
24	Evancon	Champoluc	22	0	0	22	Inf
25	Gesso	Entraque	12	0	0	12	Inf
26	Gesso della	S. Lorenzo	11	0	0	11	Inf
	Valletta						
27	Grana	Monterosso	48	0	0	48	Inf
29	Lys	D'Ejola	10	0	0	10	Inf
30	Lys	Gressoney	24	0	0	24	Inf
		St.Jean					
31	Mastallone	Ponte Folle	54	0	0	54	Inf
33	Orco	Pont	41	2	2	73	1500
		Canavese					

Code	River	Station	n	non-syst.	l	m	Threshold
35	Po	Crissolo	14	0	0	14	Inf
38	Rio Bagni	Bagni Vinadio	20	0	0	20	Inf
40	Rutor	Promise	34	0	0	34	Inf
41	San Bernardino	Santino	14	0	0	14	Inf
42	Savar	Eau Rousee	18	0	0	18	Inf
43	Scrivia	Isola del Cantone	13	0	0	13	Inf
44	Scrivia	Serravalle	26	4	3	72	1650
45	Sesia	Campertogno	38	0	0	38	Inf
47	Sesia	Ponte Aranco	15	2	4	43	2150
48	Sesia	Vercelli	22	0	0	22	Inf
50	Stura di Demonte	Fossano	33	0	0	33	Inf
51	Stura di Demonte	Pianche	18	0	0	18	Inf
52	Stura di Lanzo	Lanzo	60	4	9	81	800
53	Stura di Vi	Usseglio	11	0	0	11	Inf
56	Tanaro	Farigliano	63	2	7	75	1150
58	Tanaro	Nucetto	47	0	0	47	Inf
59	Tanaro	Ormea	13	0	0	13	Inf
60	Tanaro	Ponte Nava	42	1	10	50	181
62	Toce	Cadarese	15	0	0	15	Inf
63	Toce	Candoglia	55	4	13	70	1700
64	Varaita	Rore	58	0	0	58	Inf
66	Vobbia	Vobbietta	14	0	0	14	Inf
68	Breuil	Alpette	14	0	0	14	Inf
70	Chiavanne	Alpette	14	0	0	14	Inf
72	Dora di Rhemes	Notre Dame	14	0	0	14	Inf
80	Rutor	La Joux	29	0	0	29	Inf
91	Varaita	Castello	56	0	0	56	Inf
98	Lys	Guillemore	29	0	0	29	Inf
99	Chiusella	Gurzia	31	2	2	74	820
112	Stura di Viu	Malciaussia	48	1	1	64	33.5
115	Marmore	Perreres	15	0	0	15	Inf
118	Sermenza	Rimasco	44	0	0	44	Inf
124	Maira	S.Damiano	57	0	0	57	Inf
		Macra					
126	Maira	Saretto	13	0	0	13	Inf
128	Bormida	Valla	47	0	0	47	Inf
		Spigno					
131	Adda	Fuentes	42	0	0	42	Inf
134	Adda	Tirano	11	0	0	11	Inf
136	Aveto	Cabanne	27	0	0	27	Inf
138	Brembo	P.te Briolo	29	1	1	43	1580
164	Serio	P.te Cene	23	2	1	44	547
165	Taro	Pradella	13	0	0	13	Inf
168	Taro	Carniglia	29	0	0	29	Inf
169	Taro	Ostia	10	0	0	10	Inf
172	Trebbia	S.Salvatore	17	0	0	17	Inf
173	Trebbia	Due Ponti	20	0	0	20	Inf
174	Trebbia	Valsigiara	27	0	0	27	Inf

Code	Q_{ind}	$\sigma_{Q_{ind}}^2$	LCV	σ_{LCV}^2	LCA	σ_{LCA}^2	L_{kur}
1	12.6	2.4	0.227	0.003	0.433	0.036	0.395
2	19.4	3.3	0.266	0.002	0.274	0.013	0.229
3	256.7	1751.2	0.417	0.007	0.286	0.018	0.150
7	163.6	912.6	0.482	0.009	0.403	0.022	0.172
8	94.7	230.8	0.345	0.007	0.207	0.025	0.131
9	32.8	36.2	0.497	0.011	0.524	0.032	0.102
10	207.4	943.7	0.412	0.007	0.307	0.019	0.300
11	18.6	16.3	0.441	0.007	0.467	0.025	0.382

Code	Q_{ind}	$\sigma_{Q_{ind}}^2$	LCV	σ_{LCV}^2	LCA	σ_{LCA}^2	L_{kur}
12	36.9	20.3	0.268	0.002	0.425	0.020	0.407
13	285.5	578.1	0.238	0.002	0.184	0.013	0.104
14	93.9	11.9	0.083	0.000	-0.065	0.017	-0.001
15	812.6	2745.6	0.270	0.001	0.309	0.006	0.278
16	27.4	7.8	0.213	0.003	0.243	0.030	0.072
19	55.8	95.7	0.384	0.004	0.571	0.021	0.432
20	99.0	40.8	0.284	0.001	0.320	0.007	0.255
22	103.4	52.7	0.191	0.001	0.122	0.013	0.069
24	26.4	10.0	0.276	0.003	0.430	0.023	0.286
25	76.4	348.0	0.426	0.012	0.468	0.045	0.372
26	67.6	248.5	0.353	0.009	0.632	0.063	0.552
27	40.9	43.0	0.495	0.004	0.516	0.012	0.356
29	12.7	1.4	0.185	0.003	-0.033	0.022	-0.099
30	28.8	8.4	0.260	0.002	0.332	0.018	0.227
31	380.0	1259.5	0.368	0.002	0.271	0.007	0.157
33	498.4	3047.9	0.427	0.004	0.465	0.013	0.368
35	34.3	114.9	0.587	0.020	0.542	0.043	0.290
38	24.2	50.3	0.540	0.012	0.712	0.038	0.487
40	16.1	2.2	0.221	0.001	0.487	0.016	0.401
41	259.5	437.7	0.183	0.002	0.091	0.018	0.072
42	24.6	3.9	0.205	0.002	0.043	0.013	0.100
43	412.8	5972.6	0.365	0.008	0.406	0.037	0.266
44	641.0	4982.0	0.281	0.002	0.151	0.011	-0.048
45	160.8	575.4	0.428	0.004	0.500	0.015	0.317
47	994.1	28801.9	0.315	0.005	-0.100	0.017	-0.046
48	1673.2	32196.1	0.291	0.003	0.161	0.014	0.160
50	104.9	236.0	0.393	0.004	0.458	0.016	0.333
51	39.5	53.1	0.376	0.006	0.432	0.028	0.332
52	482.3	1387.4	0.365	0.002	0.316	0.007	0.190
53	24.1	22.7	0.365	0.010	0.438	0.046	0.237
56	712.1	3005.3	0.314	0.001	0.322	0.007	0.303
58	292.9	1215.9	0.414	0.003	0.362	0.009	0.232
59	162.2	477.7	0.255	0.004	0.411	0.037	0.327
60	130.3	339.0	0.422	0.003	0.467	0.013	0.357
62	56.9	34.4	0.197	0.002	0.367	0.030	0.387
63	1078.9	6589.4	0.328	0.002	0.210	0.006	0.135
64	41.3	39.4	0.355	0.002	0.645	0.012	0.574
66	87.6	633.8	0.497	0.014	0.545	0.043	0.412
68	13.1	0.8	0.142	0.001	0.123	0.020	0.250
70	10.2	2.1	0.297	0.005	0.290	0.028	0.302
72	13.2	0.4	0.115	0.001	0.001	0.015	0.093
80	13.8	0.8	0.179	0.001	0.328	0.014	0.301
91	18.8	1.1	0.201	0.001	0.186	0.006	0.258
98	105.2	336.9	0.448	0.006	0.429	0.017	0.309
99	233.1	880.5	0.455	0.005	0.424	0.016	0.271
112	8.4	0.4	0.272	0.001	0.445	0.011	0.458
115	20.9	16.4	0.405	0.009	0.405	0.032	0.181
118	182.6	263.7	0.333	0.002	0.194	0.007	0.109
124	67.9	60.0	0.391	0.002	0.404	0.008	0.288
126	6.7	0.2	0.129	0.001	0.203	0.025	0.372
128	138.1	325.6	0.444	0.003	0.423	0.011	0.221
131	608.3	1410.4	0.232	0.001	0.114	0.006	0.078
134	191.6	1392.2	0.333	0.008	0.517	0.053	0.374
136	112.5	234.8	0.313	0.003	0.566	0.023	0.402
138	541.4	1849.7	0.266	0.002	0.407	0.017	0.352
164	257.5	320.1	0.221	0.002	0.422	0.021	0.333
165	646.5	8181.4	0.308	0.006	0.074	0.019	0.056
168	194.9	297.0	0.275	0.002	0.154	0.010	0.075
169	574.6	10117.1	0.323	0.008	0.293	0.039	0.225
172	910.5	11268.6	0.257	0.003	0.173	0.018	0.326
173	256.2	1375.3	0.346	0.005	0.328	0.021	0.284
174	474.4	2835.3	0.321	0.003	0.197	0.012	0.142

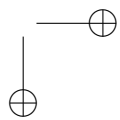
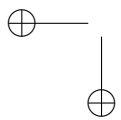
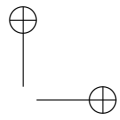
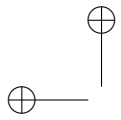
A.2. Hydrological data for flow duration curves modelling

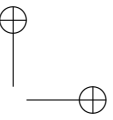
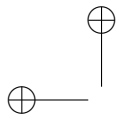
Flow duration curves analyzed in chapter 4 are based on a data set of daily discharge data available for some basins located in Italy and Switzerland. The following table provides a short summary of data consistency.

Code	River	Station	Begin year	End year	Complete years
67	Artanavaz	St.Oyen	1952	1967	16
68	Ayasse	Champorcher	1950	1953	22
94	Borbera	Baracche	1942	1961	14
93	Bormida	Cassine	1947	1958	12
77	Chisone	Fenestrelle	1942	1951	8
78	Chisone	S.Martino	1942	1971	29
76	Chisone	Soucheres Basses	1959	1970	12
87	Corsaglia	Presa C.Molline	1942	1959	18
66	Dora Baltea	Aosta	1942	1955	10
71	Dora Baltea	Tavagnasco	1951	1986	36
74	Dora Riparia	Oulx	1943	1956	10
75	Dora Riparia	S.Antonino di Susa	1942	1953	10
92	Erro	Sassello	1945	1960	16
70	Evancon	Champoluc	1949	1978	30
85	Gesso	Entraque	1952	1964	12
80	Grana	Monterosso	1942	1975	32
69	Lys	Gressoney St.Jean	1942	1953	7
61	Mastallone	Ponte Folle	1942	1965	22
72	Orco	Pont Canavese	1942	1975	29
79	Po	Crissolo	1943	1973	28
82	Rio Bagni	Bagni Vinadio	1942	1956	11
81	Rio del Piz	Pietraporzio	1942	1956	15
64	Rutor	Promise	1942	1967	20
65	Savar	Eau Rousse	1944	1962	17
95	Scriveria	Serravalle	1942	1963	14
62	Sesia	Campertogno	1942	1950	7
63	Sesia	Ponte Aranco	1942	1950	9
84	Stura di Demonte	Gaiola	1942	1965	11
83	Stura di Demonte	Pianche	1942	1955	14
73	Stura di Lanzo	Lanzo	1942	1981	38
90	Tanaro	Farigliano	1942	1985	40
89	Tanaro	Nucetto	1935	1965	29
88	Tanaro	Ponte Nava	1936	1968	30
60	Toce	Candoglia	1943	1964	21
86	Vermenagna	Limone	1942	1956	15
91	Tanaro	Montecastello	1942	1985	38
1	Broye	Payerne	1921	2000	80
2	Emme	Emmenmat	1909	2000	92
3	Ltschine	Gsteig	1920	2000	81
4	Grbe	Belp.Stockmatt	1923	2000	78
5	Sense	Thrishaus	1928	2000	73
6	Emme	Eggiwil	1931	1974	44
7	Weisse Ltschine	Zweiltschinen	1933	2000	68
8	Simme	Oberried	1949	2000	52
9	Allenbach	Adelboden	1950	2000	51
10	Gornernbach	Kiental	1950	1982	33
11	Biberenkanal	Kerzers	1956	2000	45
12	Langeten	Huttwil	1966	2000	35
13	Langeten	Lotzwil	1969	1993	24
14	Mentue	Yvonand	1971	2000	30
15	Orbe	Le Chenit	1971	2000	30
16	Poschiavino	La Rsa	1970	2000	31
17	Poschiavino	Le Prese	1931	2000	70

A.2 Hydrological data for flow duration curves modelling 117

Code	River	Station	Begin year	End year	Complete years
18	Aach	Salmsach	1962	2000	39
19	Albula	Tiefencastel	1921	2000	80
20	Birse	Moutier	1912	2000	99
21	Dischmabach	Davos	1964	2000	37
22	Ergolz	Liestal	1934	2000	67
23	Goldach	Goldach	1962	2000	39
24	Hinterrhein	Andeer	1923	1961	39
25	Hinterrhein	Hinterrhein	1945	2000	56
26	Landquart	Felsenbach	1921	2000	80
27	Landquart	Klosters	1933	1974	42
28	Landwasser	Davos	1967	2000	34
29	Murg	Wngi	1954	2000	67
30	Plessur	Chur	1931	2000	70
31	Sitter	Bernhardzell	1924	1980	57
32	Somvixer Rhein	Somvix	1977	2000	24
33	Steinach	Steinach	1962	2000	39
34	Thur	Jonschwilen	1966	2000	35
35	Thur	Stein	1964	2000	37
36	Tss	Neftenbach	1921	2000	80
37	Urnsch	Hundwil	1962	2000	39
38	Werdenberger Bin- nenkanal	Salez	1931	2000	70
39	Saaser Vispa	Zermeiggern	1923	1963	41
40	Borgne	La Luette	1926	1979	54
41	Baye de Montreux	Les Avants	1933	1974	42
42	Venoge	Lussery	1948	1978	31
43	Baye de Montreux	Montreux	1933	1973	41
44	Bavona	Bignasco	1929	1966	38
45	Reuss	Andermatt	1910	2000	91
46	Grosstalbach	Isenthal	1957	2000	44
47	Alpbach	Erstfeld	1960	2000	41
48	Engelberger Aa	Engelberg	1955	1990	36
49	Engelberger Aa	Buochs	1983	2000	18
50	Witenwasserreuss	Realp	1957	1986	30
51	Roseggbach	Pontresina	1955	2000	46
52	Berninabach	Pontresina	1955	2000	46
53	Ova da Cluozza	Zernez	1962	2000	39
54	Chamuerabach	La Punt Chamues	1972	2000	29
55	Biber	Ramsen	1942	1983	42
56	Seez	Weisstannen	1959	1991	33
57	Minster	Euthal	1961	2000	40
58	Steinenbach	Kaltbrunn	1968	2000	33
59	Ticino	Bellinzona	1942	1951	10





Appendix B.

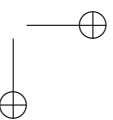
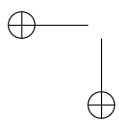
Morpho-climatic information

B.1. Parameters description

This appendix lists and describes the morphometric and climatic catchment parameters in use in the CUBIST Information System and adopted in this thesis. All of them can be computed automatically using GIS tools, using procedures developed in the Linux "bash" scripting language, that exploit together the "GRASS" GIS and the "Fluidturtle" libraries available at:
<http://www.ing.unitn.it/~rigon/indexo.html>.

The "R" statistical computing software has been also used for the computation of statistical indices. The choice of open source software, under the GNU General Public License, has been determined by the fact that all these packages are constantly updated and improved by experts of the international scientific community. Following this philosophy our script is open, easily customizable, and available at the address:
www.idrologia.polito.it/~alviglio/software/GRASSindex.htm.

The digital terrain model used for the Italian basins is the DEM SRTM (Shuttle Radar Topography Mission) released by NASA in 2000 and downloadable at:
<http://edc.usgs.gov>.



B.1.1. Geomorphological parameters

For each drainage basin, morphological parameters were calculated operating on a Digital Elevation Model (DEM) with the "Fluidturtle" libraries. These libraries also provide tools for DEM analysis like the pit removal (to ensure hydraulic connectivity within the watershed), the computation of flow directions, the delineation of channel networks and much more. The geomorphological descriptors considered are:

- X, Y [m]: coordinates of the gauging station.
- X_c, Y_c [m]: coordinates of the centroid of the plane projection of the drainage basin.
- A [km²]: area of the plane projection of the drainage basin.
- P [km]: basin perimeter.
- H_{max}, H_{min}, H [m]: maximum, minimum and mean elevation of the drainage basin above sea level.
- MA [deg]: mean geometric (vector) aspect calculated as the average of the aspect of each cell. The aspect is the direction towards which a slope faces and is important in hilly or mountainous terrain. Here it is defined as the angle of exposure of the cell (computed from the north).
- ΔH_1 [m]: difference between the maximum and the minimum elevation of the cells belonging to the basin.
- ΔH_2 [m]: difference between the mean and the minimum elevation of the basin.
- L_{OV} [km]: length of the segment joining the basin centre of mass to the basin outlet.
- O_{OV} [deg]: angle between the orientation vector and the north.

- p_m [-]: average of the slope values associated to each pixel in the DEM of the drainage basin.
- P_m [-]: mean large-scale slope, computed as $2(H - H_{min}/\sqrt{A})$. P_m is a slope measure of a square equivalent basin, and does not account for basin shape; its definition is objective, i.e. not affected by the DEM resolution.
- $IPSSx$: area-elevation curve (hypsometric curve) percentile, i.e. the curve that represents the portion of the basin area located above a given elevation. The curve is represented recording elevations corresponding to the 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95% and 97.5% of the area.
- R_c [-]: circularity ratio, i.e. the ratio between the basin area and the area of a circle having the same perimeter, equal to $4A\pi/P^2$.
- C_c [-]: compactness (Gravelius) coefficient, i.e. the ratio between the perimeter of the basin and the diameter of the equivalent circle, equal to $P/(2\sqrt{A/\pi})$.

B.1.2. River network parameters

Selected analyses can be performed on the river network, that is automatically extracted from the DEM, using the above-described drainage directions and the following constraints: (i) a pixel belongs to the network if its contributing area exceeds 1 km²; (ii) a stream belongs to the network if it is composed of more than one pixel. The river network descriptors considered are:

- MSL [km]: main stream length, i.e. the length of the longest series of streams that connects the basin outlet to the foremost source point (i.e. the upper stream end).

- *LLDP* [km]: length of the longest drainage path, i.e. the longest path between the basin outlet and the most distant point on the basin border, following drainage directions. Actually the longest drainage path corresponds to the main stream plus the path on the hillslope that connects the stream source to the basin border.
- *SLDP* [km]: slope of the longest drainage path PLDP computed as the average of the slope values associated to each pixel in the longest drainage path.
- *R_{al}* [-]: elongation ratio, i.e. the ratio between the diameter of a circle with area equivalent to the basin area and the length of the longest drainage path ($2\sqrt{A/\pi}/LLDP$).
- *F_f* [-]: shape factor, i.e. the ratio between the basin area and the square of the longest drainage path length ($A/LLDP^2$).
- *FA* [m]: width function; moments (mean, variance, skewness and kurtosis) and percentiles of the width function, which is defined as the cumulated frequency of the pixel metric distance from the basin outlet.
- *MHL* [m]: mean hillslope length, i.e. the average distance (throughout all the basin) between pixels and channel .
- *M* [-]: magnitude, i.e. number of source points of the network.
- *T_d* [-]: topological diameter, i.e. the number of links that constitute the main stream, or number of confluences to the main stream.
- Horton-Strahler ordering, i.e. number of links, average length, average contributing area and mean slope corresponding to every Horton class. These classes form an ordering classification system in which channel segments are ordered numerically from a stream's headwaters to the basin outlet. Numerical ordering begins with the tributaries at the

stream's headwaters being assigned the value 1. A stream segment that results from the joining of two 1st order segments is given an order 2. Two 2nd order streams form a 3rd order stream, and so on.

- Rh_b, Rh_l, Rh_a, Rh_s [-]: Horton ratios, i.e. the slope of the interpolation straight line (computed with the Ordinary Least Squares method) between the points given by the order and the variable (number of links, average length, average contributing area and mean slope) on a semilogarithmic diagram.
- TNL [km]: total network length, i.e. the sum of the lengths of all stream within the basin.
- D_d [km/km²]: drainage density i.e. the measure of the length of stream channel per unit area of drainage basin. Mathematically it is expressed as the total network length divided by the area of the drainage basin.

B.1.3. Soil use and permeability parameters

Five soil use indexes are defined pooling similar land-cover classes defined in the CORINE project (COoRdination of INformation on Environment, European Commission [1985]):

- LC_1 : percentage of basin area with urban areas.
- LC_2 : percentage of basin area with forests, woodlands or shrubs.
- LC_3 : percentage of basin area with grasslands or cultivated lands.
- LC_4 : percentage of basin area without vegetation.
- LC_5 : percentage of basin area with wetlands.

Moreover, permeability indexes available are:

- CN [-]: curve number related to soil permeability. The CN relative to the whole basin is calculated as the average of all the cell-values.
- c_f [-]: permeability index from VAPI Piemonte.

B.1.4. Climatic parameters

Climatic parameters are firstly computed for each raingauge station available, then the results are interpolated by means of a suitable kriging procedure over the whole area of interest in order to create a raster with the DEM resolution. The parameters at the basin scale are the average of the cell-values over the catchment area, in particular:

- a [mm/h]: coefficient of the intensity-duration-frequency (IDF) in the form $h = a \cdot d^n$.
- n [-]: coefficient of the IDF curve in the form $h = a \cdot d^n$.
- MAP [mm]: mean annual precipitation.

B.2. Parameters list

Parameters involved in the analysis are listed in the following tables.

Code Chap.	Code Chap.	River	Station	X	Y
1	67	Artanavaz	St.Oyen	360183	5075749
2	68	Ayasse	Champorcher	392518	5052993
3	94	Borbera	Baracche	500626	4951965
6	93	Bormida	Cassine	463745	4954630
7	-	Bormida di Mallare	Ferrania	446013	4912537
8	-	Cervo	Passobreve	424990	5053388
9	77	Chisone	Fenestrelle	346024	4989051
10	78	Chisone	S.Martino	363869	4971770
11	76	Chisone	Soucheres Basses	338658	4987848
12	87	Corsaglia	Presa C.Molline	407071	4904931
13	66	Dora Baltea	Aosta	371847	5065981
14	-	Dora Baltea	Ponte di Mombardone	344228	5069841
15	71	Dora Baltea	Tavagnasco	408854	5043555
16	-	Dora di Bardonecchia	Beaulard	324755	4990299
19	74	Dora Riparia	Oulx	329119	4988888
20	75	Dora Riparia	S.Antonino di Susa	362906	4996791
22	92	Erro	Sassello	456328	4926986
24	70	Evancon	Champoluc	400630	5075832
25	85	Gesso	Entraque	371058	4901682
26	-	Gesso della Valletta	S. Lorenzo	368064	4901375
27	80	Grana	Monterosso	367078	4918842
29	-	Lys	D'Ejola	407799	5078900
30	69	Lys	Gressoney St.Jean	408449	5071177
31	61	Mastallone	Ponte Folle	442083	5075487
33	72	Orco	Pont Canavese	391455	5030182
35	79	Po	Crissolo	354451	4950954
38	82	Rio Bagni	Bagni Vinadio	347105	4906010
39	81	Rio del Piz	Pietraporzio	343044	4911675
40	64	Rutor	Promise	341062	5062853
41	-	San Bernardino	Santino	462832	5089696
42	65	Savar	Eau Rousse	360242	5047894
43	-	Scrvia	Isola del Cantone	496450	4943465
44	95	Scrvia	Serravalle	488864	4952192
45	62	Sesia	Campertogno	424653	5072656
47	63	Sesia	Ponte Aranco	444885	5061708
48	-	Sesia	Vercelli	455947	5019480
50	84	Stura di Demonte	Gaiola	373774	4910281
51	83	Stura di Demonte	Pianche	349455	4907337
52	73	Stura di Lanzo	Lanzo	380917	5014005
53	-	Stura di Vi	Usseglio	359557	5010289
56	90	Tanaro	Farigliano	412701	4929373

Code Chap. 2	Code Chap. 4	River	Station	X	Y
58	89	Tanaro	Nucetto	425280	4910519
59	-	Tanaro	Ormea	413186	4889264
60	88	Tanaro	Ponte Nava	410758	4885835
62	-	Toce	Cadarese	450467	5125992
63	60	Toce	Candoglia	455247	5090657
64	-	Varaita	Rore	359268	4937167
65	86	Vermenagna	Limone	385914	4896184
66	-	Vobbia	Vobbietta	499022	4941987
68	-	Breuil	Alpette	336956	5064247
70	-	Chiavanne	Alpette	336855	5064354
72	-	Dora di Rhemes	Notre Dame	353255	5048250
80	-	Rutor	La Joux	341259	5061743
91	-	Varaita	Castello	344947	4941946
98	-	Lys	Guillemore	411049	5058165
99	-	Chiusella	Gurzia	402645	5030981
112	-	Stura di Viu	Malciaussia	354250	5007748
115	-	Marmore	Perreres	392559	5084867
118	-	Sermenza	Rimasco	427245	5078630
124	-	Maira	S.Damiano Macra	361400	4927343
126	-	Maira	Saretto	335555	4927344
128	-	Bormida Spigno	Valla	447554	4932222
131	-	Adda	Fuentes	534567	5110177
134	-	Adda	Tirano	589647	5118637
136	-	Aveto	Cabanne	527874	4927491
138	-	Brembo	P.te Briolo	547267	5067317
164	-	Serio	P.te Cene	564310	5071039
165	-	Taro	Pradella	559398	4925585
168	-	Taro	Carniglia	548372	4925569
169	-	Taro	Ostia	567898	4930371
172	-	Trebbia	S.Salvatore	530364	4955188
173	-	Trebbia	Due Ponti	520769	4931707
174	-	Trebbia	Valsigiara	524861	4944296
-	91	Tanaro	Montecastello	475365	4976584
-	1	Broye	Payerne	341678.82	5187379.5
-	2	Emme	Emmenmat	405055.84	5200949.32
-	3	Ltschine	Gsteig	413507.53	5168644.57
-	4	Grbe	Belp.Stockmatt	386142.18	5195614.03
-	5	Sense	Thrishaus	374532.98	5194211.92
-	6	Emme	Eggiwil	408945.88	5191673.2
-	7	Weisse Ltschine	Zweitschinen	415929.95	5164886.83
-	8	Simme	Oberried	382980.56	5142824.48
-	9	Allenbach	Adelboden	388856.62	5149027.17
-	10	Gornernbach	Kiental	404742.81	5155611.92
-	11	Biberenkanal	Kerzers	361087.64	5203478.25
-	12	Langeten	Huttwil	411071.17	5219604.22
-	13	Langeten	Lotzwil	408105.36	5226891.55
-	14	Mentue	Yvonand	326474.56	5183605.91
-	15	Orbe	Le Chenit	282476.46	5159373.82
-	16	Poschiavino	La Rsa	583786.33	5146344.39
-	17	Poschiavino	Le Prese	583689.81	5126419.11
-	18	Aach	Salmsach	527168.29	5266841.29
-	19	Albula	Tiefencastel	543746.86	5167664.72
-	20	Birse	Moutier	377674.11	5238045.67
-	21	Dischmabach	Davos	563075.44	5178704.93
-	22	Ergolz	Liestal	404594.43	5260193.31
-	23	Goldach	Goldach	536616.41	5260185.49
-	24	Hinterrhein	Andeer	533294.88	5163794.49
-	25	Hinterrhein	Hinterrhein	515823.8	5153103.62
-	26	Landquart	Felsenbach	545743.06	5202507.94
-	27	Landquart	Klosters	569958.38	5189883.77
-	28	Landwasser	Davos	560465.51	5178635.51
-	29	Murg	Wngi	494877.41	5262651.61
-	30	Plessur	Chur	539055.53	5190131.96
-	31	Sitter	Bernhardzell	523745.55	5260817.42
-	32	Somvixer Rhein	Somvix	498911.63	5166931.46
-	33	Steinach	Steinach	534018.82	5262795.84
-	34	Thur	Jonschwilen	505491.55	5252988.28
-	35	Thur	Stein	517455.92	5226883.43
-	36	Tss	Neftenbach	475866.25	5262350.81
-	37	Urnsch	Hundwil	522222.17	5243236.02
-	38	Werdenberger Binnenkanal	Salez	537432.01	5231115.04
-	39	Saaser Vispa	Zermeiggern	419441.31	5103655.07
-	40	Borgne	La Luette	379767.13	5113026.68
-	41	Baye de Montreux	Les Avants	342788.58	5146263.86
-	42	Venoge	Lussery	311079.02	5166958.33
-	43	Baye de Montreux	Montreux	340788.21	5144895.57
-	44	Bavona	Bignasco	469486.01	5132565.02
-	45	Reuss	Andermatt	468306.37	5166779.56
-	46	Grosstalbach	Isenthal	466700.26	5195389.14
-	47	Alpbach	Erstfeld	469120.12	5184402.43
-	48	Engelberger Aa	Engelberg	453279.14	5185302.95
-	49	Engelberger Aa	Buochs	452208.55	5201145.89
-	50	Witenwasserrenreuss	Realp	461382.99	5159557.34
-	51	Roseggbach	Pontresina	568396.94	5149838.91
-	52	Berninabach	Pontresina	570836.37	5147190.7
-	53	Ova da Chuoza	Zernez	585473.7	5171709.8
-	54	Chamuerabach	La Punt Chamues	571416.71	5158294

Code Chap. 2	Code Chap. 4	River	Station	X	Y
-	55	Biber	Ramsen	485513.6	5283770.16
-	56	Seez	Weisstannen	526895.43	5205191.04
-	57	Minster	Euthal	485858.98	5215018.7
-	58	Steinenbach	Kaltbrunn	503324.13	5228593.51
-	59	Ticino	Bellinzona	502772.8	5117525.2

Code Chap. 2	Code Chap. 4	X_c	Y_c	A	H	H_{min}	P	ΔH_1
1	67	356250	5076750	69.2	2230	1351.0	38.0	1816
2	68	387350	5052050	41.9	2364	1372.0	31.0	1747
3	94	508450	4946050	202.2	869	359.0	69.0	1313
6	93	445350	4926450	1514.4	494	117.0	237.0	1259
7	-	444150	4905150	50.3	611	361.0	42.0	669
8	-	420050	5058650	75.5	1493	586.0	41.0	1922
9	77	339150	4985050	154.1	2152	1153.0	71.0	2075
10	78	348550	4980550	580.5	1734	417.0	126.0	2811
11	76	337050	4982050	92.9	2227	1506.0	49.2	1722
12	87	406050	4897450	89.4	1529	643.0	46.0	1950
13	66	358150	5064450	1846.4	2262	547.0	296.0	4180
14	-	339850	5071550	372	2402	1011.0	115.0	3716
15	71	374950	5065150	3320	2085	259.0	350.0	4468
16	-	319450	4995650	207.4	2187	1122.0	80.0	2206
19	74	329950	4977450	257.5	2168	1079.0	87.0	2179
20	75	335150	4993350	1037.9	1898	385.0	254.0	3192
22	92	457350	4922150	92.1	600	324.0	60.0	922
24	70	402150	5080650	102.4	2635	1553.0	51.0	2604
25	85	372250	4893250	159.6	1886	815.0	61.0	2363
26	-	362050	4896750	110.9	2096	912.0	53.0	2227
27	80	360050	4917850	109.6	1534	711.0	54.0	1896
29	-	408150	5083050	29.6	3110	1812.0	25.0	2615
30	69	408750	5078650	90.4	2637	1398.0	46.0	3029
31	61	437950	5081750	146.6	1324	492.0	63.0	1925
33	72	376250	5036250	613.4	1928	429.0	145.0	3406
35	79	350350	4950750	37.3	2240	1294.0	27.0	2410
38	82	344150	4903350	61.2	2138	1257.0	35.0	1675
39	81	341550	4908350	21.6	2193	1273.0	23.0	1707
40	64	341750	5059450	45	2554	1508.0	35.0	1906
41	-	457650	5097850	121.2	1253	280.0	57.0	1965
42	65	359550	5042650	81.3	2694	1650.0	43.0	2266
43	-	502150	4933150	215.9	668	297.0	86.2	1246
44	95	502950	4941650	613.8	687	203.0	128.0	1469
45	62	416950	5076450	170.7	2113	828.0	69.0	3643
47	63	429150	5075850	702.9	1496	328.0	142.0	4143
48	-	433950	5055350	2187.9	838	119.0	251.8	4352
50	84	351050	4908650	560	1817	661.0	141.0	2319
51	83	340750	4913250	179.7	2073	972.0	73.0	2008
52	73	365150	5016450	578.4	1769	461.0	123.0	3172
53	-	355050	5010150	79.5	2375	1280.0	44.0	2237
56	90	407850	4905250	1497.2	948	239.0	199.0	2386
58	89	411850	4892350	374.9	1224	453.0	135.0	2159
59	-	402950	4886550	175.5	1512	719.0	71.0	1893
60	88	401450	4886350	148.5	1569	788.0	63.0	1824
62	-	453250	5135850	187	2144	730.0	70.0	2563
63	60	439950	5110850	1539.4	1671	203.0	263.0	4285
64	-	346850	4940850	278	2111	837.0	97.0	2863
65	86	385850	4892550	57.3	1684	963.0	37.0	1750
66	-	503650	4939050	50.6	732	355.0	35.0	1026
68	-	333550	5063850	28.4	2444	1793.0	27.1	1269
70	-	334250	5067450	21.7	2483	1793.0	23.7	1383
72	-	351250	5043050	69	2664	1717.0	41.2	1839
80	-	341750	5059150	41.4	2587	1600.0	35.4	1814
91	-	341750	4945750	67.4	2400	1575.0	37.2	1663
98	-	409850	5070550	202.9	2247	911.0	90.1	3516
99	-	399050	5039750	142.1	1358	424.0	59.6	2341
112	-	351950	5008750	25.8	2598	1792.0	24.4	1609
115	-	394650	5088350	54.8	2719	1844.0	36.0	2529
118	-	424850	5082150	82.1	1844	902.0	39.9	1975
124	-	345950	4927450	452.1	1892	716.0	110.7	2447
126	-	333150	4931150	54.9	2408	1539.0	33.4	1624
128	-	449750	4925250	68.5	471	261.0	54.0	576
131	-	578650	5124850	2578.5	1859	201.0	343.0	3762
134	-	602650	5139350	908.1	2167	425.0	169.0	3332
136	-	524250	4925350	39.7	989	815.0	38.0	497
138	-	551050	5085050	748.6	1187	260.0	149.0	2579
164	-	571150	5087150	459.8	1336	365.0	129.0	2592
165	-	550150	4923850	297.3	836	412.0	99.0	1275
168	-	541650	4922150	90.3	967	538.0	46.0	1149
169	-	553950	4924550	412	821	335.0	120.0	1352
172	-	526250	4938150	638	951	287.0	140.0	1492
173	-	516550	4930750	74.5	964	642.0	47.0	918
174	-	520950	4935050	222.7	943	442.0	75.0	1174
-	91	425050	4932750	7982.8	656	82	626	3096
-	1	335295.14	5167304.17	414.57	750	444	148.75	1537

Code Chap. 2	Code Chap. 4	X_c	Y_c	A	H	H_{min}	P	ΔH_1
-	2	412425.13	5192325.31	443.79	1076	639	126.39	1539
-	3	420434.26	5160015.76	359.97	2050	583	99.52	3486
-	4	385515.4	5183053.94	120.24	849	518	76.21	1613
-	5	375165.13	5180625.19	349.72	1078	561	113.16	1603
-	6	415934.56	5184764.57	127.65	1293	743	67.79	1435
-	7	415213.81	5155964.72	164.96	2122	654	60.61	3415
-	8	385064.34	5140214.82	35.43	2347	1116	29.53	2110
-	9	386324.49	5147594.92	28.93	1862	1335	24.64	1392
-	10	407114.67	5152995.21	25.3	2314	1334	22.48	2278
-	11	361484.77	5197724.5	48.51	543	433	45.63	253
-	12	409905.65	5215094.22	60.02	764	602	41.07	497
-	13	409004.49	5217973.72	115.35	715	500	56.06	599
-	14	324584.35	5174506.35	100.93	667	442	69.56	491
-	15	278324.44	5154254.85	47.81	1246	1094	39.07	456
-	16	581895	5144895.88	12.83	2457	1974	17.46	1034
-	17	581805	5134185	169.64	2118	977	78.45	2872
-	18	520154.03	5267205.29	47.93	475	406	42.81	169
-	19	559935.7	5171805.1	529.77	2130	1001	128.25	2236
-	20	371476.35	5234354.41	181.81	925	548	94.66	869
-	21	565153.81	5174414.71	43.74	2344	1630	34.57	1367
-	22	409544.58	5253795.6	259.82	587	306	87.81	855
-	23	535004.47	5252713.4	51.36	832	402	42.71	843
-	24	529606.96	5149755.26	509.42	2247	1087	162.02	2248
-	25	510255.84	5150743.82	55.18	2344	1602	40.66	1733
-	26	561194.09	5196915.59	616.47	1799	560	151.38	2779
-	27	575954.5	5187735.78	112.55	2298	1250	56	2089
-	28	566326.12	5179005.4	185.82	2225	1500	73.22	1601
-	29	498463.01	5255596.39	75.19	625	446	69.51	577
-	30	549044.49	5183595.85	265.64	1861	558	83.79	2386
-	31	527624.62	5243623.18	326.77	1007	497	112.52	1949
-	32	499634.78	5163794.89	22.73	2402	1388	26.01	1682
-	33	530143.79	5254061.93	21.58	716	402	38.58	681
-	34	512054.64	5236514.9	500.45	1022	538	129.13	1867
-	35	522045.62	5226344.57	84.49	1448	845	49.01	1560
-	36	487215.22	5252806.23	391.08	656	407	119.52	885
-	37	523034.85	5238676.1	65.58	1107	776	42.32	1612
-	38	532846.26	5223915.04	185.77	1029	436	75.69	1879
-	39	418454.56	5098453.93	65.99	2847	1736	35.32	2424
-	40	384343.86	5101784.76	230.55	2549	971	76.36	2902
-	41	343754.57	5147054.74	6.99	1417	933	12.85	887
-	42	302173.61	5165775.93	160.92	813	441	78.85	1220
-	43	342855.42	5146423.66	13.53	1224	609	17.36	1211
-	44	463363.44	5137873.95	120.75	1934	453	55.31	2775
-	45	465165.04	5160554.47	193.83	2265	1264	76.62	2234
-	46	462734.91	5192595.24	43.44	1808	810	35.97	2110
-	47	466425.34	5183144.76	20.68	2217	1089	21.5	2014
-	48	459405.18	5184765.65	86.79	1955	997	42.43	2117
-	49	455803.71	5188903.95	230.99	1602	447	77.74	2667
-	50	459945.65	5156054.69	30.85	2421	1577	25.76	1482
-	51	566505.19	5140933.49	67.09	2672	1752	43.75	2126
-	52	574424.37	5142285.94	108.03	2609	1837	50.82	2028
-	53	585404.42	5166314.08	27.17	2362	1535	25.5	1584
-	54	575954.77	5152544.51	74.2	2540	1702	40.67	1495
-	55	479475.6	5291774.82	163.52	552	410	77.53	427
-	56	523214.14	5202135.15	72.64	1901	952	41.56	1861
-	57	483795	5209514.42	63.78	1326	886	36.25	1364
-	58	506655.88	5228685.81	18.96	1127	500	22.85	1400
-	59	496847.49	5139585.37	1516.03	1682	232	275.35	3126

Code Chap. 2	Code Chap. 4	MSL	MHL	P_m	P_m	D_d	a	n	MAP	c_f
1	67	10.5	783.6	49.05	21.76	0.52	9.5	0.54	1080.4	0.33
2	68	10.7	669.1	40.02	32	0.65	16.4	0.59	1059.4	0.48
3	94	24.1	699.2	34.24	6.64	0.56	30.7	0.43	1331.5	0.6
6	93	133.8	641.2	20.55	1.76	0.66	24.7	0.44	1031.3	0.53
7	-	17.7	606.9	24.44	6.4	0.65	27.9	0.46	1446.2	0.6
8	-	14	715.9	52.26	20.37	0.55	27.9	0.54	1808.2	0.51
9	77	25.1	862.9	44.87	16.45	0.56	11.4	0.56	887.2	0.38
10	78	55.8	776.9	46.66	11.21	0.57	15.3	0.54	1069.6	0.4
11	76	15.9	870.6	43.05	15.13	0.52	11.6	0.56	900	0.37
12	87	17.6	633.1	44.54	18.3	0.62	21.1	0.51	1349.6	0.24
13	66	57.5	856.5	51.88	8.33	0.6	10.5	0.54	902.3	0.33
14	-	22.2	865.7	53.05	14.37	0.61	8.9	0.55	854.6	0.4
15	71	117.1	845.4	50.65	6.61	0.6	12.2	0.55	935.6	0.43
16	-	22.8	817.6	47.1	15.36	0.54	11.4	0.52	856.2	0.23
19	74	34.5	850.5	42.89	13.61	0.52	11.5	0.55	906.3	0.29
20	75	79.3	853	44.77	9.87	0.62	12.6	0.53	902.5	0.23
22	92	19.1	628.4	20.2	5.17	0.71	31.2	0.47	1394.8	0.61
24	70	12.4	794.8	47.23	20.47	0.61	11.5	0.59	944.8	0.53
25	85	17	788.6	60	17.32	0.53	18.5	0.53	1535.5	0.31
26	-	15.2	757.4	63	23.71	0.53	14.8	0.54	1427.3	0.37
27	80	19.6	683.9	46	14.38	0.56	16.2	0.5	1117.8	0.27
29	-	7	1023.7	49	45.86	0.58	11.8	0.62	1074.7	0.2

Code Chap.	Code Chap.	<i>MSL</i>	<i>MHL</i>	<i>p_m</i>	<i>P_m</i>	<i>D_d</i>	<i>a</i>	<i>n</i>	<i>MAP</i>	<i>c_f</i>
2	4									
30	69	15.6	847.5	52.09	25.03	0.6	12.8	0.6	1071.4	0.26
31	61	22.1	636.3	52.75	13.15	0.58	28	0.58	2047.4	0.59
33	72	51	796.2	54.46	12.23	0.57	20.5	0.55	1290.1	0.52
35	79	8.3	737.2	45.87	31.96	0.61	16.4	0.55	1229	0.52
38	82	8.3	745.4	53.03	23.45	0.53	13.6	0.53	1308.7	0.3
39	81	7.2	665.8	52.93	41.08	0.62	13.5	0.53	1173.5	0.29
40	64	8.5	719	38.13	31.62	0.74	9.5	0.52	905.6	0.3
41	-	22.7	690	57.31	18.02	0.57	30.2	0.58	2328.6	0.6
42	65	10.9	865.4	46.15	23.45	0.6	13	0.57	1011.9	0.39
43	-	37.8	631	32.06	4.67	0.6	37.1	0.43	1741.5	0.64
44	95	53.5	663.5	31.1	3.54	0.59	32.8	0.42	1454	0.61
45	62	20.3	807.7	59.57	19.82	0.58	16.4	0.6	1357.1	0.44
47	63	61.4	723.5	53.11	8.4	0.57	24.7	0.57	1775.4	0.54
48	-	114.8	722.9	27.49	1.92	0.76	29.7	0.44	1534.2	0.5
50	84	56.3	760.9	48.56	10.47	0.56	14.4	0.52	1162.5	0.26
51	83	25.6	758.5	50.15	17.4	0.54	13.4	0.53	1088.3	0.29
52	73	39.8	818	49.79	10.71	0.56	20.2	0.53	1270	0.58
53	-	13.3	840.8	56.53	25.33	0.54	14.3	0.56	1172.2	0.42
56	90	100.5	679	27.68	2.78	0.69	22.6	0.45	1191.8	0.47
58	89	55.8	707.8	38.04	7.22	0.6	22	0.5	1234.2	0.39
59	-	24.4	700.7	42.08	11.17	0.59	22.2	0.52	1247.9	0.36
60	88	19.8	708.7	42.62	12.01	0.58	22.2	0.52	1250.8	0.37
62	-	29.3	784.9	53.89	22.11	0.6	15.8	0.6	1527.7	0.42
63	60	82.6	813	54.67	7.75	0.57	18	0.61	1610.8	0.51
64	-	29.6	796.9	46	15.64	0.61	15.1	0.53	1057.7	0.28
65	86	9.7	703.3	43.39	18.73	0.57	21.6	0.51	1462.6	0.37
66	-	12.8	660.7	37.23	10.01	0.53	34.3	0.43	1455.4	0.52
68	-	8.9	821.8	36.81	25.35	0.51	11	0.52	1084.2	0.27
70	-	8.4	775.5	51.53	31.15	0.52	11.3	0.52	1119.2	0.32
72	-	12.2	831.7	49.1	23.83	0.6	11	0.56	965.5	0.2
80	-	7.1	719.2	35.84	30.71	0.75	9.4	0.53	905.7	0.29
91	-	11.4	787.8	47.48	20.74	0.55	15.2	0.55	1110.5	0.26
98	-	32.3	803.9	54.79	18.31	0.6	14.3	0.59	1171.3	0.5
99	-	24.8	761.5	42.03	14.77	0.57	24.5	0.5	1489.5	0.5
112	-	6.2	834	55.6	32.01	0.55	12.8	0.57	1137.6	0.33
115	-	9.9	941.1	48.05	23.5	0.57	9.7	0.6	1015.9	0.44
118	-	10.5	733.6	60.09	20.99	0.53	19.4	0.61	1602	0.55
124	-	41.4	745.6	47.04	11.28	0.59	14.4	0.51	1007.2	0.25
126	-	10	822.9	48.83	25.46	0.61	12.5	0.54	1049.9	0.26
128	-	19.8	664.5	14.42	4.59	0.63	28	0.45	1064.6	0.52
131	-	132.6	857.5	50.65	6.94	0.58	16.4	0.45	1121	0.39
134	-	62.8	866.2	48.95	12.21	0.57	13	0.46	1164	0.39
136	-	10.6	580.8	27.76	5.11	0.64	33.6	0.44	2074.3	0.45
138	-	44	741.7	45.42	6.19	0.57	27.8	0.36	1593	0.39
164	-	45.9	745.7	47.53	8.39	0.59	25	0.39	1666	0.39
165	-	34.4	730.9	23.91	4.55	0.63	32.2	0.4	1753	0.21
168	-	20.1	656.3	30.74	8.12	0.64	35.4	0.39	2143	0.48
169	-	45.5	760.9	24.21	4.49	0.61	31.3	0.39	1622	0.35
172	-	55.4	703.9	32.13	5.19	0.62	36.3	0.43	1921.4	0.47
173	-	14.2	649.9	30.09	6.74	0.61	38.8	0.44	2105.6	0.49
174	-	33.8	680.5	30.67	6.5	0.63	38.2	0.43	2060.8	0.47
-	91	224.13	674.65	21.36	0.8	0.71				
-	1	55.62	722.28	11.81	2.77	0.72				
-	2	39.55	665.56	27.68	3.55	0.63				
-	3	26.04	940.19	56.21	14.98	0.59				
-	4	26.63	820.26	20.31	4.29	0.7				
-	5	36.16	676.13	24.21	4.34	0.64				
-	6	26.31	712.89	32.55	9.15	0.6				
-	7	21.56	945.01	58.26	22.45	0.55				
-	8	7.88	860.01	44.31	45.8	0.64				
-	9	5.7	660.07	44.99	18.74	0.55				
-	10	6.76	899.91	58.5	37.85	0.54				
-	11	19.92	667.66	6.53	3.04	0.69				
-	12	11.68	609.93	13.61	4.05	0.67				
-	13	22.56	619.59	14.15	4.06	0.69				
-	14	25.19	724.57	8.64	4.2	0.67				
-	15	15.7	841.73	14.26	3.91	0.61				
-	16	5.35	770.82	48.2	27.64	0.41				
-	17	22.03	826.44	49.53	18.21	0.6				
-	18	16.82	738.05	3.98	1.73	0.79				
-	19	44.94	779.79	46.9	10.38	0.56				
-	20	30.61	784.33	26.25	5.44	0.58				
-	21	9.96	852.77	46.99	22.41	0.45				
-	22	21.34	628.7	21.18	3.29	0.63				
-	23	18.64	698.07	17.96	13	0.68				
-	24	33.96	808.07	48.2	10.74	0.52				
-	25	14.81	811.16	51.4	20.41	0.54				
-	26	45.74	818.9	43.28	10.11	0.56				
-	27	13.87	789.44	50.92	21.15	0.54				
-	28	21.44	834.64	43.37	11.22	0.51				
-	29	22.49	664.43	12.75	3.28	0.78				
-	30	33.07	798.97	41.72	16.7	0.59				
-	31	46.9	715.02	25.18	4.89	0.64				
-	32	9.14	717.47	49.49	42.54	0.59				
-	33	17.71	584.13	13.1	12.53	0.92				
-	34	50.11	703.82	27.54	3.6	0.62				
-	35	13.46	764.95	38.5	12.58	0.6				

Code Chap.	Code Chap.	<i>MSL</i>	<i>MHL</i>	<i>pm</i>	<i>Pm</i>	<i>D_d</i>	<i>a</i>	<i>n</i>	<i>MAP</i>	<i>c_f</i>
-	36	44.95	652.02	16.07	2.21	0.69				
-	37	13.99	696.57	28.23	7.24	0.69				
-	38	27.39	834.26	30.3	7.81	0.75				
-	39	9.63	911.65	45.3	27.48	0.57				
-	40	26.88	933.56	51.19	21.36	0.59				
-	41	3.02	674.21	44.97	37.23	0.5				
-	42	30.33	876.61	12.17	4.02	0.75				
-	43	5.79	686.12	44.36	33.6	0.52				
-	44	20.82	826.15	60.96	29.36	0.55				
-	45	20.42	818.68	45.9	14.98	0.52				
-	46	11.12	709.58	57.38	29.56	0.59				
-	47	6.48	824.93	61.55	50.05	0.51				
-	48	15.86	861.56	56.72	21.9	0.54				
-	49	35.56	825.78	53.14	15.13	0.57				
-	50	8.06	796.63	42.17	31.9	0.46				
-	51	16.37	973.57	52.29	23.1	0.53				
-	52	13.39	815.37	45.98	14.43	0.57				
-	53	10.24	601.97	61.16	32.34	0.56				
-	54	14.46	699.89	48.04	20.18	0.49				
-	55	25.88	680.41	10.3	1.92	0.66				
-	56	11.46	695.96	55.17	23.54	0.52				
-	57	12.79	775.43	31.58	11.07	0.59				
-	58	8.35	643.5	33.39	28.98	0.56				
-	59	79.27	817.16	55.98	7.88	0.6				