Chapter 3

Parametric Modelling of the Runoff Process

3.1 Basic Principles

How complex should a model be to describe the observed reality? The answer to this question depends on the available data and the intended final use of the proposed model. If the aim of modelling is to understand the relationships among several intertwined components, then a parametric model -for instance one aimed to describe a characteristic of the runoff process- should be as simple as possible so that the main relationships among the input variables can be fully perceived. "*The complexity of reality does not imply the need for a complex model"* (Gilchrisk 1984). Based on the knowledge provided by simple models, more complex ones can be formulated afterwards to tackle the deficiencies of the simple ones. In this context the concept of simplicity comprises the following principles (Gilchrisk 1984):

1. **Parsimony of parameters.** This principle advises that the number of parameters in a given model should be minimum. In other words: "*entities should not be multiplied unnecessarily*" (William of Ockham, ~14th century).

2. **The number of variables.** The number of selected explanatory variables should be as few as possible, but they should explain as much as possible the phenomenon represented by the explained variable.

3. **The model structure.** The functional relationships linking all variables employed in a given model should be as simple as possible. Linear relationships would be preferred to non-linear ones if the studied phenomenon allows such simplification.

4. **A good approximation to reality.** A given model that has been selected based on the previous principles should provide a good approximation to the observed phenomenon described by the collected data.

3.2 Defining the Formal System

The cumulative throughput of the water cycle or any of its derivative characteristics Q_i^t for a given basin i within the Study Area during a period t (from time $t-1$ to time t) can be defined as a function of observables[1] (Chow 1962, Rodriguez-Iturbe 1969, Raudkivi 1979, Clark 1994, Abdulla and Lettenmaier 1996 have proposed similar approaches) and/or of their derivative information as follows

$$Q_i^t = f\left(\mathbf{G}_i^t, \mathbf{U}_i^t, \mathbf{M}_i^t, \boldsymbol{\beta}\right) + \varepsilon_i^t \quad \forall \quad i = 1,\ldots,n \quad \forall \quad t = 1,\ldots,T\,, \tag{3.1}$$

where

Q_i^t the output variable measured for the spatial unit i occurred during the period t,

$\mathbf{G}_i^t$ $= \left[x_{i,1}^t \quad x_{i,2}^t \quad \cdots \quad x_{i,g}^t\right]$, a vector of size $(1\times g)$ containing g observables that describe the morphological characteristics for the spatial unit i during the period t,

$\mathbf{U}_i^t$ $= \left[x_{i,g+1}^t \quad x_{i,g+2}^t \quad \cdots \quad x_{i,g+u}^t\right]$, a vector of size $(1\times u)$ containing u input variables that describe the land cover states for the spatial unit i during the period t,

$\mathbf{M}_i^t$ $= \left[x_{i,u+g+1}^t \quad x_{i,u+g+2}^t \quad \cdots \quad x_{i,u+g+m}^t\right]$, a vector of size $(1\times m)$ containing m input variables that describe the climatic conditions for spatial unit i during the period t,

$\boldsymbol{\beta}$ $= \left[\beta_l\right]$, a vector of size $(J^*\times 1)$ containing the model parameters to be estimated.

ε_i^t an independent additive error for the spatial unit i occurred during the period t,

i a subscript for spatial units; $i = 1,\ldots,n$,

j a subscript for type of input variable; $j = 1,\ldots,J$,

t a subscript for the time period; $t = 1,\ldots,T$,

n the total number of spatial units within the Study Area,

l a subscript for each model parameter; $l = 1,\ldots,J^*$,

J^* the total number of model parameters,

J $= g + u + m$, the total number of input variables or observables,

T = 33 years, the total number of years covered by the available time series, i.e. from 1961 to 1993, and

$f(\bullet)$ a non-linear function.

The formal system (see Chapter 2) as it is stated in (3.1) is a function of all available variables. However, in a highly complex natural system such as the water cycle, where everything is related with everything else, it is highly improbable to find an observable $x_{ij}^t \in \left(\mathbf{G}_i^t \cup \mathbf{U}_i^t \cup \mathbf{M}_i^t\right)$ where $1 \le j \le J$ that is absolutely independent of the rest of the input variables. Additionally, it is also

[1] A physical property, such as weight or temperature, that can be observed or measured directly, as distinguished from a quantity, such as work or entropy, that must be derived from observed quantities (Walker, 1999).

possible that some of the input variables are more suitable to describe a characteristic of the water cycle than others due to particular reasons, or that a subset of input variables are linearly dependent among themselves, hence having a lesser number of them might be enough to explain the system's behaviour. In other words, there may be multicollinearity amongst the variables contained in a given data set (Montgomery 1982).

Based on this rationale, it is sound to assume that it is likely that a set made up of few key variables may explain the behaviour of the system almost as good as the original model described by (3.1), with the great advantages of having a fewer number of input variables to deal with and thus a much simpler system to understand. The problem is therefore, to find out which set of variables explains as much as possible the observed system's output while keeping the number of variables as small as possible. In addition to that, the selected variables have to be statistically significant as will be explained later.

Assume that a set of L variables exist and fulfils the previous conditions, thus a hydrological characteristic of the system can be represented as

$$Q_i^t = f\big(\underbrace{x_{i(1)}^t,\ \ x_{i(2)}^t,\ \ldots,\ \ x_{i(j_G)}^t}_{\in\,\mathbf{G}^t},\ \ \underbrace{\ldots,\ \ x_{i(j_U)}^t}_{\in\,\mathbf{U}^t},\ \ \underbrace{\ldots,\ \ x_{i(L)}^t}_{\in\,\mathbf{M}^t},\boldsymbol{\beta}\big) + \varepsilon_i^t\,. \tag{3.2}$$

The selected input variables are ordered (here represented by a sub index within parentheses) so that they correspond to the original variables according to the following convention

$$x_{\mathrm{i}(j)}^t \in \mathbf{G}^t \quad \forall \quad 1 \le j \le j_{\mathrm{G}}$$

$$x_{\mathrm{i}(j)}^t \in \mathbf{U}^t \quad \forall \quad j_G + 1 \le j \le j_{\mathrm{U}}$$

$$x_{\mathrm{i}(j)}^t \in \mathbf{M}^t \quad \forall \quad j_{\mathrm{U}} + 1 \le j \le L$$

with

$$3 \le L < J$$

$$j_{\mathrm{G}} \ge 1,\ j_{\mathrm{U}} - j_{\mathrm{G}} \ge 1 \text{ and } L - j_{\mathrm{U}} \ge 1\,.$$

The minimum number of variables has been fixed to three because each subcategory of the input variables has to be represented by at least one variable. This constraint will allow tackling effectively the first objective of this study, namely: to assess the effects of land cover change under continuously changing weather conditions and assuming that the physiographical characteristics of the Study Area at mesoscale level can be considered as invariant during the chosen time interval of this study.

By using this procedure it will be possible to split the observed variability of the output variable along the time axis into two independent components, one that is only explained by climatic fluctuations (some of them cyclic or even exogenous to the Study Area), and the second one that is exclusively explained by land cover changes occurring within the system. It should be noted that the model will be fitted under given physiographical characteristics for various basins within the study area.

Furthermore, since a watershed is an open system, it can be assumed that land cover changes may influence the microclimatic conditions and hence the throughput of the system, but they would have very little influence on the macroclimate of the basin, which is considered an exogenous variable of the system.

The reasoning stated above can be summarised by the following expression

$$\frac{dQ^t}{dt}=\frac{\partial Q^t}{\partial G^t}\frac{dG^t}{dt}+\frac{\partial Q^t}{\partial U^t}\frac{dU^t}{dt}+\frac{\partial Q^t}{\partial M^t}\frac{dM^t}{dt}. \qquad (3.3)$$

As was stated in Section 2.2, the physiographical factors are regarded as quasi-static, thus

$$\frac{dG^t}{dt}\approx 0 \qquad (3.4)$$

Hence

$$\frac{dQ^t}{dt}\approx\frac{\partial Q^t}{\partial U^t}\frac{dU^t}{dt}+\frac{\partial Q^t}{\partial M^t}\frac{dM^t}{dt}. \qquad (3.5)$$

3.3 Modelling the Long-term Mean of the Annual Specific Discharge

3.3.1 Introduction

In order to develop and test a methodology to solve the problem stated before, a characteristic of the water cycle, namely the 33-year annual mean specific discharge for the catchments within the Study Area is to be modelled. Such an exercise is the simplest to be carried out and therefore it will allow testing the proposed method, as well as comparing its results with those obtained by standard methods often found in the literature (e.g. in Chow 1964 or in Clark 1994).

In should be noted that the model (3.8) to be derived here will not allow us to assess the effects of land cover change because the evolution of the system during the studied period is not taken into account, but rather than this, it will show whether a variable $x_{ij}\in\mathbf{U}$ (i.e. a land cover state of the basin) contributes “on average” to describe significantly the system or not. In the present case, each variable x_{ij} is defined as

$$x_{ij}=\frac{1}{T}\sum_{t=1}^{T}x_{ij}^t \quad \forall \;\; i=1,\ldots,n \quad \forall \;\; j=1,\ldots,J, \qquad (3.6)$$

and

$$Q_i=\frac{1}{T}\sum_{t=1}^{T}Q_{i1}^t \quad \forall \;\; i=1,\ldots,n. \qquad (3.7)$$

In this case, each element x_{ij} contains the arithmetic means of the available time series for the spatial unit i and the input variable j. As a consequence of this, the time index t is not longer needed; hence, the model could be called time-independent or static. So Q would be simply represented as

$$Q_i=f\left(x_{i1},x_{i2},\ldots,x_{ij},\ldots,x_{iJ},\boldsymbol{\beta}\right)+\varepsilon_i. \qquad (3.8)$$

Assume that the variables shown in Table 3.1 may be used to describe this characteristic of the water cycle in the Study Area. In this case $m = 3,\ g = 16,\ u = 3,$ thus $J = 22$.

Table 3.1 Definition and notation of input and output variables used to describe the 33-year mean annual discharge for the Study Area.

Variable			Unit	Description
Factor	**Name**	**Index j**		
Q			[mm]	33-year mean specific annual discharge
G	x_1	1	[km²]	Area of the catchment
	x_2	2	[°]	Mean catchment slope
	x_3	3	[°]	Median of the catchment's slope
	x_4	4	[°]	Trimmed mean slope $F_{(15)}$-$F_{(85)}$
	x_5	5	[°]	Trimmed mean slope $F_{(30)}$-$F_{(70)}$
	x_6	6	[°]	Mean slope of the stream network
	x_7	7	[°]	Mean slope in floodplains
	x_8	8	[1/km]	Drainage density
	x_9	9	[-]	Shape factor
	x_{10}	10	[-]	Fraction of north-facing slopes
	x_{11}	11	[-]	Fraction of south-facing slopes
	x_{12}	12	[m]	Mean elevation of the catchment
	x_{13}	13	[m]	Difference between max. and min. elevation within a catchment
	x_{14}	14	[-]	Fraction of saturated areas
	x_{15}	15	[mm]	Mean field capacity
	x_{16}	16	[-]	Fraction of karstic formations
U	x_{17}	17	[-]	Mean fraction of forest cover
	x_{18}	18	[-]	Mean fraction of impervious cover
	x_{19}	19	[-]	Mean fraction of permeable cover
M	x_{20}	20	[mm]	Mean annual precipitation
	x_{30}	21	[°C]	Mean temperature in January
	x_{32}	22	[°C]	Mean maximum temperature in January

Based on this assumption, the task will be to find out which variables are the most and the least significant. For instance, variables such as: $x_2, \ldots, x_7$ (see Table 3-1), are all depicting the slope of the catchment's terrain using different definitions or conventions. Slope is in general a very important physiographical factor since it is related to the velocity of the surface runoff and the rate of infiltration into the soil matrix, therefore, based on these arguments, it can be assumed that a variable representing this factor should be relevant to model a long-term mean of the annual discharge. The problem is then to find the best indicator representing the slope. Similar reasoning can be applied for the other variables.

In order to solve this problem three algorithms are proposed, namely:

- Modified forward selection,
- Modified backward elimination, and
- Modified all-possible regressions approach.

These approaches are based on standard statistical procedures (Montgomery 1982) but with some modifications to overcome the difficulties imposed by the system analysed in this study.

The standard method, i.e. the Stepwise Method (Montgomery 1982, Gilchrisk 1984), use multi-linear regression analysis to rank the input variables from the weakest to the strongest, or vice versa. Using these results a model with the j^{th} strongest variables can be selected. This method estimates the parameters of a given model by minimizing the so-called unexplained deviation, commonly known as the L_2 estimator (Rousseeuw and Leroy 1987). Such estimator is defined as

$$L_2 = \sum_i (\varepsilon_i)^2 \rightarrow \quad \text{min!}. \tag{3.9}$$

The shortcomings of this procedure stem from its assumptions, namely:

- The relationship between input and output variables is assumed to be linear.
- The errors ε_i have to be independent random variables and normally distributed with zero mean and constant variance (homoscedastic) for all i (Berenson 1983, Montgomery 1982, Wonnacott 1990). Standard parametric statistical tests can be used for analysis of variance, calculation of confidence intervals, and test of independence only if these conditions are fulfilled.
- There is no guarantee that "*the best*" model has been chosen.

This means that a model describing a highly complex system such as the water cycle, which is non-linear by nature (Bonell, 1993), had to be linearized if its parameters would have to be estimated using multi-linear regression. Usually, a model is linearized by taking logarithms of (3.8). According to the assumptions, $\ln(\varepsilon_{\text{i}})$ has to be normally distributed with zero mean and constant variance [i.e. $\text{N}(0, s^2)$], which in turn implies that ε_i has to be lognormal distributed (Gilchrisk 1984), which is not true in reality.

The following algorithms will consider the following improvements to overcome these shortcomings:

1. The form of the model should have a non-linear functional form $f(.)$ and its parameters have to be estimated by a non-linear optimisation algorithm without any sort of linearization or suitable transformation.
2. The estimator Φ, which constitutes the objective function to be minimized by a non-linear optimisation algorithm, should be in general written as $\min_{\boldsymbol{\beta}} \Phi$

   with

   $$\Phi = \sum_{t=1}^{T} \sum_{i=1}^{n} w_i^t \left|\varepsilon_i^t\right|^{\varphi}, \tag{3.10}$$

   where

   $$\varepsilon_i^t = Q_i^t - \hat{Q}_i^t \quad \forall \quad i = 1,\ldots,n \quad \forall \quad t = 1,\ldots,T \tag{3.11}$$

   $$\hat{Q}_i^t = f\left(x_{i1}^t, x_{i2}^t, \ldots, x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) \tag{3.12}$$

   ε_i^t a random error with zero mean for a spatial unit i occurred during the period t,

   $\hat{Q}_i^t$ an estimate of the output variable for a spatial unit i occurred during the period t,

φ a parameter[2] greater than zero. It denotes the confidence that one has in the data set and the influence that outliers may have in the estimation of $\hat{\boldsymbol{\beta}}$. The bigger φ, the more the influence of outliers is with respect to the estimates of the output variable. Rousseeuw and Leroy (1987) have extensively documented the effect of the type of estimator with regard to the robustness of the model parameters,

w_i^t a weighting factor greater than or equal to zero corresponding to a spatial unit i during the period t introduced to correct heteroscedasticity if present in the data set, or to diminish the influence of outliers in the estimation of the model's parameters; hence, it will contribute to improve the model robustness. The same idea is used by the weighted least squares method (Montgomery 1982, Rousseeuw and Leroy 1987). This weighting factor is estimated as follows:

$$w_i^t = \begin{cases} 1 \quad \text{if} \quad \left|\dfrac{\varepsilon_i^t}{s_\varepsilon}\right| \le Z_c \\ 0 \quad \text{if} \quad \left|\dfrac{\varepsilon_i^t}{s_\varepsilon}\right| > Z_c \end{cases} \tag{3.13}$$

with

$$s_\varepsilon = \sqrt{\frac{1}{n_0 - 1}\sum_t \sum_i (\varepsilon_i^t)^2} \tag{3.14}$$

s_ε the estimated sample standard deviation of random errors provided that the expectation of ε_i^t is zero, $\overline{\varepsilon} = E\left[\varepsilon_i^t\right] = 0$,

$\overline{\varepsilon}$ the mean of random errors,

n_0 the total number of observations,

Z_c a threshold value normally ranging from 2 to 3 (Rousseeuw and Leroy 1987).

3.3.2 Modified Forward Selection

Assuming that the general model is represented by (3.1), then the expected output should be as follows

$$Q_i^t = f\left(x_{i(1)}^t,\ x_{i(2)}^t,\ \ldots,\ x_{i(J)}^t, \boldsymbol{\beta}\right) + \varepsilon_i^t, \tag{3.15}$$

where $x_{i(1)}^t$ is the strongest input variable, or in other words, the variable that alone got the minimum value for the estimator Φ presented by (3.10). The next variable $x_{i(2)}^t$ is one that makes the greatest improvement to the model (further reduction of Φ) once $x_{i(1)}^t$ has been already selected. This process is then repeated $J-2$ times (Gilchrisk 1984). In (3.15) $x_{i(J)}^t$ represents the weakest variable. Weak variables can be discarded due to their small contribution in explaining the dependent variable Q_i^t.

[2] Historically astronomers in the 18th century and then Edgeworth (1887) used $\varphi = 1$, but due to great difficulties they had trying to minimise (3.10) this criterion was abandoned and replaced by $\varphi = 2$ (first introduced by Laplace) (Gilchrisk 1984), as is actually used by the Method of Least Squares.

In general, the algorithm used for this approach is as follows:

Algorithm 1

1. Assume a functional form for $f(\bullet)$.
2. For all $j = 1,\ldots,J$.
   a. Bring into the model the variable x_j^t and estimate $\hat{\boldsymbol{\beta}}$ so that $\Phi_j \to \min!$; the model at this stage has only one input variable, namely:
   $$Q_i^t = f\left(x_{ij}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t \quad \forall \quad i = 1,\ldots,n \quad \forall \quad t = 1,\ldots,T. \tag{3.16}$$
   b. Perform a significance test (see Section 3.3.7) for the variable x_j^t.
3. Repeat step 2. $(J-1)$ times.
4. Select a variable that is significant (from step 2.) and gives the lowest estimator, i.e. $\min(\Phi_j)$. This is the **strongest variable** among a set of J variables available. Rename it as $x_{(1)}^t$ and use it always in the following steps.
5. For all $j = 1,\ldots,J \;\wedge\; j \neq (1),\ldots,(.)$.
   a. Bring the new variable x_j^t into the model $j \neq (.)$, then estimate $\hat{\boldsymbol{\beta}}$ so that $\Phi_j \to \min!$; the model at this stage is as follows:
   $$Q_i^t = f\left(x_{i(1)}^t, x_{ij}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t \quad \forall \quad i = 1,\ldots,n \quad \forall \quad t = 1,\ldots,T. \tag{3.17}$$
   b. Perform a significance test (see Section 3.3.7) for the variable x_j^t.
6. From the remaining variables (step 5.) select as in step 4. the **second strongest variable**. Rename it as $x_{(2)}^t$ and then include it in (3.16) as follows:
   $$Q_i^t = f\left(x_{i(1)}^t, x_{i(2)}^t, x_{ij}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t \quad \forall \quad i = 1,\ldots,n \quad \forall \quad t = 1,\ldots,T. \tag{3.18}$$
7. Repeat steps 5.-6. until all variables are chosen or stop it either if no more significant reduction of Φ is achieved by the inclusion of a new variable j, or if the last chosen variable is not statistically significant (step 5.b.).

3.3.3 Modified Backward Elimination

This procedure is the opposite of that presented in the Algorithm 1. In other words, this approach starts with all variables and discards in each step the variable with the lowest contribution to the model. The first variable to be discarded is the weakest variable. Then the process continues until only one variable is left, this is then called the strongest variable. In general, the algorithm is as follows:

Algorithm 2

1. Assume a functional form for $f(\bullet)$.
2. Bring all variables x_j^t into the model and estimate $\hat{\boldsymbol{\beta}}$ so that $\Phi \to \min!$; a model at this stage has J variables (sometimes it is called *saturated model*, Gilchrisk 1984) namely:
   $$Q_i^t = f\left(x_{i1}^t,\; x_{i2}^t,\; \ldots,\; x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t. \tag{3.19}$$
3. For $j = 1,\ldots,J$ that are still in the list of variables.
   a. Estimate $\hat{\boldsymbol{\beta}}_j$ eliminating only variable j at each step so that $\Phi_j \to \min!$; in general a model at this stage is as follows
   $$Q_i^t = f(x_{i1}^t,\; x_{i2}^t,\; \ldots,\; x_{i,j-1}^t, x_{i,j+1}^t,\; \ldots, x_{iJ}^t, \hat{\boldsymbol{\beta}}_j) + \varepsilon_i^t. \tag{3.20}$$

b. Select from those models obtained in step 3.a. one combination that provides the lowest estimator, i.e. $\min(\Phi_j)$. Then exclude the variable that has not been used, i.e. x_j^t. Comparing the estimator of this combination Φ_j with Φ, it is clear that the contribution of variable j has been minimal compared with the rest. For the next steps this variable, called the **weakest variable**, would be excluded.

4. Repeat step 3 $(J-1)$ times eliminating the weakest variable each time until one variable is left. The last one is called the **strongest variable**.

3.3.4 Building All Combinations

Although the two described procedures are relatively fast, they only consider a small subset from all possible combinations that can be built up from J input variables. Hence, many models, perhaps very good ones, are not evaluated by these procedures. This shortcoming for a complex system may be crucial because it may lead to choose a wrong model, or one that is not the best. In order to find the "*best model*", $2^J - 1$ combination of variables have to be evaluated (the null model, i.e. one having a constant and no variables has been excluded). As shown in Table 3.2, the total number of possible combinations considering 22 input variables is 4,194,303! Hence, this method, although convenient when the number of variables is small, is not practical due to the high computation time required when the number of variables is greater than 12, but still possible depending on computational power at hand.

Table 3.2 Total number of possible combinations of J input variables.

Number of variables	**Number of combinations**
J	$2^J - 1$
2	3
4	15
8	255
16	65,535
22	4,194,303
32	4,294,967,295

Assuming that the number of variables is small enough to use this method, then, how can the best model be selected out of hundreds or maybe thousands of possible models? In order to answer this question, firstly, it should be noted that the greater the number of input variables, the smaller will be the value of the objective function Φ (3.10) after the minimization. Hence, the value $\min\Phi(\hat{\beta})$ as an indicator of the quality of the model does not lead to find the best combination of explanatory variables (the same behaviour can be observed in multi-linear regression models: the greater J, the better the fit and the greater the value of R^2 is; to counter-balance this effect an adjusted $\overline{R}^2$ was proposed by Ezekiel in 1930). In the present case, two criteria have been implemented to solve this issue, namely:

- The Mallows' C_{p*} statistic to select a subset of best performing combinations of input variables, and
- A cross-validation test to evaluate the quality and robustness of the previously selected subset of combinations, from which the best model is to be chosen.

3.3.5 Selection of the Best Models Using Mallows' C_{p*} Statistic

The Mallows' statistic can be estimated as follows (Berenson 1983):

$$C_{p*} = \frac{\left(1 - R_{p*}^2\right)\left(n_0 - J\right)}{1 - R_J^2} + 2p * -n_0 , \tag{3.21}$$

where

$$R_{p*}^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\left(Q_i^t - \hat{Q}_i^t\right)^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}\left(Q_i^t - \frac{1}{n_0}\sum_{t=1}^{T}\sum_{i=1}^{n}Q_i^t\right)^2} \tag{3.22}$$

$p*$ the number of parameters used in a given model that contains j input variables,

R_J^2 is equal to R_{p*}^2 if $j = J$ and $p* = J^*$. In other words, the coefficient of determination associated with a model containing all input variables available (i.e. J).

This indicator showing the quality of the model, commonly known as the C_{p*} criterion, was introduced by Mallows (1973). It has the advantage, compared with an adjusted $\overline{R}^2$, that in addition to adjust the sum of squared errors, it can be demonstrated that its expectation is equal to the number of parameters used in the model (Daniel and Wood, 1980), or

$$E\left[C_{p*}\right] = p * . \tag{3.23}$$

That means that the closer the value of C_{p*} to $p *$, the lesser the bias of the fitted model, hence, the better the model fit is. Using this property, the best model or a set of best performing models can be identified as it is shown in Figure 3.1.

Other criteria such as the Akaike's Information Criteria (Akaike, 1973) can also be used for selecting models as will be discussed later.

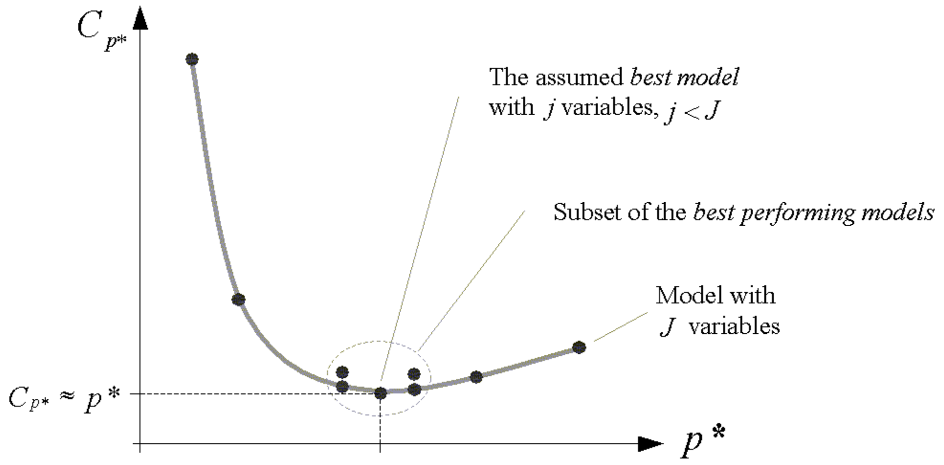

Figure 3.1 Identification of the best model using the C_{p^*} plot.

3.3.6 Model Validation

In order to evaluate the quality of the model, a *Cross-Validation Method* (Efron 1981, Simonoff 1996) is carried out for each possible model that belongs to the subset of the best performing models selected before. This procedure is a special case of the *Jackknife Method* introduced by Quenouille (1949) and Tukey (1958). It consists of dividing the data set into y groups of equal size of observations, and consecutively, it deletes one group at a time; then, it estimates the model parameters $\hat{\boldsymbol{\beta}}$ with the remaining points using the same estimation procedure previously used. A model estimated in such a way is then validated with the group of data not considered during its estimation. This procedure is then repeated for all groups, i.e y times. As a result of this procedure y *Jackknife statistics* θ_y are obtained. Finally, all y statistics are combined to obtain the *Jackknife estimator* θ. In general this estimator would indicate how robust[3] the model is; the lesser the value of θ, the more robust the model is regarding the disturbances from outliers present in the dataset.

If the number of groups is equal to the number of observations ($y = n_0$) the procedure is called *cross-validation.*

Let $\mathcal{D}$ be the original set of observations in a given case. Using the notation used before

$$\mathcal{D} = \left\{ \left(Q_i^t, x_{ij}^t \right) \,\middle|\, i = 1,\ldots,n \quad j = 1,\ldots,J \quad t = 1,\ldots,T \right\}. \tag{3.24}$$

The algorithm used to validate a model $f(\bullet)$ composed of J variables is described below (based on Efron 1981).

Algorithm 3

1. For all $i = 1,\ldots,n$.
2. For all $t = 1,\ldots,T$.

[3] The term “robust” was coined in statistics by G.E.P. Box in 1953. In general, referring to a statistical estimator, it means “insensitive to small departures from the idealized assumptions for which the estimator is optimised.” Launer and Wilkinson 1979, Huber 1981.

a. Let $\mathcal{E}_i^t = \left\{ \left(Q_i^t, \; x_{ij}^t \right) \;\middle|\; j = 1, \ldots, J \right\}$ be a subset of observations given i and t. Eliminate the subset $\mathcal{E}_i^t$ from the original data set so that a new subset $\widehat{\mathcal{D}} = \mathcal{D} - \mathcal{E}_i^t$.

b. Using $\widehat{\mathcal{D}}$ estimate $\widehat{\boldsymbol{\beta}}$ so that $\widehat{\Phi} \rightarrow \min!$.

c. Estimate $\widehat{Q}_i^t = f\left(x_{i1}^t, \; x_{i2}^t, \; \ldots, \; x_{iJ}^t, \widehat{\boldsymbol{\beta}} \right)$.

d. Calculate the Jackknife statistic for the observation i, t as follows

$$\theta_i^t = \left(Q_i^t - \widehat{Q}_i^t \right)^2 . \tag{3.25}$$

e. Repeat step 2. T times.

3. Repeat step 1. n times.
4. Calculate the overall quality indicator or Jackknife estimator for a given model as follows

$$\theta = \sum_{i=1}^{n} \sum_{t=1}^{T} \theta_i^t \;, \; \theta \geq 0 . \tag{3.26}$$

The most reliable model among the subset of the best performing models (see Figure 3.1) can be selected using the *Jackknife estimator* θ. The minimum value of θ will correspond to the best model. The exponent employed in (3.25) has been chosen equal to two because of the following reasons: 1) to make positive the difference between the calculated and the observed value; and, 2) to penalize those points where the model has large differences, hence making θ larger, and thus reducing its robustness.

3.3.7 Significance Test

A significance test has the purpose of assessing the plausibility of a scientific hypothesis (Davison and Hinkley 1997) based on a given set of data. Literally, a hypothesis should be understood as "*a proposition made as a basis for reasoning*" without reference to its value of truth, or "*as a starting-point for further investigation*" (Concise Oxford Dictionary). A significance test, however, can not prove that a hypothesis is true or false, in fact no procedure can guarantee that (Gilchrist 1984), but it will lead to conclude that based on the data available there is enough evidence to state that a hypothesis is unlikely to be true and hence can be rejected. Rejecting a hypothesis always presupposes a level of risk that can be defined as the probability that such a hypothesis is rejected when in fact it is true (Error Type I). This probability is called *level of significance* (α). By definition, a significance test is performed to infer that a hypothesis that is represented by an assumed value of a parameter called *null hypothesis* H_0 is not likely to be the true value (Lane 2001), consequently, it can be rejected in favour of an *alternative hypothesis* H_A at a given level of significance. H_A should be an important alternative of H_0 to be detected, one that is likely to be true if H_0 is not (Davison and Hinkley 1997). Often H_A is taken as the opposite of H_0.

Working with all data available to do this task is unpractical. Therefore, a *test statistic* Θ should be built so that it will satisfy the following conditions: 1) it has to summarize some aspects of the data relevant to the particular problem so that it measures the discrepancy between the data and the null hypothesis, e.g. the smaller the value of Θ, the stronger the evidence against H_0 (the opposite is also possible) is; 2) its behaviour whether H_0 or H_A is true should be remarkably different from each other; and 3) the sampling distribution of Θ must be known or at least approximately estimated under the assumption that H_0 is true (Neave and Worthington 1988).

Suppose then that a test statistic fulfils these three conditions mentioned above and that the value of the test statistic based on the available data is denoted by ϑ. In such a case, the level of evidence against H_0 is measured by the *significance probability* (Davison and Hinkley 1997) or the so-called p-value

$$p\text{-value} = \Pr(\Theta \leq \vartheta \;\Big|\; H_0). \tag{3.27}$$

If $p\text{-value} < \alpha$ two answers are plausible, namely: 1) that H_0 is true but a rare event has been observed (summarized by ϑ); or 2) that based on the strong evidence against H_0 provided by the available data, H_0 does not conform to the observed phenomenon and therefore can be considered a bad hypothesis. Hence, it can be rejected at the level of significance α. The latter answer has been adopted as the rationale of the significance test (Gilchrist 1984). Conversely, if $p\text{-value} \geq \alpha$ H_0 can not be rejected. In general, the following verbal interpretations can be formulated: if the p-value is between 1% and 5%, less than 1%, or even less than 0.1%, this would mean that there is a considerable, a very strong, or a practically conclusive evidence, respectively, in the data to reject H_0 (Neave and Worthington 1988).

In order to perform a significance test within the context of this study the following definitions are necessary. Let the set of observations be a random sample denoted by $\mathcal{D}$, whose cardinality (i.e. the number of valid observations) is

$$n_0 = |\mathcal{D}| \leq nT. \tag{3.28}$$

Based on $\mathcal{D}$, assume that an observed phenomenon in a given location i during the period t can be predicted by a model using J explanatory variables (i.e. observables and/or derivative information) and a vector of calibrated parameters $\hat{\boldsymbol{\beta}}$. Such a model is represented by

$$Q_i^t = f\left(x_{i1}^t,\ x_{i2}^t,\ \ldots,\ x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t. \tag{3.29}$$

In this case there would be J null hypotheses H_0 that require to be tested within the scope of the present study, which also implies J corresponding alternative hypotheses to be formulated. The objective of the j-th null hypothesis is to test whether the variable x_j in the model (3.29) is independent with respect to the explained variable Q considering the J-dimensional space ($\mathbb{R}^J$) where the model has been defined. In other words, to infer that based on the sample data these variables are certainly not independent at the level of significance α, or that the sample does not indicate at the level of significance α that the variable x_j has been chosen by chance when such a model was assessed.

The j-th null hypothesis and its corresponding alternative one can be written up as follows

$H_0^{(j)}$: Variables Q and x_j are independent in $\mathbb{R}^J$, given a functional $Q_i^t = f\left(x_{i1}^t,\ x_{i2}^t,\ \ldots,\ x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t$, and the random sample $\mathcal{D}$.

$H_A^{(j)}$: These variables are not independent under the previous conditions. (3.30) or the same but using conditional probabilities as

$$\begin{aligned}
H_0^{(j)} \quad : \quad & \Pr\left(Q_i^t = f\left(x_{i1}^t,\ldots,x_{i(j-1)}^t,x_{i(j)}^t,x_{i(j+1)}^t,\ldots,x_{iJ}^t,\hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t \left| x_{i(j)}^t \right.\right) \\
& = \Pr\left(Q_i^t = f\left(x_{i1}^t,\ldots,x_{i(j-1)}^t,x_{i(j+1)}^t,\ldots,x_{iJ}^t,\hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t\right) \\
H_1^{(j)} \quad : \quad & \Pr\left(Q_i^t = f\left(x_{i1}^t,\ldots,x_{i(j-1)}^t,x_{i(j)}^t,x_{i(j+1)}^t,\ldots,x_{iJ}^t,\hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t \left| x_{i(j)}^t \right.\right) \\
& \neq \Pr\left(Q_i^t = f\left(x_{i1}^t,\ldots,x_{i(j-1)}^t,x_{i(j+1)}^t,\ldots,x_{iJ}^t,\hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t\right) \quad .
\end{aligned} \tag{3.31}$$

As mentioned above, one of the prerequisites to perform a significance test is to know in advance the sampling distribution of Θ in order to calculate the exact $p\text{-value}$. In the present case, due to the complexity of the relationships among the components of the system this may be very difficult or even impossible considering that the test statistic has an unknown J - dimensional distribution under the null hypothesis.

To overcome this problem without the simplistic and sometimes doubtful assumption that the sampling distribution of Θ under H_0 is approximately equal to a known theoretical distribution (e.g. normal, exponential, t-student, χ^2 among others) a *resampling method* [4] can be used to estimate a reasonable approximation for the exact $p\text{-value}$ of the test statistic Θ. These methods, sometimes termed as *Monte-Carlo test*, *randomisation test*, *permutation test*, or *bootstrap*, are suitable to estimate confidence intervals and significance probabilities for problems with very limited datasets and unknown or -at most- partially known distribution function (Dudewicz 1992, Canty 1998). The *permutation test* is a nonparametric or *distribution-free* test, and will be employed here because of the following reasons: first, it allows using any test statistic that may be considered meaningful, and second, it can be used even if the size of the population is finite (Good, 2000).

As already mentioned, in order to test the hypothesis given by (3.31), a test statistic Θ that measures the level of dependence between the variables is needed. Furthermore, it should consider that x_j and Q are not alone, but there are $J-1$ additional explanatory variables. Thus, the simplest test statistic in such a case would be the estimator Φ defined in (3.10). The test statistic $\Theta = \Phi$ is a large number under $H_0^{(j)}$, and conversely very small if $H_0^{(j)}$ should not be true.

The rationale of this test is as follows: since F -the distribution function of Θ under the null hypothesis- is unknown, $\hat{\mathrm{F}}$ -an EDF[5] obtained from the simulated datasets under the null hypothesis-

[4] Though the resampling methods were an old idea, they were not extensively used until the late 1970's mainly due to lack of computer power not commonly available in those days. Despite the fact that fast computers did not exist until the 1960's, the first real use of such a method was carried out by W. S. Gosset ("Student") in 1908 to corroborate its famous t-distribution. Later on, in 1935, R. A. Fisher applied for first time a randomisation test to estimate p-values and some years later, Fermi, von Neumann, N. Metropolis and S. Ulman introduced the term Monte Carlo Simulation around 1948 (Hammersley and Handscomb 1964; Dudewicz 1992). For the reasons mentioned above, Monte Carlo Simulations and related techniques have been vastly used for spatial analysis (Davison and Hinkley 1997), especially since its reintroduction by Efron in 1979.

[5] Empirical Distribution Function.

is said to be *minimal sufficient* for F (Davison and Hinkley 1997). In order to estimate $\hat{\mathrm{F}}$, R batches of artificial data sets, each of size n_0, have to be generated *without replacement* from $\mathcal{D}$ (Wilks 1995).

Let the $r\text{-th}$ simulated data set be denoted by $\mathcal{D}_r^*$, with $n_0 = |\mathcal{D}| = |\mathcal{D}_r^*| \leq nT \quad r = 1,\ldots,R$. As x_j is supposed to be independent from Q under the null hypothesis, a random permutation of x_j, denoted by x_j^* , should not produce any effect in the selected test statistic, had x_j been replaced by x_j^* in the original set $\mathcal{D}$. In the present case, the result of such substitution is called the $r\text{-th}$ simulated data set $\mathcal{D}_r^*$. Further on, the test statistic will be evaluated using $\mathcal{D}_r^*$ and the result will be denoted by ϑ_r^*. Since $\mathcal{D}$ is a random sample, there are $n_0!$ equally likely permutations of x_j. As $n_0!$ is a large number, for practical reasons R will be limited to $R = 1000$ or perhaps $R = 10\,000$ randomly selected permutations. Based on these results, the EDF that mimics the unknown distribution function (F) can be calculated, and from it, the proportion of the random ϑ_r^* that are smaller than or equal to the observed ϑ is finally estimated. Such proportion is called the Monte Carlo $p\text{-value}$. Formally it can be calculated by

$$p\text{-value} = \frac{\#(\vartheta^* \leq \vartheta)}{n_0!} \cong p_{\mathrm{mc}} = \frac{\#(\vartheta^* \leq \vartheta)}{R+1}. \tag{3.32}$$

Where p_{mc} is the Monte Carlo $p\text{-value}$, and # denotes the number of permutations in which the event $\vartheta^* \leq \vartheta$ occurs.

In general, the algorithm for the significance test is as follows:

Algorithm 4

1. Given a functional form $Q_i^t = f\left(x_{i1}^t,\ x_{i2}^t,\ \ldots,\ x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t$, and the random sample $\mathcal{D}$, estimate $\hat{\boldsymbol{\beta}}$ so that $\Phi \rightarrow \min!$ The test statistic is then $\vartheta = \Phi$.
2. For all $r = 1,\ldots,R$.
   a. Generate x_{ij}^{t*} as a random permutation of x_{ij}^t, with $i = 1,\ldots,n \quad t = 1,\ldots,T$.
   b. Generate the simulated data set $\mathcal{D}_r^*$ replacing x_{ij}^t by x_{ij}^{t*}.
   c. Based on $\mathcal{D}_r^*$ estimate $\hat{\boldsymbol{\beta}}_r^*$ so that $\Phi_r^* \rightarrow \min!$ The test statistic is then $\vartheta_r^* = \Phi_r^*$.
3. Sort ϑ among $\vartheta_r^* \quad r = 1,\ldots,R$ so that
$$\vartheta_{(1)}^* \leq \cdots \leq \vartheta_{(r-1)}^* \leq \vartheta \leq \vartheta_{(r)}^* \leq \cdots \leq \vartheta_{(R)}^*. \tag{3.33}$$
4. Estimate the Monte Carlo $p\text{-value}$ as in (3.32). In this case the one sided test statistic is equal to
$$p\text{-value} \cong p_{\mathrm{mc}} = \frac{r-1}{R+1}. \tag{3.34}$$
5. Select a level of significance (say, $\alpha = 5\%$).
6. Make a decision:

   If $p\text{-value} \leq \alpha$ then,

   ⇒ Reject $H_0^{(j)}$ in favour of $H_A^{(j)}$ at the level of significance α, then

   ⇒ **Conclusion:** At this level of significance variables Q_i^t and x_{ij}^t are ***certainly not independent***.

   Else, $H_0^{(j)}$ can not be rejected.

3.3.8 Analysis of Results

The empiric probability density functions (PDF) of the explanatory variables used to model the long-term mean of the annual specific discharge are far from being normally distributed or closer to any other theoretical distribution as can be seen in Figures 3.2 and 3.3. The same is true for the explained variable.

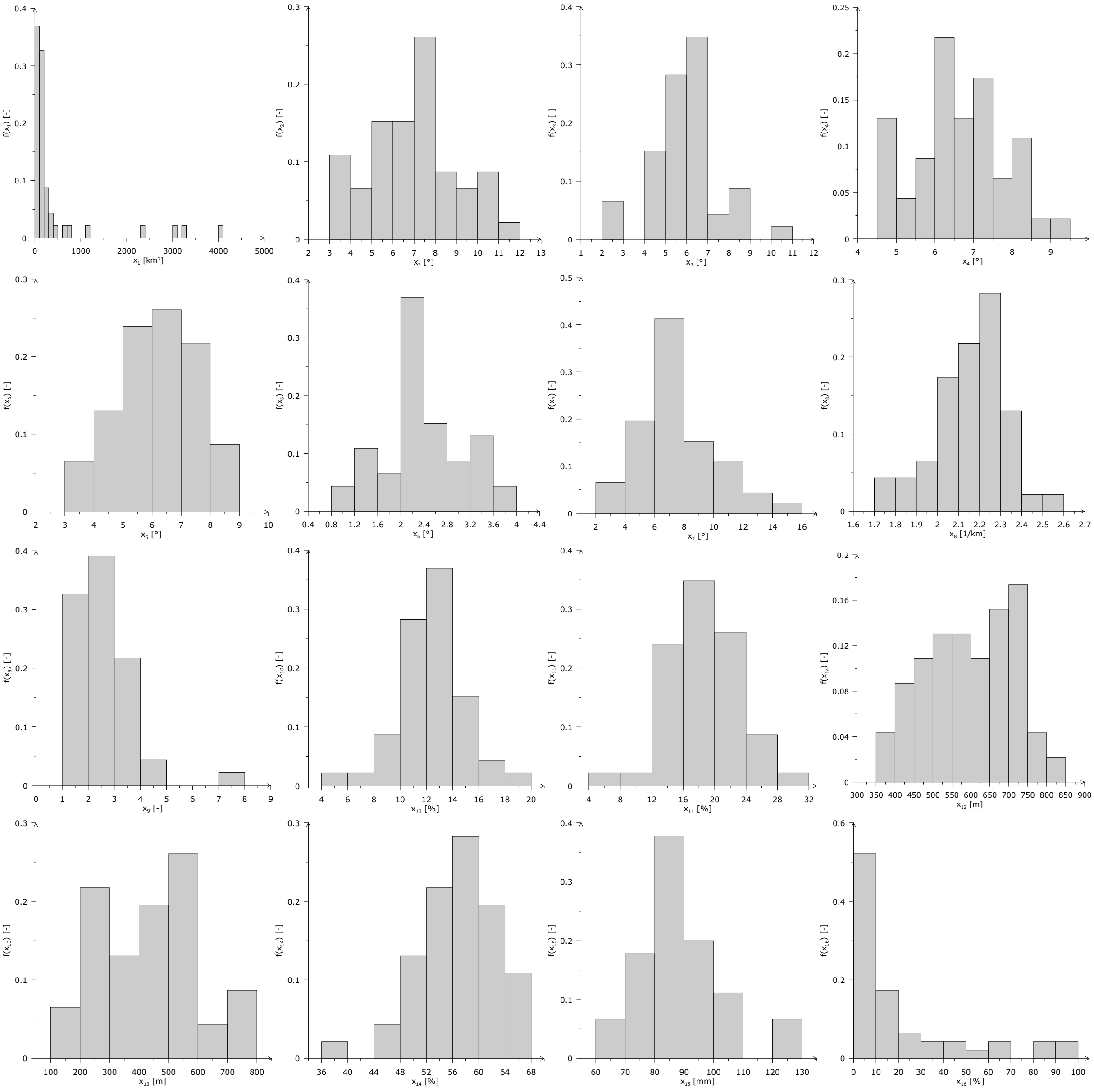

Figure 3.2 Histograms depicting the empiric PDF of all physiographic explanatory variables considered in this study.

Using the modified Backward Elimination (BE) and the modified Forward Selection (FS) (Section 3.3.3 and Section 3.3.2) the relative importance of the variables can be assessed. The results using a nonlinear model such as $Q_i = \beta_0 \prod_{i=1}^{J} x_{ij}^{\beta_j} + \varepsilon_i$ with $\varphi = 2 \ \wedge \ w_i = 1 \ \ \forall \ i = 1,\ldots,n$ are shown in Table 3.3.

Table 3.3 Relative importance of variables used to model the long-term mean specific discharge according to BE and FS approaches.

	strongest																					weakest	
BE→	x_{20}	x_4	x_3	x_{15}	x_{13}	x_7	x_1	x_5	x_{11}	x_{17}	x_{19}	x_{10}	x_8	x_6	x_9	x_{18}	x_2	x_{30}	x_{32}	x_{14}	x_{12}	x_{16}	
	x_{20}	x_4	x_3	x_{15}	x_{13}	x_{17}	x_1	x_5	x_{11}	x_{19}	x_2	x_{10}	x_{18}	x_{30}	x_7	x_8	x_9	x_6	x_{32}	x_{14}	x_{12}	x_{16}	←FS

Table 3.3 shows a direct consequence of the non-linearity of the water cycle, i.e. the different rankings obtained by using Algorithms 1 (FS) and 2 (BE) independently. The former begins with the strongest variable until the weakest variable is found, whereas the latter does the opposite. Results have shown that these procedures differ always in a number of cases (not shaded in the Table). In this case however, they have agreed on the five strongest and the four weakest variables.

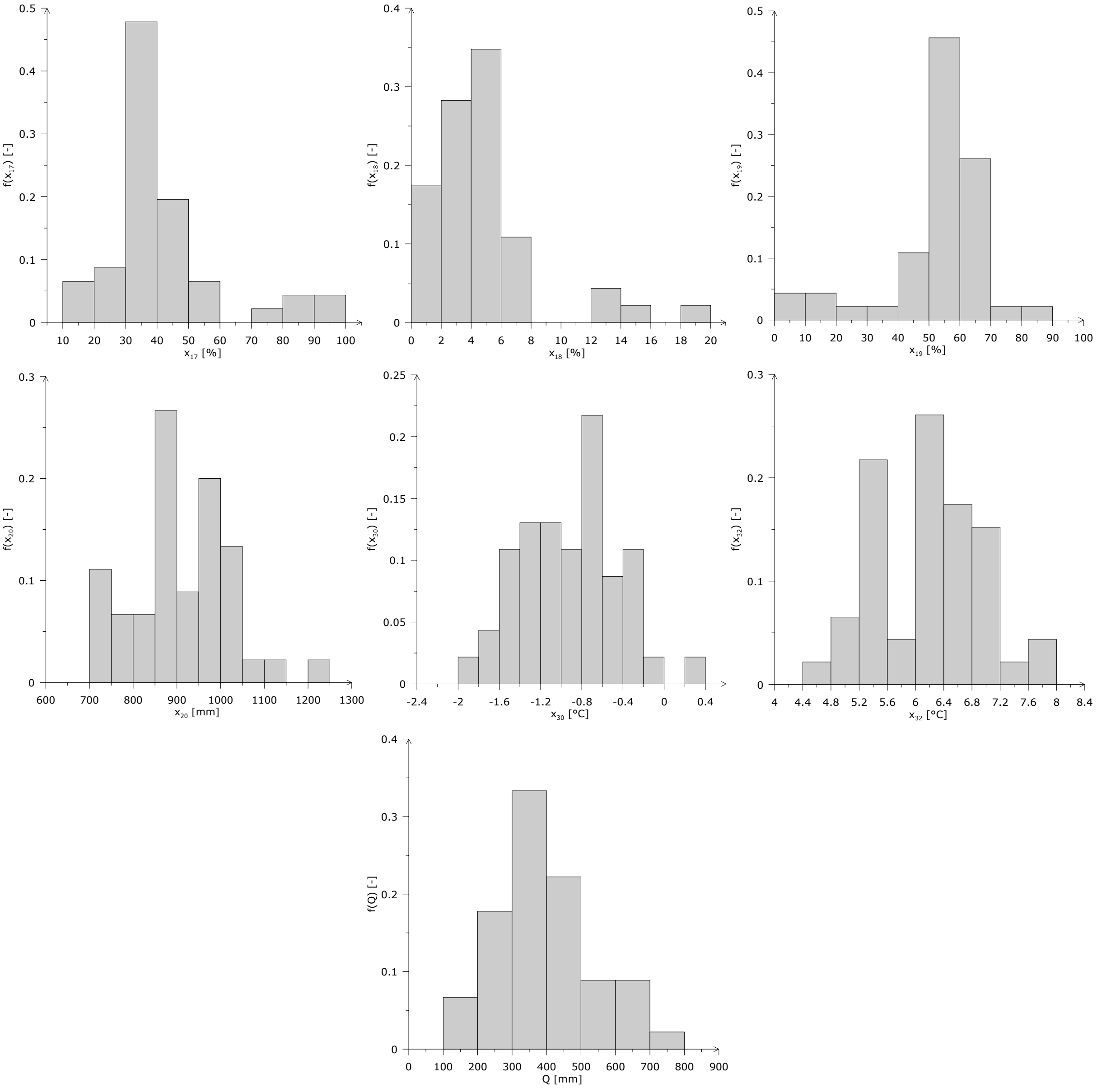

Figure 3.3 Histograms depicting the empiric PDF of the land cover and meteorological variables, as well as the specific annual discharge (the explained variable) considered as long term averages from 1961 to 1993.

It should be noted that both methods have ***only*** calculated $J\left(J+1\right)/2$ combinations of input variables, which in this case is equal to 253 out of the 4,194,303 possibilities. This represents a big disadvantage for both approaches because many 'good' models could have not been evaluated.

As can be inferred from the previous example, selecting the best model can be stated as a combinatorial problem with the following objective function: given a random sample, find the minimum number of significant variables that explain as much of its variance as possible. To solve such problem stochastic optimization methods such as simulated annealing or neural networks can be used.

Since the number of possible models is very high in the present case and hence very costly in calculation time (e.g. a computer employing one second per model would need about 48.5 days to evaluate all combinations), the previous methods may help to discard some variables that represent the same factor but have been calculated in a different way, as it is the case with the variables $x_2,\ldots,x_7$. In other words, these methods may help to assess the relative importance of the variables among each different sub-group of factors. So, using this procedure only the variables $\left\{x_1,x_4,x_8,x_9,x_{11},x_{13},x_{15},x_{17},x_{18},x_{19},x_{20},x_{30}\right\}$ have been selected for the next step, i.e. '*to find the best model*'.

In this case, building all possible models still is a feasible approach because only 12 variables have been left after the first screening. The results obtained can be appreciated by means of a C_{p^*} plot shown in Figure 3.4. Additionally, the composition of some of the best performing models has been presented in Table 3.4.

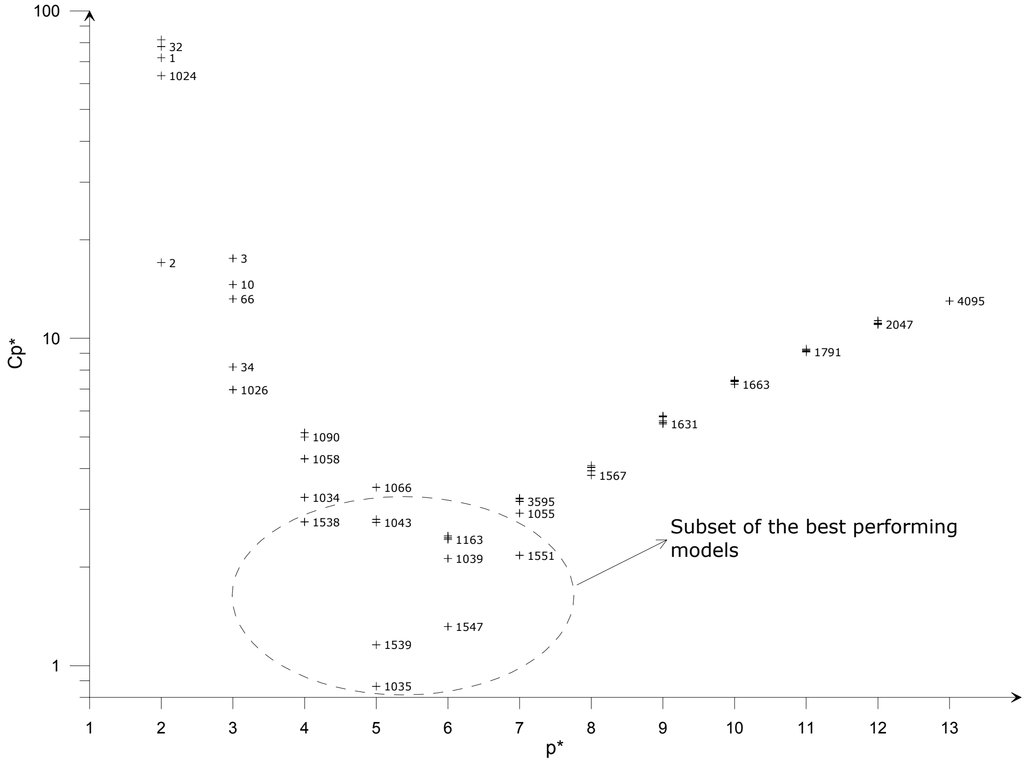

Figure 3.4 C_{p^*} vs. p^* plot showing the best 5 models for each p^* using the following variables $\left\{x_1,x_4,x_8,x_9,x_{11},x_{13},x_{15},x_{17},x_{18},x_{19},x_{20},x_{30}\right\}$. The number at the right of the marker (+) indicates the model's number. For models where $p^* \geq 8$ only the number of the best model is shown.

The last row in Table 3.4 shows the relative frequency of occurrence of a variable only with regard to the subset of best performing models shown in Table 3.4 by (⌘).

These frequencies show that the most common variables among those of the subset are mean precipitation and trimmed slopes 15-85 (x_{20}, x_4), followed by mean temperature in January (x_{30}); then by mean fraction of impervious cover and drainage density (x_{18}, x_8); then another land cover related variable, namely the fraction of permeable cover (x_{19}), and then all the rest. These results are not surprising because the system is mainly driven by precipitation, topography, and macroclimate; thus they appear as the most commonly used variables. What is more interesting is the fact that one of the variables representing land cover is very often used as an explanatory variable describing the mean discharge of mesoscale basins.

Table 3.4 Design matrix showing the composition of some of the best models depicted in Figure 3.4 (1 ≡ a variable is included in the model, 0 ≡ otherwise). For each model the value of the estimator Φ and the Jackknife statistic θ is also presented (⌘ ≡ Subset of the best models).

Model Number	x_1	x_4	x_8	x_9	x_{11}	x_{13}	x_{15}	x_{17}	x_{18}	x_{19}	x_{20}	x_{30}	Φ	θ	Description
2											1		0.5427	0.5948	
1026		1									1		0.4299	0.4889	
1538		1	1								1		0.3717	0.4697	⌘
1035		1							1		1	1	0.3354	0.4300	⌘
1539		1	1								1	1	0.3381	0.4729	⌘
1043		1						1			1	1	0.3529	0.4755	⌘
1570		1	1				1				1		0.3535	0.4764	⌘
1547		1	1						1		1	1	0.3209	0.4896	⌘
1039		1							1	1	1	1	0.3285	0.4270	⌘
1163		1			1				1		1	1	0.3313	0.4741	⌘
3587	1	1	1								1	1	0.3316	0.4849	⌘
1099		1				1			1		1	1	0.3319	0.4680	⌘
1551		1	1						1	1	1	1	0.3102	0.4885	⌘
1567		1	1					1	1	1	1	1	0.3068	0.5315	
1631		1	1			1		1	1	1	1	1	0.3037	0.5835	
1663		1	1			1	1	1	1	1	1	1	0.3015	0.7538	
1791		1	1		1	1	1	1	1	1	1	1	0.3000	0.8293	
2047		1	1	1	1	1	1	1	1	1	1	1	0.2994	0.8766	
4095	1	1	1	1	1	1	1	1	1	1	1	1	0.2992	0.9248	Saturated model
Frequency [%]	2	22	12	0	2	2	2	2	12	4	22	18	Only considering models showing a ⌘.		

Another important conclusion that can be drawn from Table 3.4 is that using all available variables for a given phenomenon does not always lead to the best model. This fact is related with the problem's dimensionality[6]. In the present case, the dimensionality of the system is around 7. This indication

[6] The dimensionality of a system is the minimum number of linear combinations of principal components that explain as much as, say 95%, of the total variance observed in its correlation matrix.

suggests the adequate number of variables that a model should have. In the present case, the subset of best performing models has between 3 and 6 variables.

The next step is to select from the short list of "good" candidates the best one. In other words, which is the most reliable model within this subset that satisfy the constrains given by (3.2) and has variables with a level of significance, say $\alpha = 5\% = 0.05$?

As can be seen in Table 3.4, the estimator Φ alone does not lead to the best model, which from this point of view only, is the saturated model (No. 4095) since it exhibits the minimum value for the estimator (0.2992). The answer to the first part of the question can be given by calculating the Jackknife or cross-validation statistic θ also depicted in Table 3.4. Using this indicator, the robustness of a model can be assessed. Not surprisingly, the saturated model gets the highest value (0.9248), this means it is to be considered the least reliable model. Therefore, a trade-off between Φ and θ should be taken into account in order to make a wise selection decision, which leads to pick models No. 1039 and No. 1035 as those with the lowest and second lowest Jackknife statistic θ (0.4270 and 0.4300 respectively).

The final step is then to determine whether all variables are significant or not at a certain level of significance chosen beforehand. The described simulation technique explained before delivers the estimates for the p-value shown in Table 3.5.

Table 3.5 Results of the permutation test for models No. 1035 and No. 1039 using R=500. The tabulated figures are the Monte Carlo p-values as fractions.

Model Number	x_4	x_{18}	x_{19}	x_{20}	x_{30}
1035	0.0020	0.0160	-	0.0000	0.0360
1039	0.0080	0.0260	0.3740	0.0000	0.0440

These results lead to the final decision, namely: model number No. 1035 is selected as "*the best*" one, since all its variables have successfully passed the significance test. Hence, all $H_0^{(j)}$ can be rejected in favour of the corresponding $H_A^{(j)}$, for all $j = 4,18,20,30$ at 5% level of significance. As a conclusion it is possible to state that all its variables are certainly not independent of the explained variable at the given level of significance.

The most significant variable in model No. 1035 is precipitation ($p_{mc} \simeq 0.0\% < 5\%$) and the least significant mean temperature in January ($p_{mc} \simeq 3.6\% < 5\%$). Model No. 1039, although with the best cross-validation statistic, has one variable (x_{19}) failing to pass the significance test and thus it is dropped out. This variable corresponds to the fraction of permeable areas whose $p_{mc} \simeq 37.4\% > 5\%$.

Finally, the model that best describes the mean specific discharge occurring within a catchment located in the Upper Neckar Basin, based on the provided information can be written explicitly as

$$Q_i = 0.559 \times 10^{-2} \left(x_{i4}^{0.6709} \; x_{i18}^{0.1089} \; x_{i20}^{1.8860} \; x_{i30}^{-1.4226} \right) + \varepsilon_i \tag{3.35}$$

This model has been calculated considering all observations contained in the sample (n=46). The relationship between observed values and calculated ones are depicted in Figure 3.5. This picture shows also two likely outliers encircled by a doted line. These points, which may contain big errors, e.g. due to faulty measurements, can influence drastically the model performance. They should be carefully checked, and if the errors persist then they should be removed from the data set. The identification of outliers and the utilization of more robust estimators will be explored in the next chapter.

The proposed model shows clearly that land cover is a significant variable with regard to the estimation of the long term mean specific discharge, but, since it is a static model, it can not be used to assess the hydrological impacts triggered by land cover changes [see (3.3) to (3.5)]. It is presented here because it helps to show the advantages of the proposed method using a practical but computationally simple example rather than to provide an answer to the research question stated in Chapter 1.

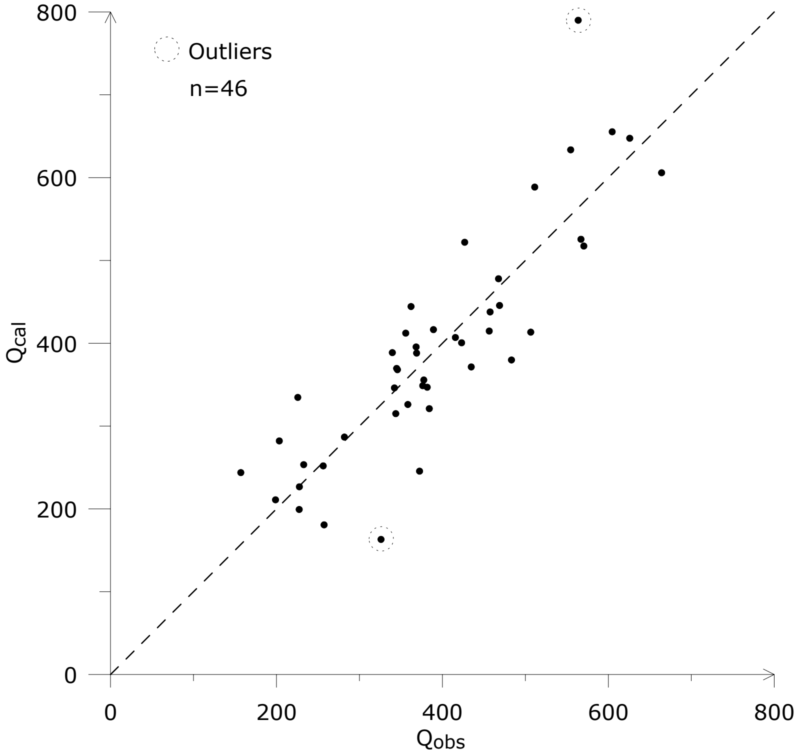

Figure 3.5 Scatterplot showing the relationship of Q_{obs} vs. Q_{cal} using the model (No. 1035) given by (3.35). A sample of size $n = 46$ was used in the calculation. Outliers have not been removed.

In order to provide an answer to the research question, time dependent models should be calibrated using the proposed method. Chapter 4 will be devoted to this task. Models aimed at estimating the specific discharge, the specific volume of high flows, the specific peak discharge, among others, at annual or seasonal basis will be presented afterwards.