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

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

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

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

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

$$\frac{dG^t}{dt} \approx 0 \tag{3.4}$$

Hence

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

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

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

and

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

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

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

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

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

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, $\ln(\varepsilon_i)$ has to be normally distributed with zero mean and constant variance [i.e. $\mathrm{N}(0, s^2)$], 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 $\min_{\hat{\boldsymbol{\beta}}} \Phi$

   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:

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

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 $\Phi_j \rightarrow \min!$; 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 $\Phi_j \rightarrow \min!$; 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 $\Phi \rightarrow \min!$; 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 $\Phi_j \rightarrow \min!$; 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 J	**Number of combinations** $2^J - 1$
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):

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

where

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

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

R^2_J is equal to R^2_{p*} 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.

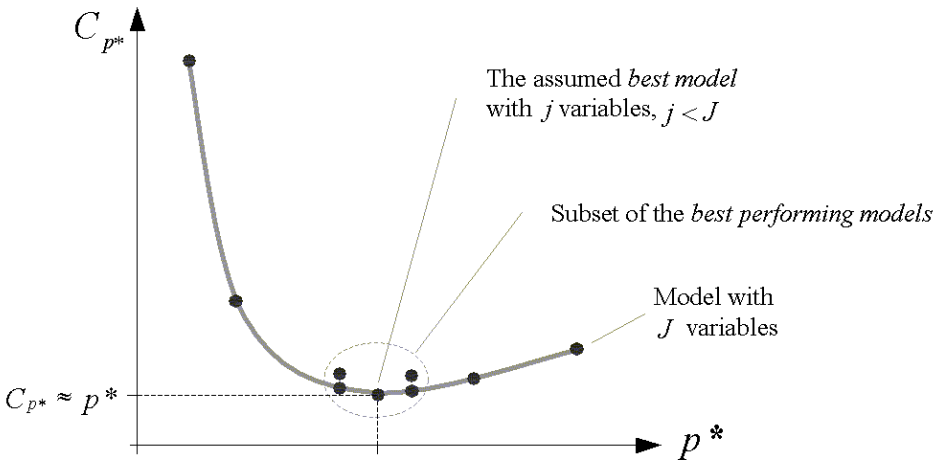


Figure 3.1 Identification of the best model using the C_{p*} plot.

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-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-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

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

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

$$\begin{aligned} H_0^{(j)} &: \text{ Variables } Q \text{ and } x_j \text{ are independent in } \mathbb{R}^J, \text{ given a functional} \\ &\quad Q_i^t = f\left(x_{i1}^t,\;\; x_{i2}^t,\;\; \ldots,\;\; x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) + \varepsilon_i^t, \text{ and the random sample } \mathcal{D}. \\ H_A^{(j)} &: \text{ These variables are not independent under the previous conditions.} \end{aligned} \tag{3.30}$$

or the same but using conditional probabilities as

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

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-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-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-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-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-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-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-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-value $\leq \alpha$ then,

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

   $\Rightarrow$ **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.

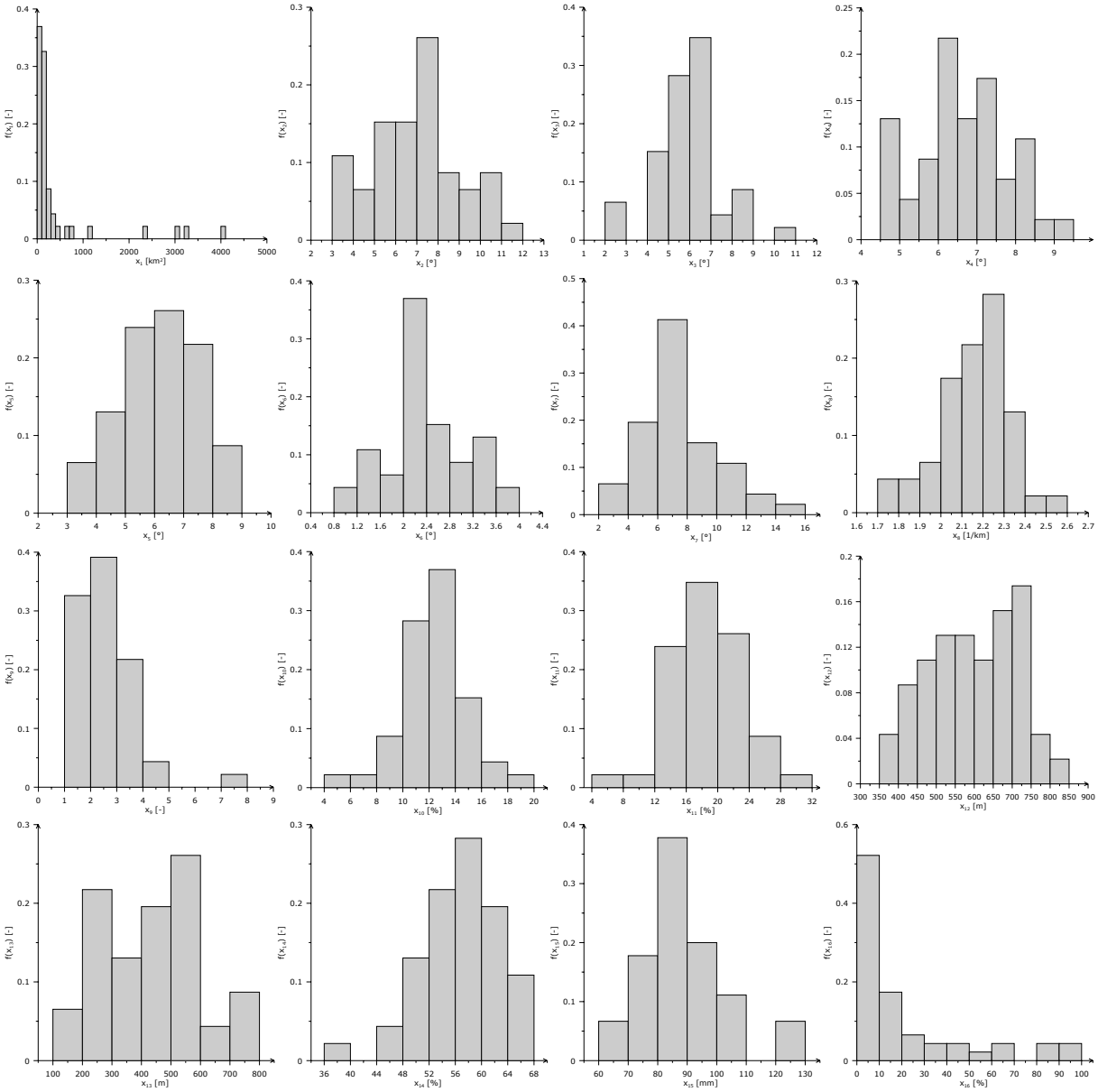


Figure 3.2 Histograms depicting the empiric PDF of all physiographic explanatory variables considered in this study.

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.

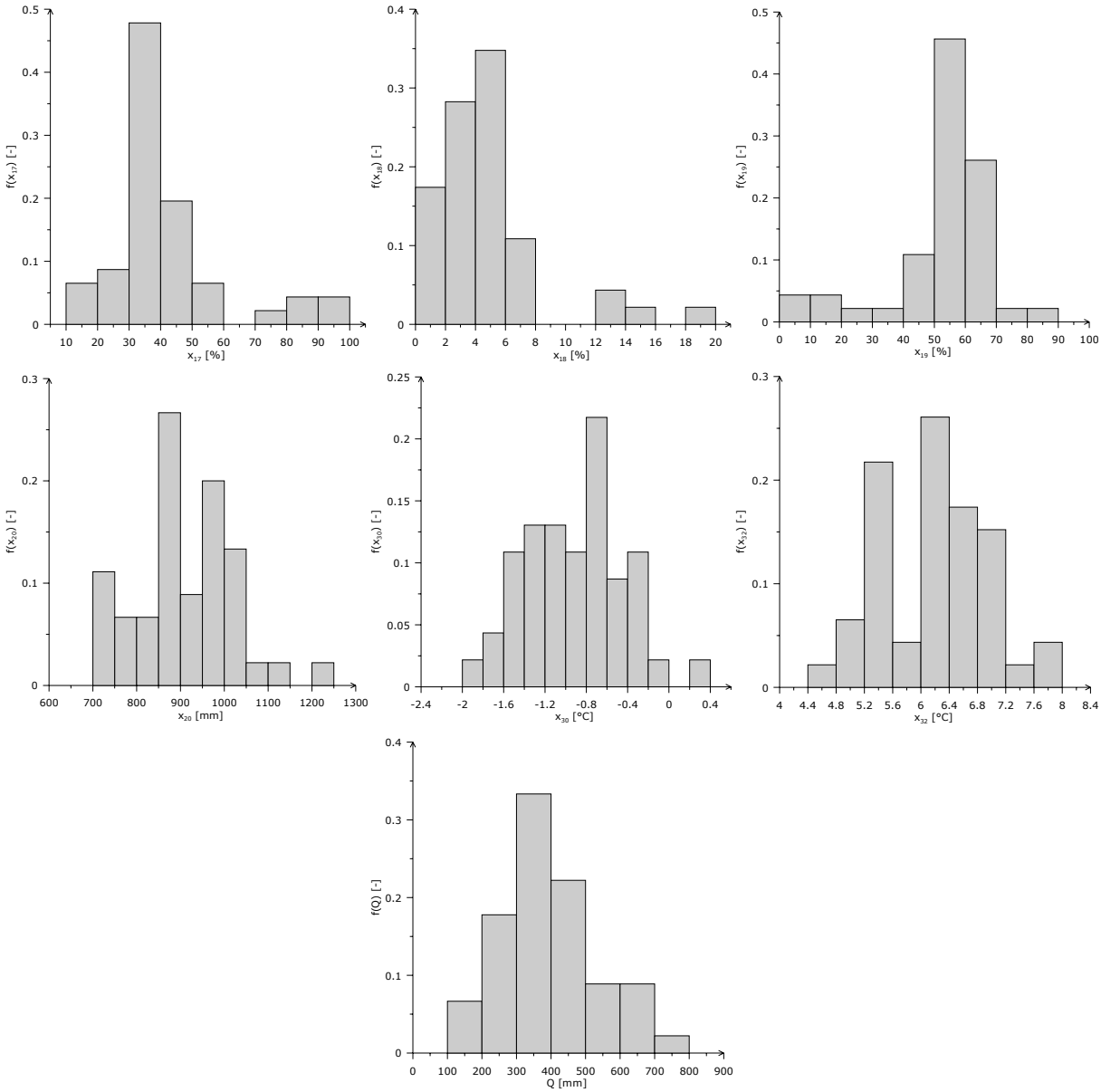


Figure 3.3 Histograms depicting the empiric PDF of the land cover and meteorological variables, as well as the specific annual discharge (the explained variable) considered as long term averages from 1961 to 1993.

It should be noted that both methods have ***only*** calculated $J(J+1)/2$ 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 $\{x_1,x_4,x_8,x_9,x_{11},x_{13},x_{15},x_{17},x_{18},x_{19},x_{20},x_{30}\}$ 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.

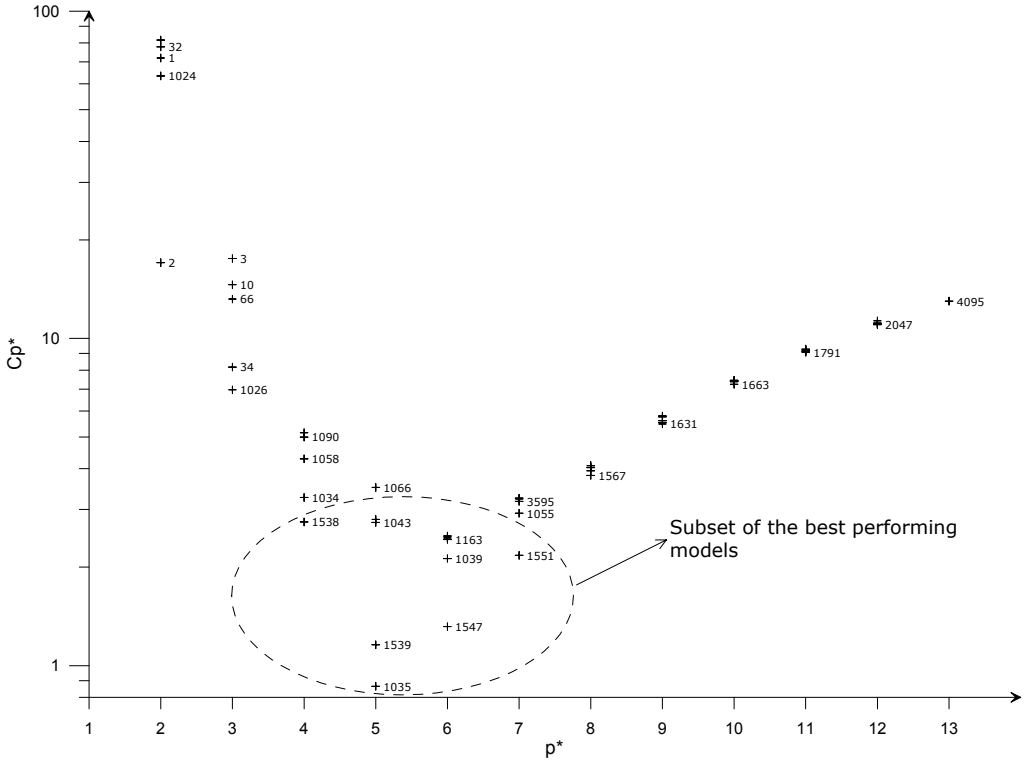


Figure 3.4 C_{p^*} vs. p^* plot showing the best 5 models for each p^* using the following variables $\{x_1,x_4,x_8,x_9,x_{11},x_{13},x_{15},x_{17},x_{18},x_{19},x_{20},x_{30}\}$. The number at the right of the marker (+) indicates the model's number. For models where $p^* \geq 8$ only the number of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

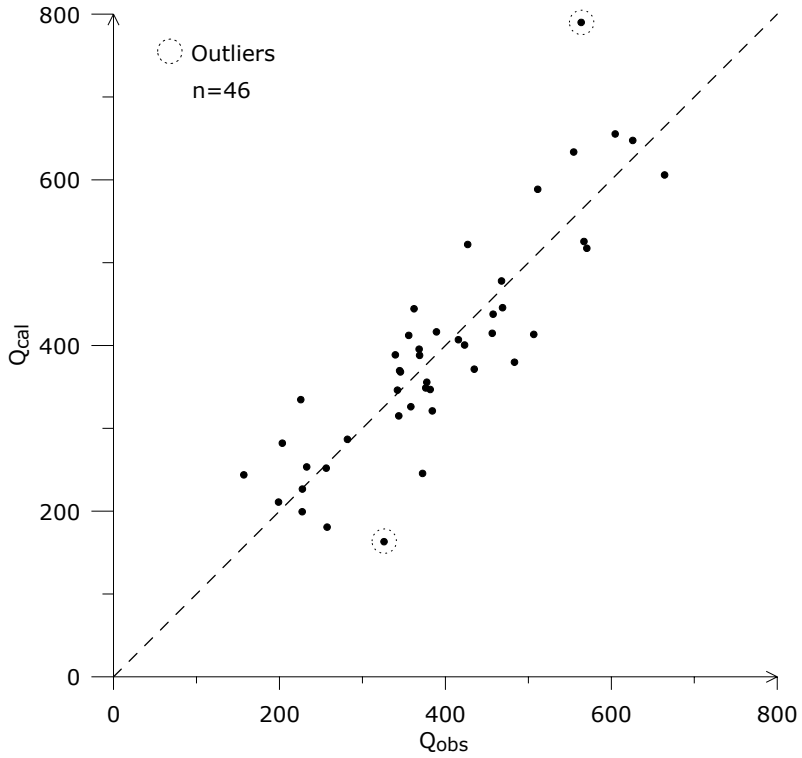


Figure 3.5 Scatterplot showing the relationship of Q_{obs} vs. Q_{cal} using the model (No. 1035) given by (3.35). A sample of size $\mathrm{n} = 46$ was used in the calculation. Outliers have not been removed.

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.

Chapter 4

Modelling Characteristics of the Runoff Process with Time-Dependent Data

4.1 Annual Specific Discharge

The influences of the land cover change, as was stated before, can only be detected when some variables involved in the model reflect the transformations occurred in the system during a significant time span (e.g. from 1960 to 1993). A reasonable time interval in which the climatic factors should be accumulated or evaluated seems to be a six-month interval, which corresponds to the water-seasons of a given year, i.e. winter and summer (see Section 2.7). By doing so, two important conditions can be fulfilled, namely: 1) the short-term auto correlation of climatic factors becomes insignificant; and, 2) the seasonal fluctuations of the climatic factors can be clearly set down.

Two models are to be formulated in order to attain the previous conditions, namely

$$Q_{il}^{t} = f\left(x_{i1}^{t},\ x_{i2}^{t},\ \ldots,\ x_{iJ}^{t}, \boldsymbol{\beta}\right) + \varepsilon_{i}^{t} \quad l = 2,3 \quad i = 1,\ldots,46 \quad t = 1961,\ldots,1993\,, \quad (4.1)$$

for winter ($l = 2$) and summer ($l = 3$) respectively. The selection of robust models fulfilling the constraints stated in Section 3.2 is to be described in following paragraphs.

4.1.1 Description of Time-Dependent Variables

At this stage and before any attempt to model the seasonal specific discharges (4.1) is carried out, it is useful to visualize the empiric PDF of the time-dependent factors for both winter and summer. Figure 4-1 shows histograms for the percentages of a given land cover type whereas Figure 4-2 depicts histograms of some climatic factors as well as specific discharge for winter and summer.

Figure 4-1 resembles the upper row of histograms shown in Figure 3.3, but there is an essential difference in the current ones. Histograms shown in Figure 4-1 do not depict the PDF of 33-year mean for each land cover type as it was in the previous case but rather than that the PDF of the time series of land cover types (see Figure 2.17) considering all spatial units. All distributions are unimodal and have a sample size equal to 184. Location and dispersion statistics for these distributions are summarised in Appendix 3. Comparing coefficients of variation among these three variables (i.e. land cover shares) it is clear that the variable representing impervious cover has the greatest value, and hence the largest relative dispersion of the data. This statement is also corroborated by the histogram depicting its empiric PDF (see Figure 4.1). The other two land cover variables are also skewed but in a lesser

degree than the distribution of "impervious" cover. The ranges of the sample PDFs for forest, impervious and permeable cover are [8.5, 98.7], [0.0, 31.0], and [1.3, 87.9] % respectively. Variables whose PDF are shown in Figure 4.1 have been evaluated at basin level, i.e. $\mathcal{L}_i \subseteq \Omega_i$ in equations (2.23) to (2.27).

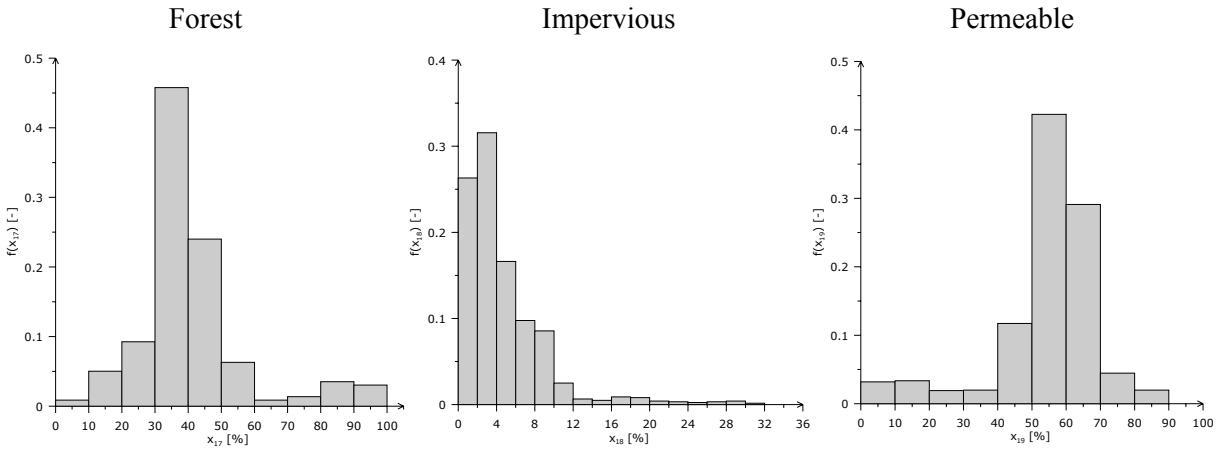


Figure 4.1 Histograms depicting the empiric PDF of the land cover types for all spatial units ($\mathcal{L}_i \subseteq \Omega_i$) from 1961 to 1993 (Number of observations for each histogram = 184).

The sample PDF for the specific precipitation in winter shown in Figure 4.2 exhibits a positive skewness whereas the PDF of this variable in summer is almost symmetrical. Both, the maximum and the minimum semi-annual specific precipitation occur in winter; hence, its standard deviation, as well as its coefficient of variation, in this season, is greater than that estimated in summer. In spite of this, the mean specific precipitation in winter is less than that in summer, and conversely, the mean specific discharge in winter is greater than that in summer. Due to this fact, the coefficient of variation of the specific discharge in summer is greater than that in winter. These characteristics of the water budget can be visualised in Figure 4.2. (Location and spread measures for all distributions shown in Figure 4.2 are summarized in Appendix 3). Such different behaviours of the water cycle fully justify the previous proposal [see point 2) above] to estimate two models, one for each water season.

PDFs for the maximum and the mean temperatures in January and July respectively are skewed and multimodal, but their relative variability in both cases during summer (July) is smaller than that in winter (January) (see the coefficient of variation in Appendix 3).

4.1.2 Assessing the Dimensionality of the System

In a complex system, such as the one being analysed here, where each explanatory variable x_j is mutually correlated with all the rest, it is very important to estimate the maximum number of variables a model should have in order to reduce as much as possible the effects of the existing *multicollinearity*. If a model has an excess of predictors, i.e. overparametrization, the sampling distributions of the estimated parameters $\hat{\beta}$ become very broad. This, in turn, may lead to confusions, errors in estimation, and even worse, to apparent contradictions when an estimated parameter comes up from the optimisation process with the opposite sign as the one expected (Rousseeuw and Leroy 1987, Wilks 1995). One viable approach to address such difficulty is presented below.

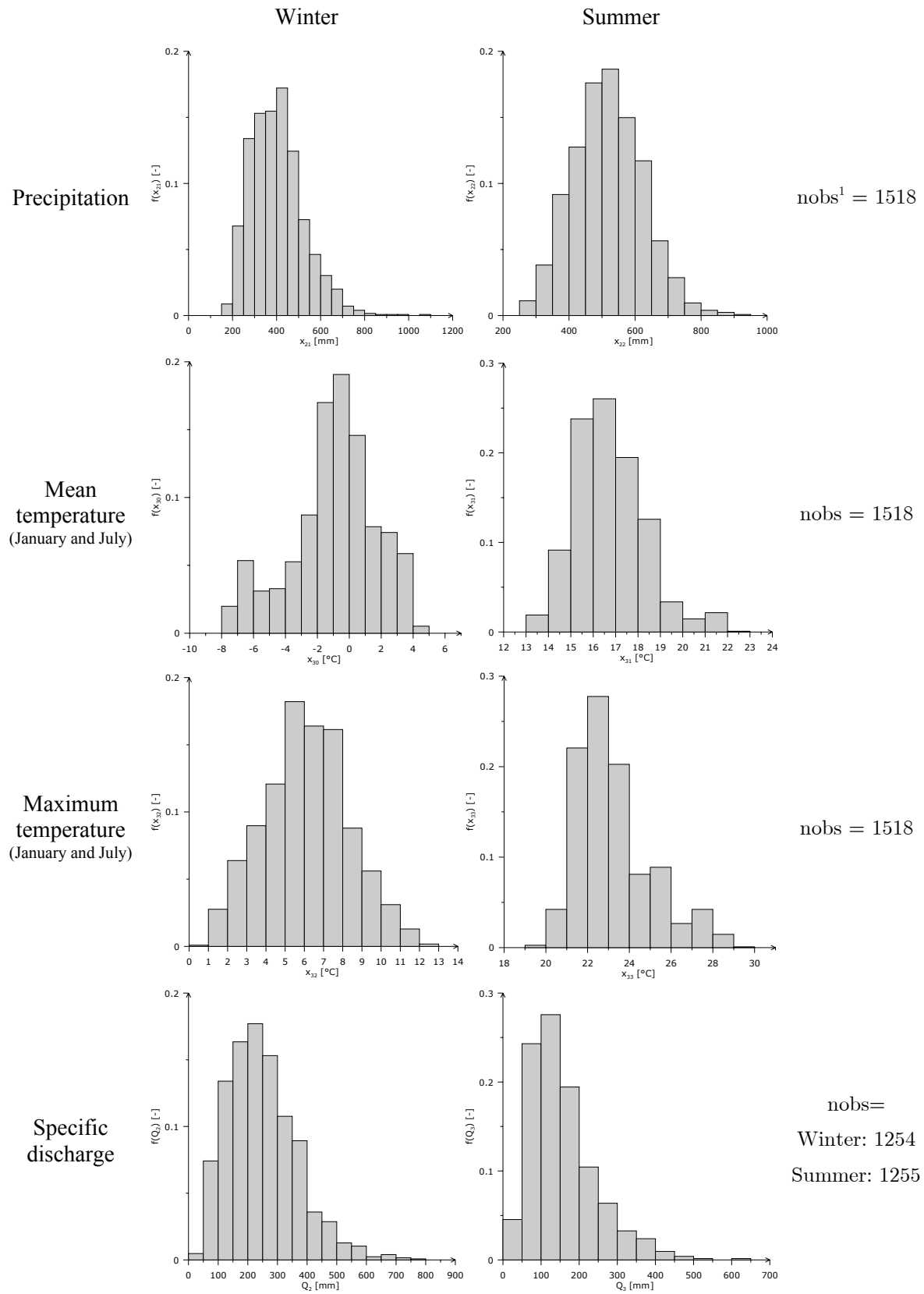


Figure 4.2 Histograms depicting the empiric PDF of climatic factors and specific discharge for all spatial units from 1961 to 1993.

[1] Number of valid observations in the corresponding sample.

Let the correlation matrix of all potential explanatory variables $(x_1, x_2, \ldots, x_J)$ be represented by $[\mathbf{R}]$, a non-singular and symmetric matrix. Based on this matrix, J eigenvectors $\mathbf{e}_j$ and their corresponding eigenvalues ω_j can be calculated, which should satisfy the equation

$$[\mathbf{R}]\mathbf{e}_j = \omega_j \mathbf{e}_j . \tag{4.2}$$

Subsequently, the eigenvalues are arranged in descending order, namely $\omega_1 \geq \omega_2 \geq, \ldots, \omega_J$. Based on them, the *dimensionality* of the system is the index k that satisfies the following relationship

$$\mathrm{v}(k) = \frac{\sum_{j=1}^{k} \omega_j}{\sum_{j=1}^{J} \omega_j} \geq \nu , \tag{4.3}$$

where $\mathrm{v}(k)$ is the proportion of the total variance retained by the first k eigenvectors and ν a threshold parameter. For instance, $\nu = 0.9$ means that at least 90% of the total observed variance in the system is described with k eigenvectors. Hence, it implicitly gives an insight into the maximum number of variables that a model should contain in order to retain a certain minimum amount of information describing the variability of the system. In general, ν lays within the interval $0.85 \leq \nu \leq 0.95$.

In the present case, the matrix $[\mathbf{R}]$ has been calculated using the following set of variables, $\{x_1, x_7, x_8, x_9, x_{11}, x_{12}, x_{14}, x_{15}, x_{16}, x_{17}, x_{19}, x_{21}, x_{30}\}$ $\forall\, i = 1, \ldots, 46 \quad t = 1961, \ldots, 1993$ whose results for the winter season are shown in Table 4.1. In this matrix, only those variables exhibiting the highest correlation with Q_2 have been included. For example, from the subset of variables describing slope, only x_7 has been selected because it has the highest correlation with the explained variable among the subset comprised by $\{x_2, x_3, x_4, x_5, x_6, x_7\}$. The same has been done with those describing aspects, elevation, temperature, and land cover.

Table 4.1 Correlation matrix $[\mathbf{R}]$ for the winter season. Additionally, a vector containing the correlation of each variable with the output variable Q_2 has been included at the left.

	Q_2	x_1	x_7	x_8	x_9	x_{11}	x_{12}	x_{14}	x_{15}	x_{16}	x_{17}	x_{19}	x_{21}	x_{30}
x_1	-0.0079	1.0000												
x_7	0.3501	-0.1185	1.0000											
x_8	-0.0599	0.1402	-0.6907	1.0000										
x_9	-0.2256	-0.0204	-0.0097	-0.0330	1.0000									
x_{11}	0.0619	-0.0788	0.5371	-0.7303	0.1269	1.0000								
x_{12}	0.3352	-0.0539	0.2628	-0.0726	-0.4182	-0.1512	1.0000							
x_{14}	0.1954	-0.1252	0.7575	-0.6671	0.0403	0.6237	0.1382	1.0000						
x_{15}	-0.3527	0.0570	-0.2276	-0.1519	0.1428	0.1786	-0.5277	-0.2578	1.0000					
x_{16}	0.3781	-0.1458	0.8233	-0.4100	-0.1788	0.2785	0.2877	0.4858	-0.1509	1.0000				
x_{17}	-0.1871	-0.1238	0.4692	-0.6627	0.2642	0.4883	-0.0914	0.5410	0.0618	0.0820	1.0000			
x_{19}	0.2174	0.1170	-0.4008	0.5710	-0.3122	-0.3990	0.1765	-0.4636	-0.0317	-0.0394	-0.9707	1.0000		
x_{21}	0.7100	0.0182	0.1166	-0.1259	-0.2379	0.0780	0.2805	0.0597	-0.0942	0.1059	-0.1150	0.1741	1.0000	
x_{30}	0.1336	0.0133	-0.0508	0.0179	0.0612	0.0131	-0.1605	-0.0189	0.0860	-0.0676	0.0321	-0.0517	0.1810	1.0000

Symmetric

As shown in Table 4.1, the correlation coefficients, either positive or negative, indicate that each explanatory variable is in higher or in lesser degree related with everything else. Based on this result it can be inferred that finding linear independent observables to describe a complex system seems to be improbable.

The eigenvalues of matrix $[\mathbf{R}]$ (i.e. for winter season) are

$$\mathbf{e}_j^{\mathrm{T}} = [4.295 \;\; 2.426 \;\; 1.282 \;\; 1.086 \;\; 0.970 \;\; 0.875 \;\; 0.691 \;\; 0.544 \;\; 0.391 \;\; 0.217 \;\; 0.154 \;\; 0.056 \;\; 0.014]\,.$$

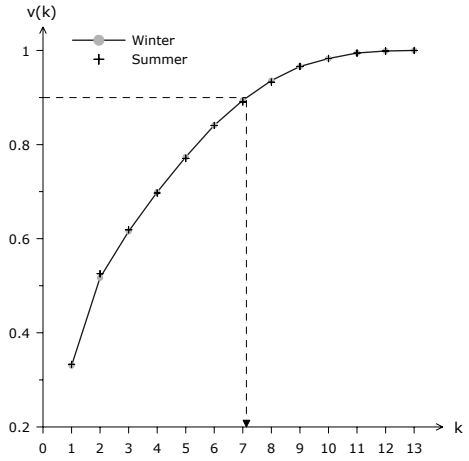


Figure 4.3 Curve showing the relative variance retained by the k first eigenvectors of the matrix $[\mathbf{R}]$ for winter. Additionally, the crosses show the results for the summer season. The correlation matrices have been calculated with time series from 1961 to 1993.

In order to assess the dimensionality of the system, it would be worthwhile to plot the index k versus $\mathrm{v}(k)$. Figure 4.3 illustrates the results of applying (4.3) to the previous eigenvalues. The horizontal dashed line in this Figure shows the threshold level chosen for this analysis, i.e. 0.9. This line, in turn, intersects the heavier line at a point whose abscissa lies in the interval [7,8]. The crosses depicted in Figure 4.3, which illustrate the values obtained for the summer season, show a very high level of agreement with the ones obtained for winter. This corroborates that the basic laws governing the system, either in winter or in summer, are the same, even if the climatic variables behave quite differently. Thus, the dimensionality of this system, given the available information, is about seven. This indicates that a conservative number of variables aimed to describe the system should be around this value in order to restrict, to a large extent, the existing and unavoidable multicollinearity amongst the explanatory variables.

Which variables should then be selected? One approach may be to use the first seven uncorrelated principal components as predictors as proposed by Jolliffe (1986). This option, although it filters the "noise" present in the data, has the following shortcoming: the principal components often have no physical interpretation, and thus would not allow in this case isolating the effects of land cover change. Instead, the method described before is to be proposed to tackle this issue. The next paragraph will describe this procedure in detail.

4.1.3 Finding a Robust Model

In essence, the selection procedure used in this case is quite similar to that employed in Section 3.3.8, although there are some differences, namely

1. Firstly, convex and continuously differentiable functions should be proposed. Three types are suitable for this case. The first one is a potential model (shortened to POT) that considers all possible explanatory variables as having nonlinear relationships with the explained variable. The second model type, thereafter called MLP1, regards the climatic variables x_{21} and x_{22} as the only ones having a nonlinear relationship with the explained variable whilst the rest are considered linearly related with the explained variable. Lastly, the third model type (shortened to MLP2) regards the land cover variables as the only ones exhibiting linear relationships with the output variable. These models can be written explicitly as

$$Q_{il}^t = \beta_0 \prod_j \left(x_{ij}^t\right)^{\beta_j} + \varepsilon_i^t \quad , \tag{4.4}$$

$$Q_{il}^t = \beta_0 + \sum_{\substack{j \\ j \neq j'}} \beta_j x_{ij}^t + \beta_{j'} \left(x_{ij'}^t\right)^{\beta_{j'}} + \varepsilon_i^t \quad , \tag{4.5}$$

and

$$Q_{il}^t = \beta_0 + \sum_{j \in \mathbf{U}} \beta_j x_{ij}^t + \beta_{J*} \prod_{\substack{j \\ j \notin \mathbf{U}}} \left(x_{ij}^t\right)^{\beta_j} + \varepsilon_i^t \ , \tag{4.6}$$

where

$$\mathbf{U} = \left\{x_j \ j = 17,18,19\right\}$$
$$l = 2,3$$
$$j, j' \in \{1,\ldots,J\}$$
$$j' = \begin{cases} 21 & \text{if} \quad l = 2 \\ 22 & \text{if} \quad l = 3 \end{cases}$$
$$J* = J + 1$$
$$J = 37$$
$$\beta_0, \beta_j, \beta_{J*} = \text{coefficients to be optimised.}$$

2. The estimators or objective functions to be minimised in both cases are twofold, one with $\varphi = 1$, and the other with $\varphi = 2$. This will allow assessing the sensitivity of the models with regard to existing outliers.

3. A weighting factor for each observation is to be used according to (3.13). For such equation the threshold $Z_c = 2.5$.

4. The goodness of the fit of all models pre-selected by both the Mallows' C_{p*} and the Jackknife statistics should be additionally assessed by the following quality measures (Bárdossy 1993, Lettenmaier and Wood 1993, Wilks 1995)

$$E_1 = \bar{\hat{Q}}_l - \bar{Q}_l \ , \tag{4.7}$$

$$E_2 = \frac{1}{n_0}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\hat{Q}_{il}^t - Q_{il}^t\right)^2 \ , \tag{4.8}$$

$$E_3 = \sqrt{E_2} \ , \tag{4.9}$$

$$E_4 = \frac{E_3}{\bar{Q}_l} \ , \tag{4.10}$$

$$E_5 = \frac{1}{n_0}\sum_{t=1}^{T}\sum_{i=1}^{n}\left|\hat{Q}_{il}^t - Q_{il}^t\right| \ , \tag{4.11}$$

$$E_6 = \frac{E_5}{\bar{Q}_l} \ , \tag{4.12}$$

$$E_7 = \frac{\mathrm{cov}(\hat{Q}_{il}^t, Q_{il}^t)}{\sqrt{\mathrm{var}(\hat{Q}_{il}^t)\,\mathrm{var}(Q_{il}^t)}} \ , \tag{4.13}$$

where

$$\bar{\hat{Q}}_l = \frac{1}{n_0}\sum_{t=1}^{T}\sum_{i=1}^{n}\hat{Q}_{il}^t \ , \tag{4.14}$$

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

$$\bar{Q}_l = \frac{1}{n_0}\sum_{t=1}^{T}\sum_{i=1}^{n}Q_{il}^t \ , \tag{4.15}$$

$\bar{\hat{Q}}_l =$ The mean of the calculated values based on the optimised model.

$\bar{Q}_l =$ The mean of the observed values.

$E_1 =$ The degree of correspondence of the calculated mean and the observed mean, often termed as BIAS.

$E_2 =$ Mean square error, or simply MSE, represents the mean of the square of the differences of the calculated and the observed values.

$E_3 =$ The positive square root of mean square error (RMSE).

$E_4 =$ The relative root mean square error (RRMSE).

$E_5 =$ The mean absolute error (MAE).

$E_6 =$ The relative mean absolute error (RMAE).

$E_7 =$ The Pearson product-moment coefficient of linear correlation (r) between $\hat{Q}_l$ and Q_l.

5. A supplementary criterion to assess the relative information contained in a given model compared with the so-called *saturated model* is to be incorporated into this analysis. The goal being that this criterion should complement and strengthen the selection of best performing models carried out by the Mallows' C_{p^*} statistic as well as the Jackknife estimator θ.

A suitable criterion constitutes the *Akaike Information Criterion* (or simply AIC), which was introduced by Akaike (1973) for evaluation of autoregressive models in time series analysis. According to Akaike, a statistic that is proportional to the sum of both the maximum log-likelihood of the model with respect to the observed data and its number of parameters provides an adequate basis for the comparative evaluation of the model. Within the context of this study, i.e. a model with j explanatory variables, the AIC can be calculated as follows (based on Venables and Ripley 1997)

$$\mathrm{AIC}_j = n_0 \ln\left(\frac{\Phi_{p^*}}{n_0}\right) + 2p^* , \tag{4.16}$$

where Φ_{p^*} and p^* have the same definitions as in Section 3.3.5. The best model according to the Akaike's criterion minimises AIC_j.

6. Each observation, either in winter or in summer, that is to be used to model (4.1) must satisfy a water budget constraint; otherwise, it will be considered as an outlier, and hence will be excluded from the optimisation process. Based on the continuity equation (i.e. conservation of mass), the water balance equation of a given basin during a time interval can be stated as follows: precipitation should be equal to the sum of evapotranspiration, runoff, water withdrawal from or water transfer to the basin (negative), and the change in water storage in both groundwater and surface reservoirs, all expressed in [mm].

   This balance of mass can be further simplified. Firstly, water withdrawals or transfers are not significant in the present case; and secondly, changes in water storage, whose estimation proves to be very difficult due to its non-steady character, can be neglected when the water balance equation is applied for long term intervals as is the case in the present study (Refsgaard et al. 1989, Dooge 1992).

   Based on these simplifications and the available statistical data for the Upper Neckar Basin (e.g. expected annual evapotranspiration is about 560 mm), two constraints can be formulated with a 99% level of significance

$$80 \le x_{i21}^t - Q_{i2}^t \le 190 \ [\mathrm{mm}] , \tag{4.17}$$

$$260 \le x_{i22}^t - Q_{i3}^t \le 590 [\mathrm{mm}] . \tag{4.18}$$

   The interpretation of (4.17) and (4.18) is as follows: the evapotranspiration in a given basin i and at time t should be greater than or equal to 80 and 260 mm, and less than or equal to 190 and 590 mm in winter and summer respectively, at the given level of confidence. Additionally, these constraints filter out information from those basins where the underground catchment does not match with its surface counterpart (e.g. derived from basin's topographic features), which in turn, induce severe problems in the water balance of the basin. This situation normally occurs in basins within karstic geological formations.

The procedure and criteria employed to select the best model and to rank them according to their degree of robustness and overall quality is described below.

Algorithm 5

1. Select $f(\bullet)$ and optimise[2] all possible models (i.e. $\min \Phi_{p^*}$) given a set of variables (e.g. in this case $J = 13 \Rightarrow 8191$ models) using two estimators: one with $\varphi = 1$, and another with $\varphi = 2$.
2. Select all models whose $C_{p^*} \leq C_{J^*}$; where C_{J^*} is the Mallows' statistic of the saturated model. These models constitute the subset of the best performing ones estimated for a given φ.
3. Calculate for the previously selected subsets the Jackknife statistics $\theta_{(\varphi=1)}$ and $\theta_{(\varphi=2)}$.
4. Rank models in ascending order with regard to their combined validation statistics $\theta = \theta_{(\varphi=1)} + \theta_{(\varphi=2)}$ and chose as the most robust model for a given functional type (POT, MLP1 or MLP2) the model that exhibits the minimum combined value.
5. The best model, and hence the most suitable function among the three attempted, is to be selected from the short list of robust models based on the results obtained for their respective quality measures [see (4.7) to (4.13)]. Additionally, all variables constituting the best model should have a p-value ranging from 5% to 10%.

The procedure described above as well as the method employed to optimise, select, test, and validate these models has been implemented within a set of programs written in Visual Fortran. These programs have been compiled along with a graphical user interface that helps the user through the modelling steps as can be seen in Appendix 6. The final product has been called **MDS**, which stands for **M**odel **D**evelopment and **S**imulation. Its modular structure would also allow including new subroutines and model types, if required, with minimum effort.

4.1.4 Selecting a Robust Model for Winter

The starting point consists of selecting among the available observables described in Chapter 2; those of them which are logically suitable to be considered as potential explanatory variables of the specific discharge in winter Q_2. These variables are in this case $\{x_j \quad j = 1,\ldots,19,21,30,32\}$. Afterwards, modified forward selection can be applied to rank this set of variables from the strongest to the weakest and then to use this information together with a correlation matrix derived from the same set

---

[2] The non-linear unconstrained optimization of the objective function Φ was carried out with the Generalized Reduced Gradient method originally proposed by Wolfe (1963) and later generalized by Abadie and Carpentier (1969) [There are many Fortran subroutines available for this method, e.g. in IMSL Fortran Libraries (1997), or the GRG algorithm, among others]. This procedure is iterative and employs a Hessian estimated by central differences and a quadratic extrapolation technique. The problem under consideration can be formulated as

$$\min \Phi = g(Q, f(\mathbf{x}, \beta))$$
$$\text{Subject to} \quad -\infty \leq \beta \leq \infty \quad ,$$

where g and f are convex and continuously differentiable functions.
In order to ease and speed up the convergence of the solution, the domain of the input data Q and $\mathbf{x}$, originally in $[0, \mathbb{R}^+]$ has been transformed to the interval $[\varepsilon, 1]$. Those values originally equal to zero have been modeled as a very small positive number, e.g. $\varepsilon = 1 \times 10^{-10}$, just to avoid likely indeterminations during the calculations. All parameters after the optimization are transformed back to their original domains.

of data to pre-select a short list of potential explanatory variables. This list, ranked according to the modified forward selection criterion, consists of $\{x_j \quad j = 21,16,11,19,15,7,8,14,12,17,9,30,18\}$. This procedure, as it happens with all stepwise algorithms, would not necessarily select the best model (Draper and Smith 1981). However, it can be used to reduce the size of potential predictors before all possible models are estimated.

The proposed method (Section 4.1.3) can be applied to this dataset aiming at obtaining a robust model for winter, which, in turn, delivers the results summarized in Table 4.2. It should be noted that this Table only shows the three best models for each type ordered in decreasing order of robustness (out of a total of 49,146 models generated and evaluated for winter).

From the original dataset, a number of outliers have been isolated by means of constraints given by (4.17). This, in turn, has reduced the sample size to 643. The nature of the high uncertainty present in those flawed observations cannot be addressed in this study, but in general, they can be attributed either to errors in measurement and/or interpolation techniques, or to divergence between the morphological and the underground catchments due to complex geological formations (e.g. a karstic formation).

The non-linear relationship between the C_{p^*} and the AIC statistics can be clearly seen in Figure 4.4. This result with respect to a model's performance implies that the Mallows' statistic is much more sensitive than the Akaike's information criterion. This does not mean that they show contradicting results. In fact, in both cases good models can certainly be found at low values. Due to this fact, further analysis will only show one of them as a measure of relative performance.

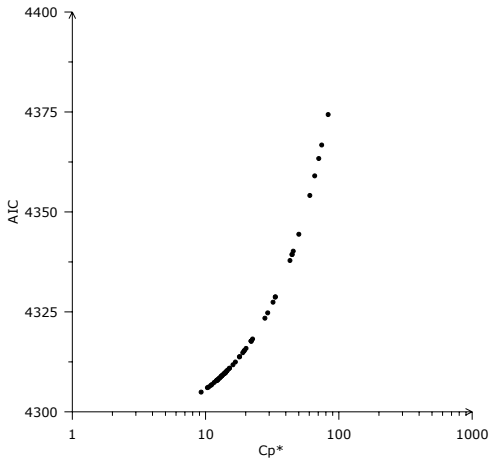


Figure 4.4 Curve depicting the non-linear relationship between the Mallows' C_{p^*} statistic and the AIC for the sample of best performing models described in Table 4.2. The best models in both cases exhibit small values.

From Table 4.2 it can be assessed that the most frequent variables within the subset of more robust models are those variables representing the specific seasonal precipitation, mean slope in floodplains and buffer zones of streams, mean field capacity, and fraction of south-facing slopes. Less frequent are the land cover related factors, but not by far with the latter. It can be also seen in this Table that there is no model within this subset that does not have at least one land cover variable.

The significance test for those models marked with a '✖' in Table 4.2 shows that all variables, with the exception of x_8, are definitely significant at the 10% level, and in some cases even at 1% . Hence, the null hypotheses can be safely rejected at the 10% level of significance in favour of the alternative hypotheses, i.e. these variables are certainly not independent from the explained variable. Results of the Monte Carlo simulations carried out with 500 replicates are shown in Table 4.3.

Table 4.2 Sample of the best models for winter (1 = a variable is included in the model, otherwise it is omitted). Values of the optimum estimators (minimum) with $\varphi = 2$ and $\varphi = 1$ are presented, as well as the results for the cross validation and the Akaike's information criterion. The most robust models are highlighted with the symbol ✖ . All values are dimensionless since the optimisation has been carried out in the interval $(0,1]$.

Model	x_7	x_8	x_9	x_{11}	x_{12}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{21}	x_{30}	$\varphi = 2$				$\varphi = 1$		Obs.
														Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																				
3729	1	1		1			1				1	1		0.967	12.6	4309.2	0.999	20.55	0.992	✖
3829	1	1		1	1	1	1		1		1	1		0.953	9.5	4306.1	0.986	20.24	1.004	
3837	1	1		1	1	1	1	1	1		1	1		0.949	8.5	4304.9	0.984	20.24	1.006	
Multilinear-potential models: MLP1																				
7827	1	1		1			1			1	1	1	1	0.940	5.1	4296.6	0.971	20.33	0.995	✖
7318	1			1			1		1	1		1	1	0.942	5.1	4296.6	0.970	20.35	0.996	
7315	1			1			1			1	1	1	1	0.942	5.1	4296.6	0.970	20.35	0.996	
Multilinear-potential models: MLP2																				
3733	1	1		1			1		1		1	1		0.934	4.8	4291.0	0.962	20.29	0.978	✖
3734	1	1		1			1		1	1		1		0.934	4.7	4291.0	0.962	20.29	0.983	✖
3731	1	1		1			1			1	1	1		0.934	4.7	4291.0	0.963	20.30	0.986	

It is important to remark that the best models presented in Table 4.3, which have been selected from thousands of possibilities because of their outstanding performance in comparison with the others, have between 6 and 8 explanatory variables. This range fits extremely well with the previously suggested number of variables that this system should have based only on the analysis of the dimensionality of the system.

Table 4.3 Results of the permutation test for models No. 3729, No. 7827, No. 3733 and No. 3734 using R=500. The tabulated figures are the Monte Carlo p-values as fractions. The estimator has been minimised with $\varphi = 2$.

Model	Type	x_7	x_8	x_{11}	x_{15}	x_{17}	x_{18}	x_{19}	x_{21}	x_{30}
3729	POT	$\simeq 0$	0.002	0.008	$\simeq 0$	-	-	$\simeq 0$	$\simeq 0$	-
7827	MLP1	$\simeq 0$	0.148	0.016	0.016	-	$\simeq 0$	$\simeq 0$	$\simeq 0$	0.080
3733	MLP2	$\simeq 0$	0.042	$\simeq 0$	$\simeq 0$	$\simeq 0$	-	0.008	$\simeq 0$	-
3734	MLP2	$\simeq 0$	0.064	$\simeq 0$	0.002	$\simeq 0$	0.002	-	$\simeq 0$	-

Subsequently, a model should be chosen among those shown in Table 4.3. Models number 3733 and 3734 are very good candidates since their estimators and validation indicators are the lowest and the second lowest according to Table 4.2. Both models are of type MLP2 and have in common all variables

with the exception of x_{18} and x_{19}. This means that both can be used depending on the requirements since the former relates the fraction of forest and permeable cover whereas the latter relates forest and impervious cover with the explained variable.

By inspection of Table 4.4 it can be established that both models (No. 3733 and No. 3734) perform much better than models No. 3729 and No. 7827 with regard to BIAS, MSE, RMSE, RRMSE, MAE, RMAE and r. Model 3734 is even better than model No. 3733 in some respects, but for practical purposes both can be used indifferently.

The potential model has the tendency to overestimate its predictions as can be inferred from the positive value of its bias (E_1). On the contrary, MLP1 and MLP2 models tend to underestimate predictions, though their bias is two or three orders of magnitude less than that of the potential model.

Table 4.4 Quality measures for the most robust models with $\varphi = 2$.

Model	Type	E_1 [mm]	E_2 [mm^2]	E_3 [mm]	E_4 [–]	E_5 [mm]	E_6 [–]	E_7 [–]
3729	POT	0.45	813.0	28.5	0.12	23.6	0.10	0.96
7827	MLP1	0.00	789.8	28.1	0.12	23.5	0.10	0.96
3733	MLP2	0.00	785.4	28.0	0.12	23.4	0.10	0.96
3734	MLP2	0.00	785.4	28.0	0.12	23.4	0.10	0.96

RMSE (E_3) or the square root of MSE (E_2) can be thought of as a typical magnitude for predicted errors, thus the lower the value the better the fit would be. Once again, selected models exhibit the lowest values. RRMSE (E_4) relates the overall magnitude of errors with the mean of all observations, and therefore can be expressed as a percentage. In this case, the error of MLP2 models is 12.16% with respect to the mean of the observations. This value is more sensible to outliers because it is derived from the MSE. In this case also the lower the value the better the fit is. MAE and RMAE (E_5 and E_6 respectively) are less sensitive to errors as compared with MSE and RMSE respectively. The percentage error with respect to the mean is in this case equal to 10.16%. Finally, the correlation coefficient (E_7) confirms what has been stated before, i.e. that models No. 3733 and No. 3734 are among those models showing a high correlation but not the highest, which always corresponds to the saturated model. The interpretation of this quality measure should be done cautiously since it reflects the association between observed and calculated values but does not account for biases present in the predictions (Wilks, 1995).

Based on all these results, it can be stated that multi-linear models have performed much better than the pure potential one. Moreover, models having precipitation as the only variable of the potential sub-model and the rest in the linear one are in general better than pure potential models; but, they are not as good as those having only land cover in the linear sub-model. This, in turn, indicates that based on the evidence provided by the sample, land cover factors are linearly related with the total specific discharge in winter at a high degree of certainty, say at least 99%.

The optimised parameters for both models are shown in Table 4.5. Both potential sub-models have almost the same values and share the same sign. However, that does not occur in the linear sub-

models. The signs of these coefficients correspond with the perception one can have about this natural system. For instance, precipitation and mean slope without doubts should have a positive sign. In other words, the higher their values, the bigger the specific discharge from a given basin will be. Field capacity, on the contrary, should have a negative sign because the higher its average value, the bigger the quantity of water stored in the soil matrix, and hence, the lesser the expected runoff.

Table 4.5 Optimized parameters (with $\varphi = 2$) for models No. 3733 and No. 3734.

Model	β_0	β_{17}	β_{18}	β_{19}	β_{J*}	β_7	β_8	β_{11}	β_{15}	β_{21}
3733	36.783	-1.1663	-	-0.8487	0.2227	0.0903	0.2051	0.0887	-0.1149	1.1987
3734	-47.587	-0.3159	0.8551	-	0.2186	0.0904	0.2078	0.0898	-0.1156	1.2010

Regarding the sign of land cover variables, one could expect based on hydrological considerations that forests and permeable covered surfaces (e.g. grassland, cropland, meadows, etc.) have to have both higher evapotranspiration and infiltration rates than impervious covered surfaces. Additionally, the overall roughness of the former is higher than that of the latter, and hence, longer concentration times and lesser runoff volumes can be expected. Due to this rationale, forest and permeable cover would tend to reduce the seasonal specific yield (thus, a negative sign should be expected in the case of a linear sub-model) whereas impervious cover would tend to evaporate less and hence increase the seasonal specific yield (thus, a positive sign should be expected in a linear sub-model).

Although it is sometimes difficult to interpret signs of the terms in empirical models, mainly because of multicollinearity among explanatory variables, the selected models agree with the assertions mentioned above.

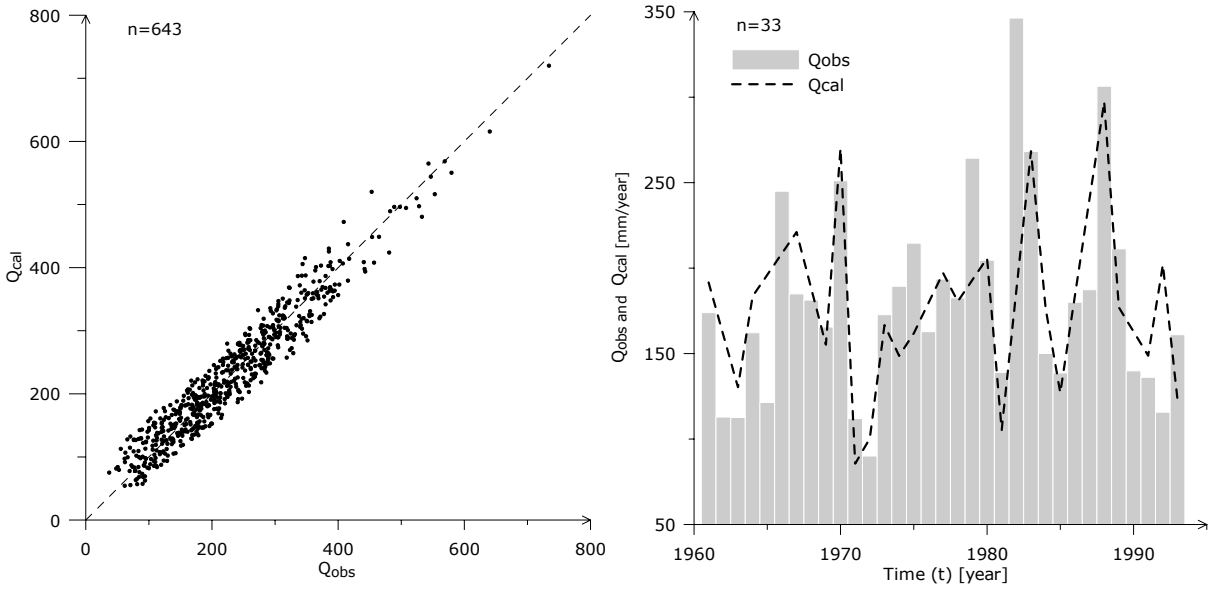


Figure 4.5 At the left panel, a scatterplot shows the relationship between observed and calculated values using model No. 3733 for winter. The samples size is 643. The right panel illustrates a time series of the observed specific discharge in winter and their corresponding calculated values for Basin No. 13.

The quality of the fit achieved by one of the proposed models (e.g. No. 3733) can be visualized in the scatterplot shown in Figure 4.5 (left panel). At the right panel of Figure 4.5, a time series of both the

observed specific discharge in winter for basin No. 13 located within the Study Area and the corresponding predicted values are displayed. This graph shows that the model No. 3733 has been able to simulate the positive trend present in the observed data and relates it with land cover variables apart of climatic and morphologic factors. It does not estimate, however, quite accurately some peaks and low values present in the time series.

Additionally, a plot of the standardized errors versus observations is shown in Figure 4.6 (right panel). This figure is very important because it illustrates at first glance that the errors are homoscedastic at least in the interval about [50,450] [mm]. Outside this interval, since there are few observations, nothing can be inferred; however, it is assumed that they also have the same error distribution. As was stated earlier, errors should be randomly distributed with zero mean and constant variance (i.e. be homoscedastic); otherwise, a proposed model is considered biased.

A plot depicting the distribution of standardized residuals over the time axis is also important when dealing with time series because it can help to spot long term cyclic variation patterns. For the model No. 3733 (model No. 3734 as well), as it is shown in Figure 4.6 (left panel), that does not seem to be the case. Based on both graphs shown in Figure 4.6, it can be concluded that the proposed model complies with those conditions stated above.

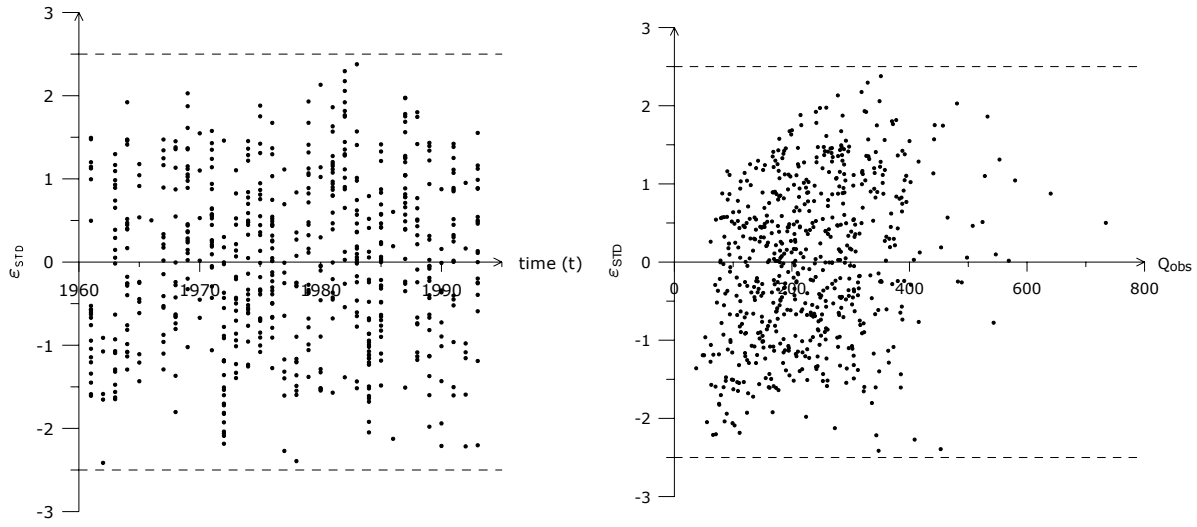


Figure 4.6 The left panel shows a plot of the standardized residuals for winter obtained with model No. 3733 versus time. At the right panel, a standardized residual plot for the same model is presented.

4.1.5 Selecting a Robust Model for Summer

Selecting a model that fits the observed specific discharge for summer during the period 1.11.1960 to 31.10.1993 for the Study Area based on observables described before would involve the calculation of $2^{22}-1$ possible combination of variables, and thus an equal number of likely models. Such a demanding task with regard to computing time can be simplified in the following way.

Firstly, a correlation matrix relating $\{(Q_3, x_j) \ \ \forall j = 1,\ldots,19,22,31,33\}$ was calculated based on the existing dataset that fulfils the constraints given by (4.18). This dataset has a cardinality equal to 1150.

Using this information and the criteria explained and used before (e.g. Section 3.3.8), variables having the highest correlations with the explained variable were pre-selected to form a short list of observables with which a robust model is to be found. This short list should also contain the first J strongest variables (limited here to 12 because of computing limitations) according to the modified forward selection procedure. This short list ordered from the strongest to the weakest is composed of $\{x_j \quad j = 22,15,7,14,17,9,16,18,13,33,19,10\}$. This pre-selection presupposes that variables having very little correlation with the explained variable would not contribute much to explaining the observed variance of Q_3, while on the contrary, they would complicate the calculation by increasing the computing time, introducing 'noise' to the solution, and probably increasing the multicollinearity. It should be observed that these variables fulfil all conditions stated in (3.2) regarding the components of the system.

Since likely effects of land cover are to be disclosed, three variables have been taken into account, namely $\{x_j \quad \forall j = 17,\ldots,19\}$. These variables, with the exception of variable x_{18}, have been evaluated at basin label (i.e. $\mathcal{L}_i \equiv \Omega_i$).

Based on the correlation matrix, it was found that the correlation coefficient between Q_3 and x_{18} depends on the domain where the latter is evaluated. For instance, if the fraction of impervious land cover (x_{18}) is estimated at a domain comprised by riparian zones and floodplains along the stream network (i.e. $\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$), then its correlation coefficient with Q_3 is about 8.4 times greater than that obtained if this variable is evaluated at basin level (i.e. $\mathcal{L}_i \equiv \Omega_i$). An explanation for such fluctuation is the fact that new settlements, industrial states, and major transportation infrastructure within the Study Area tend to be closer to both existing transportation axes and traditional urban agglomerations which, according to historic evidence, have a great probability to be located along the valleys with moderate slopes that surround main rivers and their tributaries. On the contrary, it is very unlikely that land use types with a higher percentage of impervious areas would occur at a random place with poor accessibility and sheer slopes. Thus, estimating the fraction of impervious areas within a catchment using its whole area may underestimate the effects of this land cover on the hydrological cycle and hence the impacts of its change over time. This is, in turn, reflected by its low coefficient of correlation. Conversely, if the reference area becomes smaller and additionally is set to correspond to highly sensitive ecosystems as those mentioned above, the correlation coefficient increases. Because of that x_{18} has been evaluated in this case within the domain $\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$.

Moreover, it was also found that the correlation coefficient between x_{18} and Q_3 in winter does also depend on the area of reference of the former variable, but in this case the opposite occurs, namely $\mathrm{r}(x_{18}(\mathcal{L}_i \equiv \Omega_i), Q_2)$ is 1.6 times greater than $\mathrm{r}(x_{18}(\mathcal{L}_i \equiv \mathcal{B}_i), Q_2)$.

A summary of the results obtained after applying the proposed method (see Section 4.1.3) to the variables of the short list is shown in Table 4.6. This Table reveals that the uncertainty of the system in summer is much higher than that in winter, and because of that, a model in general requires more variables to explain the observed variance; for instance, the minimum number of variables in this case was eight whilst the most robust model found (No. 3965) has ten explanatory variables. Because of the high uncertainty of the system in summer, optimum estimator values (see Tables 4.2 and 4.6) are higher in summer than those in winter, and so are the cross-validation statistics.

From Table 4.6 two models have been selected according to the guidelines mentioned above, namely: model No. 3965 and No. 3967, whose types are POT and MLP2 respectively. The significance tests displayed in Table 4.7 show that all variables, with the exception of x_{33} in model No. 3967, are significant at 10%. This drawback makes the latter less reliable than model No. 3965.

Table 4.6 Sample of the best models for summer (1 = a variable is included in the model, otherwise it is omitted). Values of the optimum estimators (minimum) with $\varphi = 2$ and $\varphi = 1$ are presented, as well as the results for the cross validation and the Akaike's information criterion. The most robust models are highlighted with the symbol ✖. All values are dimensionless since the optimisation has been carried out in the interval $(0,1]$.

Model	x_7	x_9	x_{10}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{22}	x_{33}	$\varphi = 2$				$\varphi = 1$		Obs.
													Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																			
3965	1	1		1	1	1	1	1	1		1	1	7.249	9.9	8143.4	7.433	70.83	7.501	✖
4093	1	1	1	1	1	1	1	1	1		1	1	7.246	11.5	8145.0	7.449	70.82	7.493	
3967	1	1		1	1	1	1	1	1	1	1	1	7.246	11.5	8145.0	7.443	70.81	7.524	
Multilinear-potential models: MLP1																			
3967	1	1		1	1	1	1	1	1	1	1	1	8.244	12.2	8291.0	8.457	74.97	8.556	
4095	1	1	1	1	1	1	1	1	1	1	1	1	8.242	14.0	8292.7	8.476	75.03	8.540	
3455		1		1	1	1	1	1	1	1	1	1	8.279	15.0	8293.8	8.477	75.09	8.560	
Multilinear-potential models: MLP2																			
3967	1	1		1	1	1	1	1	1	1	1	1	7.518	16.6	8188.1	7.736	71.49	7.791	✖
4095	1	1	1	1	1	1	1	1	1	1	1	1	7.487	14.0	8185.5	7.719	71.48	7.809	
4028	1	1	1		1	1	1	1			1	1	7.567	19.9	8191.4	7.762	71.77	7.791	

Table 4.7 Results of the permutation test for models No. 3965 and No. 3967 using R=500. The tabulated figures are the Monte Carlo p-values as fractions. The estimator has been minimised with $\varphi = 2$.

Model	Type	x_7	x_9	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{22}	x_{33}
3965	POT	$\simeq 0$	$\simeq 0$	0.012	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	-	$\simeq 0$	0.054
3967	MLP2	$\simeq 0$	$\simeq 0$	0.004	$\simeq 0$	$\simeq 0$	$\simeq 0$	0.010	0.060	0.018	$\simeq 0$	0.194

The reliability of the potential model is confirmed by comparing the quality measures shown in Table 4.8. According to these results, model predictions in both cases tend to underestimate observations since their respective bias (E_1) is negative. The potential model has a bias whose absolute value is greater than that of the multi-linear one, but its relative root mean square error (E_4) is a bit smaller than that of the latter (i.e. about 25.3% and 25.8% respectively). Additionally, the correlation coefficient between observed and calculated values for the potential model ($E_7 \approx 0.87$) is almost as high as that obtained for the saturated one. This is a good advantage because having two variables less makes a model relatively simpler.

Based on these arguments, it seems adequate to opt for model No. 3965 instead of model 3967. The optimised parameters for the chosen model are shown in Table 4.9. It is important to emphasize that two land cover variables, i.e. forest and impervious cover, have been selected by the proposed algorithm as significant variables to explain the annual specific discharge of a basin during summer. Their relative influence on the system is somehow reflected in this model by the order of magnitude of

the coefficients and their signs. The fraction of forest cover within a spatial unit has a coefficient in model No. 3965 whose absolute value is one order of magnitude higher than the coefficient for the fraction of impervious cover evaluated in $\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$, i.e. within a buffer zone of the stream network.

Table 4.8 Quality measures for the most robust models with $\varphi = 2$.

Model	Type	E_1 [mm]	E_2 [mm^2]	E_3 [mm]	E_4 [–]	E_5 [mm]	E_6 [–]	E_7 [–]
3965	POT	-0.11	1515.5	38.9	0.25	31.0	0.20	0.88
3967	MLP2	-0.01	1580.9	39.8	0.26	31.4	0.20	0.87

Both coefficients have negative signs, which may have the following interpretation. Land cover variables in this study are indicators of both intensity and type of land-atmosphere interactions. Forested areas would tend to evaporate more water than those portions of the basin with other land cover types (e.g. impervious, grassland, cropland) under the same climatic and morphologic conditions because of the high transpiration rates attributed to the tree physiology. This assertion has been confirmed by long-term controlled catchment experiments in several locations around the globe and with different types of tree species. Studies carried out or reported by Law 1956, Bosch and Hewlett 1982, Kirby et al. 1991, Eeles and Blackie 1993, and Jones 1997 indicate that afforestation would lead to a considerable reduction of annual runoff yield, or conversely, that deforestation would augment the yield of a given catchment. Such conclusions imply an inverse relationship between x_{17} and Q_3 or between x_{17} and Q_2. This kind of inverse relationship is represented in model No. 3965 by the negative exponent of variable x_{17}.

Table 4.9 Optimized parameters (with $\varphi = 2$) for model No. 3965.

Model	β_0	β_7	β_9	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{22}	β_{33}
3965	20.235	0.6473	0.1346	0.0954	-1.8215	-0.6539	0.0066	-0.2994	-0.0161	1.9459	-0.2331

As stated before, impervious areas would evaporate water to the atmosphere due to the absorption of heat provided by the sun, but in much smaller amounts than the latter because they lack of a very important component of the evapotranspiration process, namely the transpiration of vegetal tissue. As a result, a higher yield should be expected at the outlet of such areas. This relationship is denoted in model No. 3965 by the negative sign of the exponent of variable x_{18}, and its smaller absolute value in comparison with that of variable x_{17}. In fact, these exponents are in the following ratio $\beta_{17} : \beta_{18} = 18.7 : 1$.

It is noteworthy to express that the relationship between land cover variables is certainly highly non-linear in summer, whereas in winter, due to almost no physiological activity of vegetation, the relationship between specific discharge and land cover is very close to linear. This is why a multi-linear potential model containing these variables in the linear sub-model was chosen as the most robust one in winter, whereas in summer, all models of type MLP2 and MLP1 performed badly compared with those of type POT (see Table 4.6) with the additional advantage, in general, that the

latter needs less variables than the former. Because of this, a potential model was selected as the most robust one based on the available data.

Other variables such as x_7 or x_{15} appear in almost all models both in winter and summer (see Tables 4.2 and 4.6). According to the selected models, the following assertions can be done. Firstly, the higher the mean slope within $\mathcal{B}_i \subset \Omega_i$ is, the higher the seasonal runoff yield of the basin Ω_i would be, and secondly, the higher the mean field capacity of the basin, the lower its specific discharge. These statements make sense also from a theoretical point of view.

The goodness of the fit achieved by model No. 3965 can be visualized by the scatterplot depicted in Figure 4.7 (left panel) or by means of a time series shown in Figure 4.7 (right panel) which relates predicted and observed values for the basin No. 13 within the Study Area. The latter shows clearly that model No. 3965 is able to simulate the majority of peaks and valleys of the observed time series based on the input data. The cases where the model has failed may indicate an underestimation of the spatial distribution of precipitation. In these cases, the proposed model has also been able to simulate the positive trend observed in the data (see Figure 4.7 right panel).

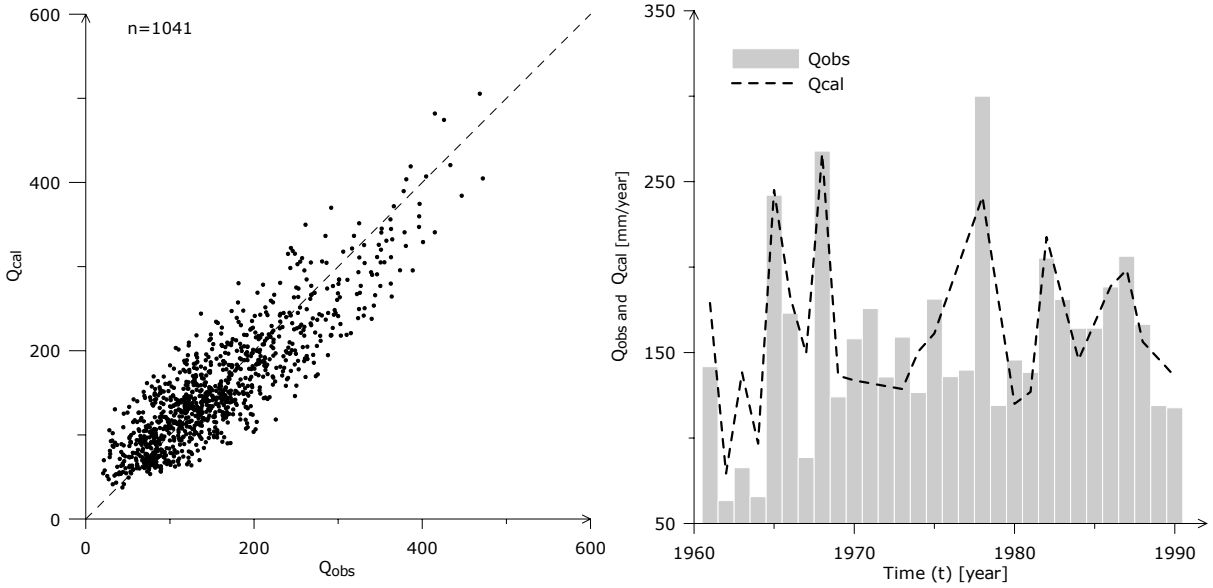


Figure 4.7 The left panel shows a scatterplot of the observed values versus calculated ones for summer obtained with model No. 3965. The right panel illustrates a time series of the observed specific discharge in summer and their corresponding calculated values for Basin No. 13.

4.1.6 Visualizing the Effects of Land Cover Change on Annual Runoff

A good example for visualizing the effects of land cover change is the drainage area of the River Körsch (in the present study named as Basin No. 13), whose gauging station is located at Denkendorf-Sägewerk. This area, because of its vicinity to Stuttgart, has endured a fast land use change triggered mainly by anthropologenic driving forces. Because of them, impervious areas have grown from about 7.3% of the total area in 1961 to about 30.9% in 1993. That means an average annual growth rate of about 4.6%. Forest grew slowly since 1961 to the middle of the 70s and then a smooth decline has begun as can be seen in the graph on top of Figure 4.8.

During the same period, precipitation in this basin has endured a continuous decline as it is illustrated by the trend line shown in Figure 4.8 (dashed line). This climatic factor, which is composed of x_{21} and x_{22} in the present case, has a marked periodicity but, in general, its average is decreasing at the rate of 1.1 mm/year. Conversely, the seasonal specific discharge has increased at the rate of 0.83 mm/year during the same period (see the graph at the bottom of Figure 4.8).

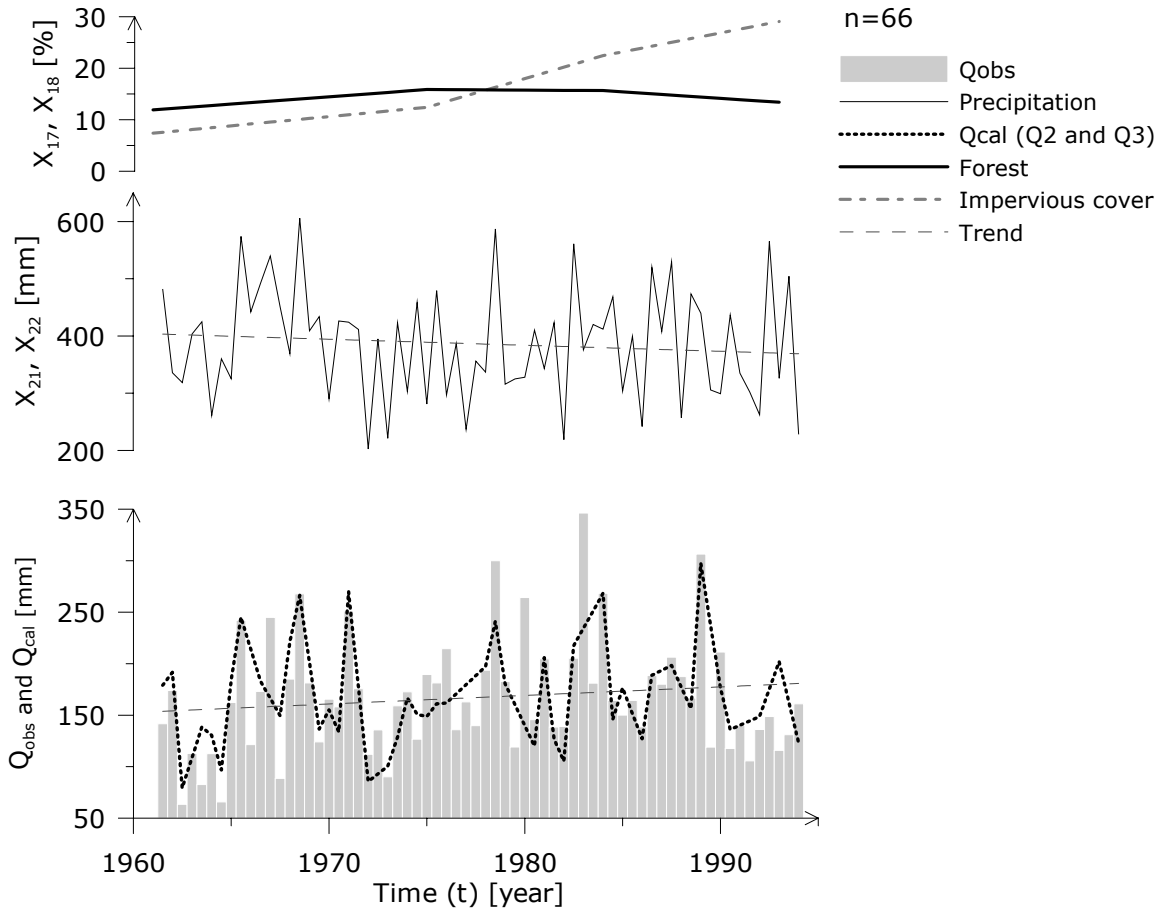


Figure 4.8 Comparison of time series of land cover, precipitation and specific discharge in winter and summer for Basin No. 13. Calculated values using models No. 3733 for winter and No. 3965 for summer are also displayed.

Based on these facts, and considering that other factors are quasi-constant or reveal no trend at all, an upward tendency of the specific discharge can only be attributed to influences stemming from land cover changes occurring in the basin since 1961. This assertion has been corroborated by the models presented before. They not only predict an upward trend as can be seen in Figure 4.8, but they also relate the specific discharge with two land cover variables, whose tests of independence with the explained variable can be rejected even at levels of significance lower than 1% according to the Monte Carlo simulations carried out.

Moreover, it should be noted that the selected models represent a regionalization for all basins within the study area, and because of this, the models might fail to predict with high certainty a peak or a nadir at a given time point. However, they have an advantage; i.e. they can perceive upward or downward tendencies of those variables included in the model, and hence, predict an expected value for the explained variable based on such trends.

4.2 Specific Peak Discharge

4.2.1 Description of Some Time-Dependent Variables Employed

In the present section, variables that have not been described before and are deemed potential predictors for peak flows within a basin are to be described. According to Chow (1964) and others, suitable potential predictors for peak flows are antecedent indices for both precipitation and temperature. In the present case, their maximum seasonal values will be employed because of their high correlation with the explained variable. Figure 4.9 illustrates the PDFs of such variables for winter and summer as well as the PDFs of the observed peak flows.

The PDFs of the maximum API for both winter and summer have a skewness of approximately 1.1 and 0.8 respectively, which means that they are clearly skewed to the right as can be seen in Figure 4.9. Their dispersion is, however, different in winter from that of summer. In fact, the range, the standard deviation, and the coefficient of variation in winter are higher than the corresponding figures in summer (see Appendix 3). The parameters on which API is based have been calibrated so that the maximum correlation with the explained variable can be achieved. So, for winter the parameters are $\kappa = 0.95$ and $C = 90[\mathrm{days}]$, whereas for summer $\kappa = 0.85$ and $C = 30[\mathrm{days}]$.

With regard to maximum ATI, its PDF in winter is almost symmetrical (skewness equal to 0.2), while in summer it is positively skewed (0.9). This index has been evaluated using temperature in degrees Kelvin [K] for the convenience of having positive numbers. The range of this variable is very small in both winter and summer, although the range in winter is higher than that in summer. The coefficients of variation are quite small compared with other variables, which may indicate that this variable is of little use in explaining the variance of the specific peak flow.

Finally, Figure 4.9 shows, at the bottom, the PDFs of the specific discharge in winter and summer, which are the explained variables in this section. These variables have a skewness of about 1.9 and 3.5 for winter and summer respectively. The kurtosis of these variables are very high also, namely 7.9 and 24.0, for winter and summer respectively. In other words, their PDFs are very peaky and positively skewed. In reality, such distributions show that very high values may occur but their probability is very small. The challenge is then to determine whether the occurrence of these high values is somehow linked with the land cover variables.

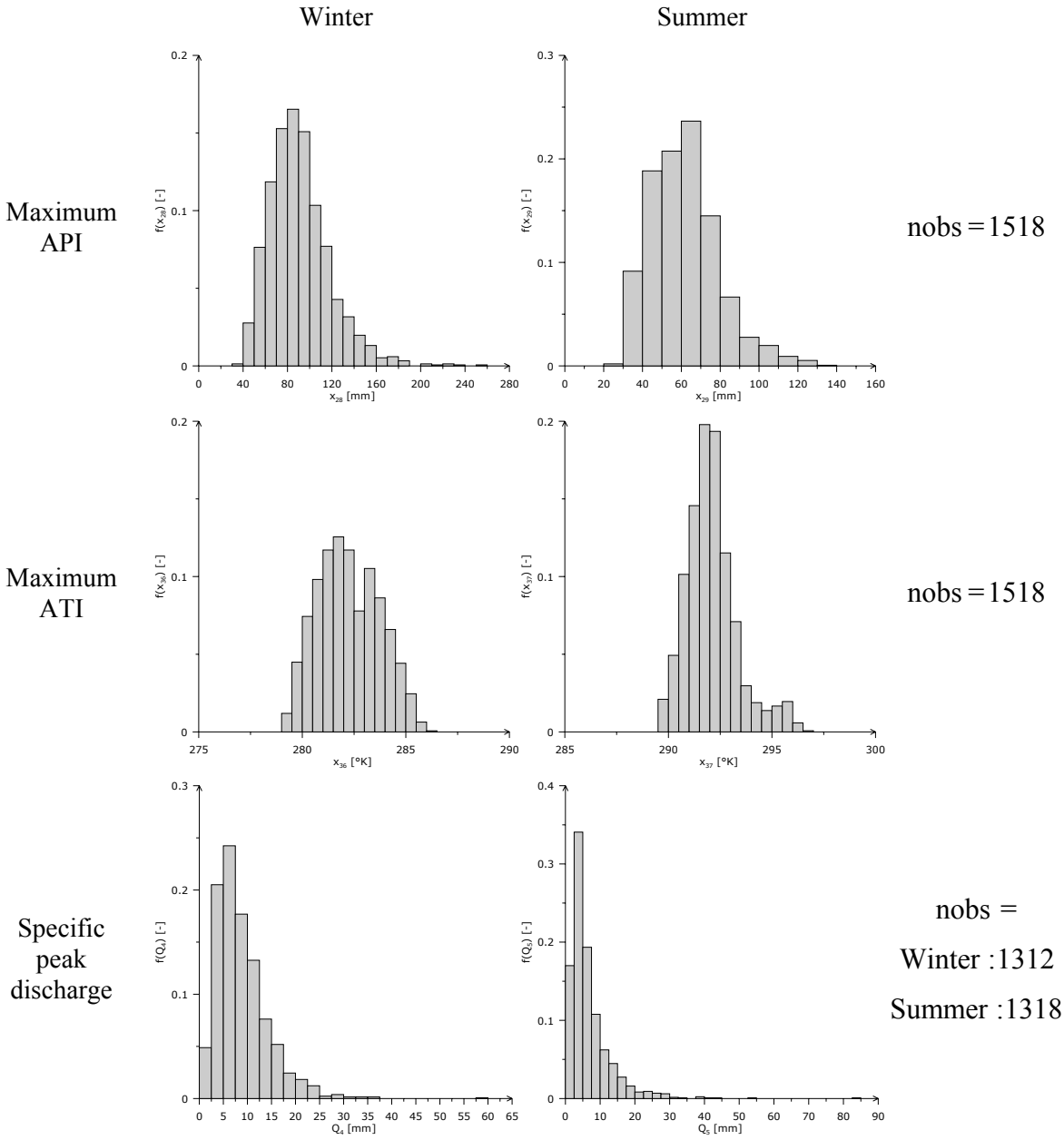


Figure 4.9 Histograms depicting the empiric PDFs for both maximum API and ATI indices for winter (left panel) and summer (right panel), as well as the specific peak discharge considering all spatial units during the period from 1.11.1960 to 31.10.1993.

4.2.2 Selecting a Robust Model for Winter

The first step consists of selecting potential predictors of the explained variable from the available dataset. In this case, specific peak flows in winter (Q_4) are assumed to have functional relationships with the following set of predictors based either on previous experience or common sense, namely $\{x_j \; j = 1,\ldots,19,21,24,28,30,32,36\}$. This long list of predictors should be shortened somehow because of the reasons already explained. Applying the same procedure used before, a short list composed of the twelve strongest predictors was found, i.e. $\{x_j \; j = 28,12,15,19,30,9,16,17,3,1,11,18\}$. This short list of predictors does not only simplify the calculation proposed in paragraph (Section 4.1.3), but also satisfies the restriction established by (3.2). The cardinality of the sample data to be employed consists

of 1182 valid observations spread along the time axis from 1.11.1960 to 31.10.1993. In this case also, all land cover variables, i.e. $\{x_j \ \forall j = 17,\ldots,19\}$ have been evaluated within the domain $\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$ due to the same reasons explained in Section 4.1.5.

In the present situation, three simple functional forms similar to those proposed before in (4.4), (4.5) and (4.6) are suitable to model Q_4 and Q_5 (see next paragraph). There are, however, some differences with subscripts l and j', namely

$$l = 4,5$$

$$j' = \begin{cases} 28 & \text{if} \quad l = 4 \\ 29 & \text{if} \quad l = 5 \end{cases} . \tag{4.19}$$

These three model types are adopted for this section and will be investigated in the subsequent analysis because they fit the characteristics of the problem at hand, for example, they can tackle the non-linear relationships among some predictors and the explained variable. It is also important to notice that a number of empirical studies, for instance those carried out by Chow (1964), Clarke (1994), Abdulla and Lettenmaier (1997), and Ayros (2001) have corroborated their applicability to model this characteristic of the discharge originated in a given drainage basin. Additionally, it should be stated that they all satisfy the guidelines suggested by the concept of simplicity stated before.

Using the short list of observables, the proposed method can be applied in order to assess which model type and which variables are needed to obtain a robust model based on the existing information for the Study Area. A summary of the results obtained are illustrated in Table 4.10.

Table 4.10 Sample of the best models for specific peak discharge in winter (1 = a variable is included in the model, otherwise it is omitted). Values of the optimum estimators (minimum) with $\varphi = 2$ and $\varphi = 1$ are presented, as well as the results for the cross validation and the Akaike's information criterion. The most robust models are highlighted with the symbol ✖. All values are dimensionless since the optimisation has been carried out in the interval $(0,1]$.

Model	x_1	x_3	x_9	x_{11}	x_{12}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{28}	x_{30}	$\varphi = 2$				$\varphi = 1$		Obs.
													Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																			
1401		1		1	1	1	1			1	1		7.989	7.7	2623.7	8.148	74.47	8.256	✖
1881	1	1		1		1	1			1	1		7.995	8.6	2624.6	8.151	74.39	8.259	
1817	1	1				1	1			1	1		8.021	10.3	2626.4	8.146	74.54	8.280	
Multilinear-potential models: MLP1																			
4091	1	1	1	1	1	1	1		1	1	1	1	7.600	11.6	2534.6	7.779	72.73	7.835	✖
4094	1	1	1	1	1	1	1	1	1		1	1	7.600	11.6	2534.6	7.779	72.73	7.841	
4093	1	1	1	1	1	1	1	1		1	1	1	7.600	11.6	2534.6	7.779	72.74	7.842	
Multilinear-potential models: MLP2																			
1308		1				1	1	1			1		7.913	4.3	2609.1	8.032	74.90	8.236	✖
1310		1				1	1	1	1		1		7.906	5.3	2610.2	8.041	74.90	8.240	
1820	1	1				1	1	1			1		7.903	4.8	2609.6	8.036	74.83	8.304	

Table 4.10 shows that the pure potential models (POT) have in general a relative poorer performance if compared with the multi-linear potential ones (MLP1, and MLP2). This finding suggests that not all variables, with the exception of x_{28}, have a strong non-linear relationship with the explained variable Q_4.

Models of type MLP1 in general and model No. 4091 in particular exhibit the lowest values of the cross-validation statistics, the latter for instance got 7.779 and 7.835 for estimators $\varphi = 2$ and $\varphi = 1$ respectively (see Table 4-10); therefore, they are comparatively more robust and thus more reliable than the other model types. They have, however, one disadvantage if compared with models of type MLP2, namely, they have almost two times as many variables as models of type MLP2. According to the concept of simplicity, model No. 1308 is preferable to model No. 4091 because the former has only five predictors and performs almost as good as the latter; in fact, its cross validation statistics are at most about 5.1% greater than those of the model No. 4091.

In order to take the final decision and select a robust model, the test of significance, whose results are displayed in Table 4-11 for the previously selected models, should also be taken into account. These Monte Carlo simulations show that models No. 1401 and No. 4091 have some variables for which the null hypotheses of the significance test cannot be rejected at 5 or 10% level of significance. This means that based on the sample, there seems to be no evidence of a functional dependence among these variables and Q_4. Model No. 1308, on the contrary, has variables significant at even less than 1%. These results confirm that all variables contained in the model are certainly not independent of the explained variable. Hence, the model No. 1308 is selected as a robust model to predict the specific peak in winter.

Table 4.11 Results of the permutation test for models Nos. 1401, 4091, and 1308 using R=500. The tabulated figures are the Monte Carlo p-values as fractions. The estimator has been minimised with $\varphi = 2$.

Model	Type	x_1	x_3	x_9	x_{11}	x_{12}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{28}	x_{30}
1401	POT	-	$\simeq 0$	-	0.046	0.100	$\simeq 0$	$\simeq 0$	-	-	$\simeq 0$	$\simeq 0$	-
4091	MLP1	$\simeq 0$	$\simeq 0$	0.210	0.038	0.030	0.022	$\simeq 0$	-	0.022	$\simeq 0$	$\simeq 0$	0.261
1308	MLP2	-	$\simeq 0$	-	-	-	$\simeq 0$	$\simeq 0$	$\simeq 0$	-	-	$\simeq 0$	-

Model No. 1308, as displayed in Table 4.12, has a very small positive bias (i.e. 8×10^{-4}), which means that this model would tend, although in a very small measure, to overestimate its predictions. This model, nevertheless, does not exhibit the smallest values with regard to other quality measures, but they are very close to the minimum, which in this case corresponds to model No. 4091.

Table 4.12 Quality measures for the most robust models with $\varphi = 2$.

Model	Type	E_1 [mm]	E_2 [mm^2]	E_3 [mm]	E_4 [–]	E_5 [mm]	E_6 [–]	E_7 [–]
1401	POT	0.01	9.08	3.01	0.35	2.34	0.27	0.78
4091	MLP1	0.00	8.38	2.89	0.33	2.25	0.26	0.79
1308	MLP2	0.00	9.00	3.00	0.34	2.33	0.27	0.78

The relative root mean square error of model No. 1308 is about 34.4%. This figure is 2.8 times greater than the corresponding value obtained for the annual specific discharge in winter. This is partly because the PDF of Q_4 is very skewed and has a relatively small average (about 8.8 mm). It could also be due to the uncertainty involved in predicting peak flows. It also implies that this model tends to be more accurate when predicting values greater than the observed mean. Because of these inaccuracies, the correlation coefficient between observed and calculated values using an estimator with $\varphi = 2$ is about 0.78 (only).

It is important to remark that the optimised coefficients (see Table 4.13) for the selected model exhibit inverse relationships for variables, x_{15}, x_{16}, and x_{17}; and direct relationships with the remaining ones. Such relationships make sense from a physical point of view, for instance, the higher the field capacity, the more rainwater is retained in the soil matrix, and hence, the smaller the peak. Conversely, the higher the specific precipitation, the higher the peak to be expected. Furthermore, the larger the forested areas in a basin, the higher the evapotranspiration, and hence, the lower the peak discharge tends to be. This kind of rationale has been extracted from the sample data by the selected model.

Table 4.13 Optimized parameters (with $\varphi = 2$) for model No. 1308 without removing heteroscedasticity.

Model	β_0	β_{17}	β_{18}	β_{J*}	β_3	β_{15}	β_{16}	β_{28}
1308	-0.2173	-0.0361	-	0.1873	0.2249	-0.2893	-0.0070	1.0847

A condition for an unbiased estimator function is that $E[\Phi] = 0$ and the $\mathrm{var}(\Phi) = const.$ (with Φ given by (3.10), Nolsøe et al. (2000). Unfortunately, these very important conditions are sometimes not fulfilled by a chosen model. This is the case with the selected model No. 1308, whose standardized errors exhibit a nonlinear variation of the variance, or in other words, they are heteroscedastic with respect to the predictor x_{28} and the explained variable $\hat{Q}_4$ as it is shown in Figure 4.10.

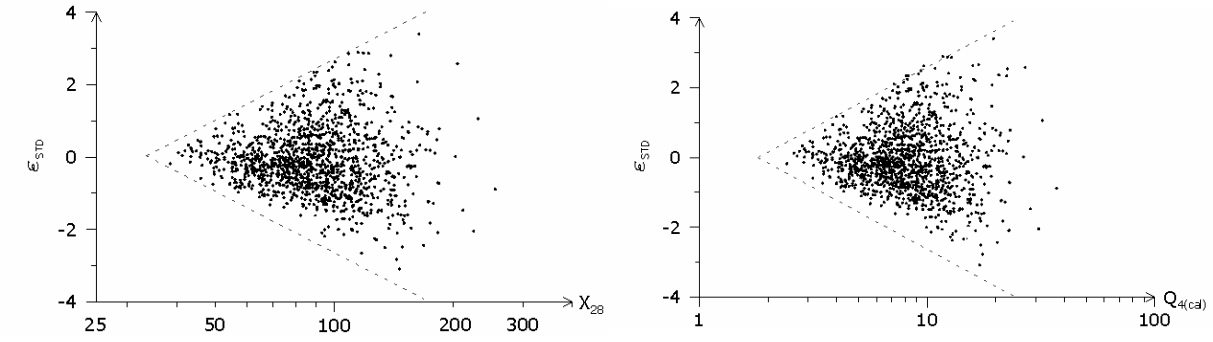


Figure 4.10 Scatterplot of residuals shows a clear heteroscedasticity of the errors with respect to variable x_{28} and the estimated values $\hat{Q}_4$ using model No. 1308.

According to Gentleman (1974), Draper and Smith (1981), Montgomery and Peck (1982), among others, this problem can be addressed by weighting the residuals in the objective function according to their reliability. As Figure 4.10 shows, in the present case the higher the predictor x_{28}, the greater the variance of the residual, and hence the less reliable the observation will be. In such a case, the inverse

of the predictor powered to a given exponent can be used as a robust weighting scheme. Thus, equation (3.10), which is the objective function to be minimised, can be written in general as

$$\Phi = \sum_{t=1}^{T} \sum_{i=1}^{n} w_i^t \left| \varepsilon_i^t \right|^{\varphi} , \tag{4.20}$$

where

$$w_i^t = \begin{cases} \left| x_{ij*}^t \right|^{\varphi_w} & \text{if} \quad \left| \dfrac{\varepsilon_i^t}{s_\varepsilon} \right| \le Z_c \\ 0 & \text{if} \quad \left| \dfrac{\varepsilon_i^t}{s_\varepsilon} \right| > Z_c \end{cases} , \tag{4.21}$$

φ_w an exponent to be calibrated, and

ε_i^t, Z_c, and s_ε variables defined in (3.11), (3.13), and (3.14) respectively.

In this case, the heteroscedasticity of the model No. 1308 with respect to the variable $j^* = 28$ has been greatly attenuated using the exponent $\varphi_w = -2$. The selection of this exponent has been done by trial and error, although other possibilities can be found in the literature (e.g., Draper and Smith 1981).

In order to visualize whether the new estimator stabilizes the variance, a new plot of residuals is needed. In this case, it would be appropriate to examine the pattern of distribution of a pair of variables such as $\{\sqrt{w_i^t}\hat{Q}_4, \sqrt{w_i^t}\varepsilon_i^t\}$ in order to be consistent with the definition of the estimator given by (4.20) and (4.21). The weighted residuals are, of course, standardised. Figure 4.11 (left panel) depicts the distribution of these variables obtained for model No. 1308. The residuals plots in Figure 4.11 reveal that the spread of the error term is roughly the same along the response. In other words, the weighted estimator appears to be effective in this case.

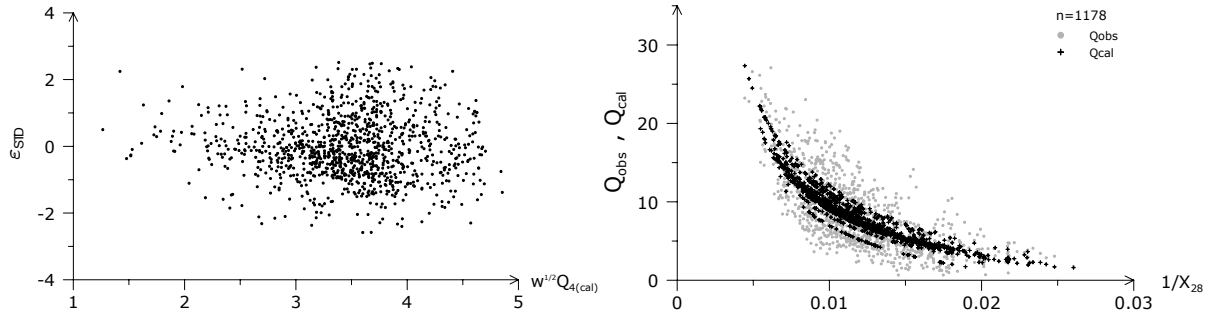


Figure 4.11 The left panel shows a scatterplot of residuals obtained for model No. 1308 using the estimator described by (4.20) and (4.21). The graph at the right panel shows the nonlinear relationship among the calculated/observed specific peak in winter and the inverse of the maximum precipitation index.

The goodness of the fit between the observed Q_4 and the calculated explained variable $\hat{Q}_4$ along the domain of the input variable x_{28} can be visualised in the right panel of Figure 4.11. The inverse of the variable has been employed here with two purposes: 1) to enhance the nonlinear relationship between

the variables, and 2) to stabilize the variance of this explanatory variable so that the plot can contain Q_4 and $\hat{Q}_4$.

The set of parameters that minimise the objective function (4.20) is shown in Table 4.14. The modulus of these parameters is different from those shown in Table 4.13, but their sign is the same. Table 4.14 also shows the optimised coefficients for model No. 1310, which may be interesting to analyse since it is composed of all variables of model No. 1308 plus one that represents the fraction of impervious cover in the floodplains. Although model No. 1310 has not achieved the best performance, it may be interesting to see the effect of this land cover variable upon the specific peak discharges in winter.

Table 4.14 Optimized parameters (with $\varphi = \varphi_w = 2$) for models No. 1308 and No. 1310 after removing heteroscedasticity.

Model	β_0	β_{17}	β_{18}	β_{J^*}	β_3	β_{15}	β_{16}	β_{28}
1308	-4.5505	-0.0149	-	1.9570	0.0814	-0.4167	-0.0040	0.8214
1310	-4.6254	-0.0110	0.0497	1.3327	0.1135	-0.3812	-0.0043	0.8515

It is interesting to see in the previous table that all constants have preserved their signs after the inclusion of variable x_{18}, however, their magnitude is affected in several intensities. The coefficient obtained for variable x_{18} is positive and its module is about 4.5 times greater than that obtained for variable x_{17}. Furthermore, after removing the heteroscedasticity of these models, all variables, with the exception of x_{17}, remain significant at the 5% level as can be seen in table 4.15. The latter is significant at the 10% level.

Table 4.15 Results of the permutation test for models No. 1308 and No. 1310 using R=500. The tabulated figures are the Monte Carlo p-values as fractions. Heteroscedasticity has been removed using the estimator described in (4.20) with $\varphi = \varphi_w = 2$.

Model	Type	x_3	x_{15}	x_{16}	x_{17}	x_{18}	x_{28}
1308	MLP2	0.050	$\simeq 0$	$\simeq 0$	0.022	-	$\simeq 0$
1310	MLP2	0.042	$\simeq 0$	$\simeq 0$	0.098	0.020	$\simeq 0$

The implication of having a positive coefficient for variable x_{18} in model No. 1310 is that if all other terms of this model remain constant, an increment of impervious cover in sensible areas of the catchment, such as the floodplains, would certainly increase the specific peak flow in winter. Conversely, based on models No. 1308 and No. 1310, an increment in forested areas in those places would tend to reduce the specific peak in winter.

4.2.3 Selecting a Robust Model for Summer

Based on the available data, a set of potential predictors of the variable Q_5 is composed of the following variables $\{x_j \quad j = 1,\ldots,19,22,25,29,31,33,37\}$. Due to the reasons already explained, a pre-selection procedure similar to that described in Section 4.1.4 can be used to reduce the number of variables to a maximum 12. This procedure yields, in the present case, the following subset of potential predictors: $\{x_j \quad j = 29,9,12,10,19,18,4,14,15,1,17,31\}$. In this case, the sample data contains 1187 observations distributed during the period 1.11.1960 to 31.10.1993.

The parameter σ used in the definition of variable x_4 is taken equal to 0.3. Regarding those variables that represent the fractions of each land cover type, it was found that x_{17} and x_{19} are more significant if they are evaluated within the domain $\mathcal{L}_i \equiv \Omega_i$, whereas x_{18} gives better results if it is estimated within a buffer zone of the streams that comprise floodplains and riparian wetlands, i.e. $\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$.

Three model types similar to those defined in Section 4.2.2 (4.19) are regarded as suitable for modelling the specific peak discharge in summer. Having the model types and a subset of observables as potential predictors of Q_5 the proposed method can be applied. As a summary of the results, Table 4.16 was compiled from the several thousand possible combinations of predictors and estimators that have been calculated in this case. This table only presents the best three combinations for each model type considering basically their performance using two estimators, namely $\varphi = 2$ and $\varphi = 1$. The weighting function is the same as that shown in (3.13). Initially the distribution of the term ε_i^t in the models described in (4.19) is regarded as homoscedastic.

Table 4.16 Sample of the best models for specific peak discharge in summer (1 = a variable is included in the model, otherwise it is omitted). Values of the optimum estimators (minimum) with $\varphi = 2$ and $\varphi = 1$ are presented, as well as the results for the cross validation and the Akaike's information criterion. The most robust models are highlighted with the symbol ✖. All values are dimensionless since the optimisation has been carried out in the interval $(0,1]$.

Model	x_1	x_5	x_9	x_{10}	x_{12}	x_{14}	x_{15}	x_{17}	x_{18}	x_{19}	x_{29}	x_{31}	$\varphi = 2$				$\varphi = 1$		Obs.
													Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																			
3954	1	1		1	1	1				1	1	1	6.345	11.2	2640.1	6.591	62.32	6.724	✖
3441		1		1	1	1			1		1	1	6.450	28.7	2657.4	6.653	62.53	6.679	✖
4082	1	1	1	1	1	1				1	1	1	6.339	12.1	2640.9	6.614	62.18	6.764	
Multilinear-potential models: MLP1																			
3967	1	1		1	1	1	1	1	1	1	1	1	11.345	13.5	2665.0	11.635	82.79	11.778	
3583		1	1	1	1	1	1	1	1	1	1	1	11.339	13.9	2665.4	11.654	82.53	11.791	
3567		1	1	1	1		1	1	1	1	1	1	11.396	17.8	2669.4	11.689	82.49	11.760	
Multilinear-potential models: MLP2																			
3447		1		1	1	1		1	1	1	1	1	9.435	11.3	2635.7	9.752	63.83	7.029	
3959	1	1		1	1	1		1	1	1	1	1	9.414	10.6	2635.0	9.740	63.69	7.045	
3953	1	1		1	1	1			1		1	1	9.555	24.3	2648.8	9.858	63.86	6.983	

Model No. 3954 is regarded as the most robust model based on the quality indicators shown in Table 4.16. It is, however, necessary to check some additional conditions. The first one is to confirm whether the random error of the model exhibits a uniform distribution with zero mean and a constant variance. The easiest way to do this is by depicting the residuals versus a predictor or the estimated value in a scatterplot in the same way as it was done before. Since the specific peak in winter did exhibit a marked heteroscedasticity with respect to the antecedent precipitation index, it would also be convenient to check whether the standardised residuals in this case have the same behaviour with respect to x_{29}. The results of these tests shown in Figure 4.12 are stunning. The variance of the residuals of model No. 3954 increases non-linearly with an increase of the predictor x_{29}.

Models No. 3447 and No. 3953, which may also be interesting to be analysed because they consider that land cover variables have a linear relationship with the explained variable, also show a marked heteroscedasticity with respect to the variable mentioned above. Hence, before proceeding with the analysis, such an anomaly should be removed (see Figure 4.12 right panel). This irregular behaviour does not occur with the remaining variables of these models.

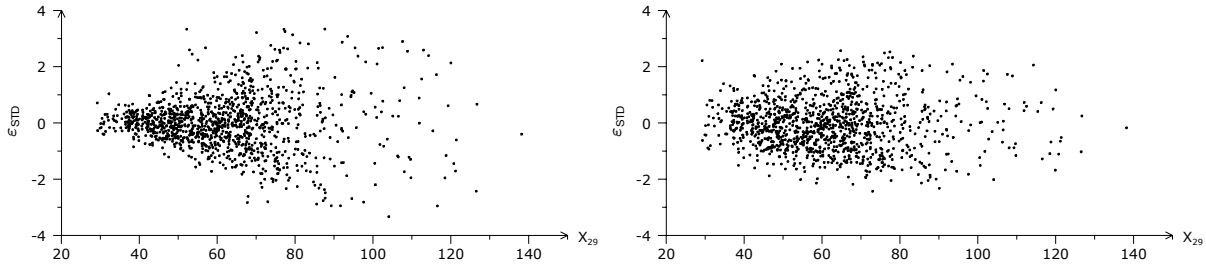


Figure 4.12 Scatterplots of residuals of model No. 3954 before (left panel) and after (right panel) the heteroscedasticity of the errors with respect to variable x_{29} has been removed.

It is also necessary to apply a significance test to corroborate that the variables contained in a given model are not just noise but that they are in some way linked to the explained variable. Such a test will, in turn, help to reduce even further the short list of 'good' models mentioned above. If a model contains non-significant variables, it should be eliminated. The results of the significance test are presented in Table 4.17.

Table 4.17 Results of the permutation test for models No. 3954 and No. 3441 using R=500. The tabulated figures are the Monte Carlo p-values as fractions. Heteroscedasticity has been removed using the estimator described in (4.20) with $\varphi = 2$ and $\varphi_w = 2.5$.

Model	Type	x_1	x_5	x_{10}	x_{12}	x_{14}	x_{18}	x_{19}	x_{29}	x_{31}
3954	POT	0.010	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	-	0.010	$\simeq 0$	$\simeq 0$
3441	POT	-	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	-	$\simeq 0$	$\simeq 0$

From Table 4.17 it can be concluded that all these models have variables that are certainly not independent from the explained variable at the level of significance of 1%, and in some cases, the null hypothesis can even be rejected at smaller levels of significance. Put differently, any of these models is a good choice, but one of them should exhibit relatively better quality indicators. Let us therefore analyse the calculated quality measures of the selected models shown in Table 4.18 in order to see which of them is the most reliable.

The information contained in Table 4.18 indicates that model No. 3954 has performed better than model No. 3441 because all quality measures, with the exception of the bias (E_1), calculated for the former are smaller than that of the latter. Additionally, both models tend to overestimate the observations since their bias is a positive value. The coefficient of correlation of the most robust model (No. 3954) is about 0.82; the RMSE (E_3) of this model is about 7.1 mm and its RRMSE is about 1.1. These relatively high values are the result of the high uncertainty present in the system when the climatic variable x_{29} exhibits higher values. It is worth noting that potential models predicting peak flows in summer have performed much better that the multi-linear potential ones, as

can be seen in Table 4.16. Such behaviour of the system is different from that found for the same runoff characteristic (explained variable) during winter (see Section 4.2.2).

Table 4.18 Quality measures for the selected robust models with $\varphi = 2$ and $\varphi_w = 2.5$.

Model	Type	E_1 [mm]	E_2 [mm^2]	E_3 [mm]	E_4 [–]	E_5 [mm]	E_6 [–]	E_7 [–]
3954	POT	0.01	50.8	7.13	1.07	5.52	0.83	0.82
3441	POT	0.00	51.1	7.15	1.07	5.54	0.83	0.81

Why is this happening? A plausible answer is the following: the linkage between land cover, the atmospheric process (e.g. evapotranspiration, precipitation) and the resulting runoff within a spatial unit during winter can be modelled with a linear sub-model mainly because of the small physiological activity of the vegetal tissue during this season. The opposite occurs in summer because the peak of biomass production is reached during this season. This, in turn, would increase evapotranspiration, and thus, reduce the specific peak flows in a given catchment. Such relationships seem to be non-linear at a mesoscale level as the previous models suggest. This fact can be corroborated with many studies carried out at a microscale; for example, the Penman-Monteith concept (Penman 1948, Monteith 1965) regards evapotranspiration as a non-linear function of many factors, one of which is land cover.

The optimised coefficients for the most robust model found for the specific peak flow in summer are shown in Table 4.19.

Table 4.19 Optimized parameters (with $\varphi = 2$ and $\varphi_w = 2.5$) for model No. 3954 after removing heteroscedasticity.

Model	β_0	β_1	β_5	β_{10}	β_{12}	β_{14}	β_{19}	β_{29}	β_{31}
3954	3003.8	-0.0309	1.1850	0.5410	-0.3101	-3.3029	-0.0694	2.0880	-0.9061

Assuming that there is no high multicollinearity among the different factors employed, the following interpretation of the sign of the variables can be stated. The variable area (x_1) exhibits an inverse relationship with the specific peak discharge; in other words, the bigger the drainage area is, the smaller the peak discharge would be expected. This result agrees with other empirical studies carried out by several authors (e.g. Chow, 1964).

Trimmed mean slope (x_5) has come up as a statistically significant factor with a direct relationship to the explained variable. From the physical point of view, this relationship makes sense since the higher the slope in a given basin is, the faster is the expected flow of water through the hillslopes and stream networks, hence the lesser the concentration time, and consequently the higher the discharge would be. It is interesting to note that the selected robust model is not related with the mean slope of the basin (x_2) but with a trimmed mean that excludes the 30% of the observations at both ends of the PDF of (x_2). This finding is remarkable because it is in those locations of the basin that have mild slopes where a land cover change is most likely to occur as it is depicted in the left panel of Figure 4.13. The right panel of Figure 4.13 shows that land cover change occurs more or less with the same likelihood

between 340 and 680 m above sea level; in this elevation range are located the majority of the urban settlements and major infrastructure within the Study Area.

The fraction of north-facing slopes in a basin (x_{10}) exhibits a direct relationship with the explained variable. This link may be explained from a physical point of view as follows. North-facing slopes in the North Hemisphere get less radiation per square meter than those south-facing ones. This, in turn, implies that in such locations of the basin, less evapotranspiration will be produced, and thus a tendency to get higher runoff may be expected.

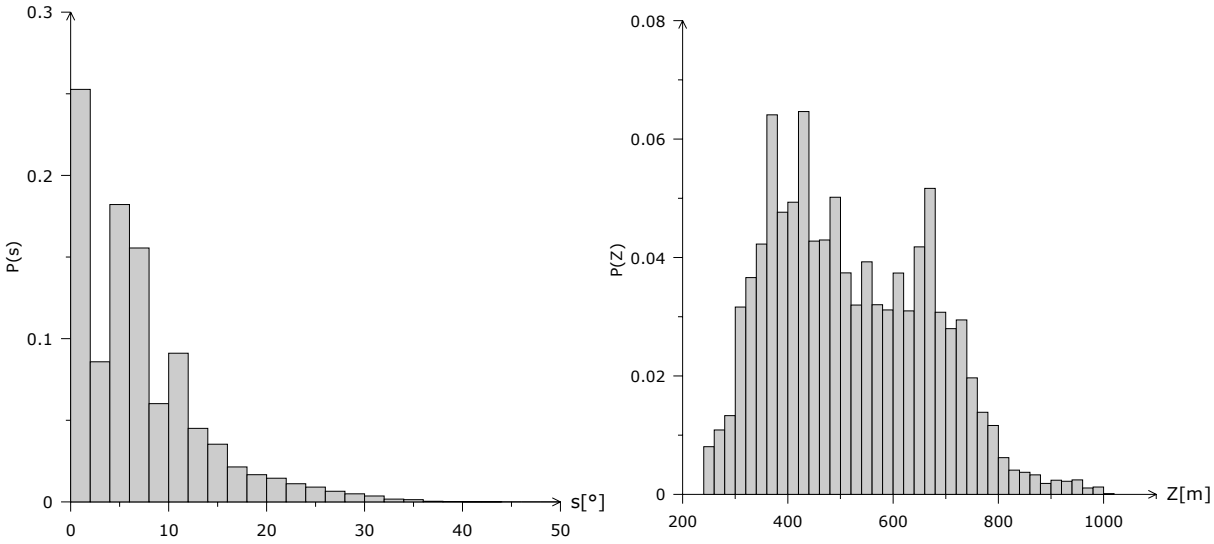


Figure 4.13 PDF showing the likelihood of a given place to endure a land cover change based on its slope and elevation. These curves take into account all locations that have undergone a land cover change from 1960 to 1993.

On the contrary, the share of permeable cover (x_{19}) within a given basin has an inverse relationship with the specific peak flow. This relationship makes sense from a hydrological point of view because the higher the share of such areas within a basin, the higher the infiltration rate to the underground, and therefore, the smaller the runoff tends to be in a given basin. Additionally, taking into account that locations with permeable surfaces would likely have vegetation cover, their overall roughness will be higher, and hence, smaller peaks and longer concentration times can be expected. The vegetal tissue likely present in this land cover category would also tend to diminish the runoff because of the increment in evapotranspiration.

The direct relationship of the precipitation index (x_{29}) is evident. The higher the specific precipitation, the higher the antecedent precipitation index, and hence, the higher the specific runoff. Mean temperature (x_{31}), on the contrary, has an inverse relationship with peak flows. The reason is as follows. The higher the mean temperature in a given basin is, the higher the evapotranspiration, and thus, the smaller the specific peak runoff expected.

The relationship between observed and the calculated values for the selected model are shown in Figure 4.14. It illustrates that the uncertainty of the model widens at higher levels. This phenomenon may have some relationship with the fast and high intensity rainstorms typical in summer whose occurrence, magnitude and consequences has proved to be very difficult to predict.

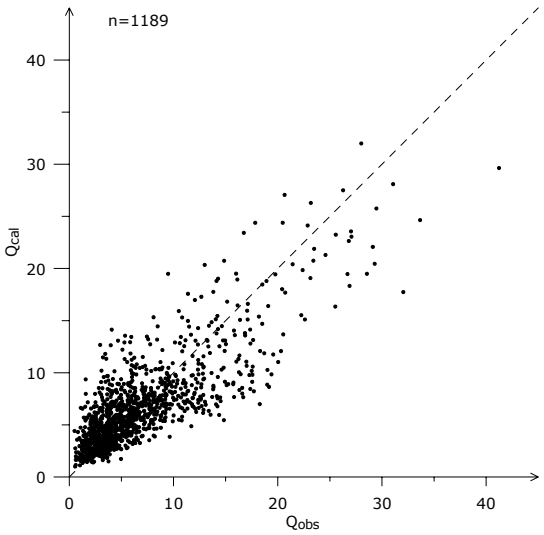


Figure 4.14 This scatterplot shows the relationship between calculated and observed specific peak flows using the potential model No. 3954.

4.3 Specific Volume of the Annual Peak Event

The PDF of the cumulative specific discharge of the annual peak event (Q_6) is positively skewed (1.14) and has a kurtosis of about 2.28. The sample size used to calculate the histogram shown in Figure 4.15 is 1307. Moreover, this variable has a range of about 118.3 mm and a coefficient of variation of about 0.53. The right tail of the PDF shows that rare events with a period of return greater than 800 years have occurred during the reference period. In this case, having such a big sample has given some advantages: 1) it allows determining its empirical distribution more accurately; 2) it reduces the uncertainty with regard to the occurrence of some extraordinary events; and, 3) it increases the reliability of the model because its parameters would have narrower confidence intervals at the same level of significance.

Determining the period of return of extraordinary events as well as investigating whether land cover changes have influenced their frequency of occurrence are crucial tasks in hydrology because they are tightly linked with planning and investment of the key infrastructure of a region. In this stage of the study, however, only the magnitude of this variable will be considered. The frequency of occurrence and its related period of return will be analysed afterwards.

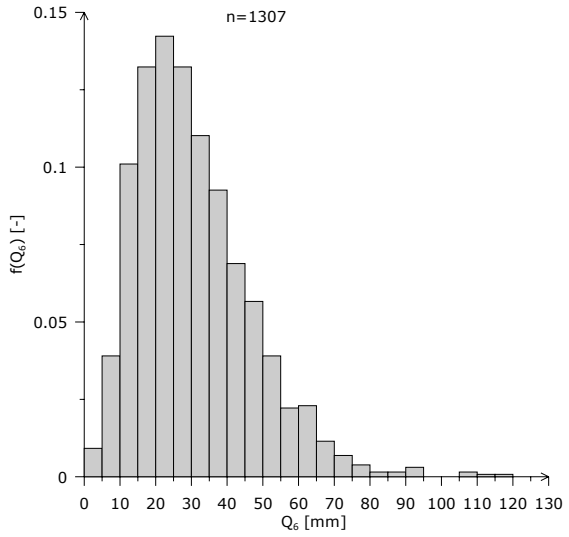


Figure 4.15 Histogram depicting the PDF of the cumulative specific discharge of the annual peak event (Q_6) considering the time series from 1.11.1960 to 31.10.1993 for all spatial units.

The explained variable Q_6, as can be seen in the time series at the bottom of Figure 4.16, exhibits a cyclic behaviour during the period of investigation. In order to visualize possible trends in the data, a 5-year moving average has been applied to this time series and is depicted in the same graph mentioned above. The same procedure has been applied for the explanatory variables x_{26} and x_{27}, whose results are shown in the top and middle graphs of Figure 4.16. Based on this presentation of the data, the following characteristics can be mentioned. Q_6 has a long-term cycle whose lowest value occurs in 1974. From this time until 1993 this variable has had the tendency to increase, albeit potential climatic explanatory variables, such as the annual maximum precipitation index (x_{27}) and the corresponding precipitation index (x_{26}) at the time of occurrence of the peak event, show a slightly negative trend in case of the former and no trend in case of the latter. Nevertheless, the cyclic behaviour of all these random variables is analogous. Consequently, based on this empirical evidence and the principle of causality that governs natural systems (Casti, 1990), one may conclude that there must be reasons that explain such deviations from the mean value. What are they? The next part of this section will be devoted to answer this question.

Based on a similar procedure described before (see Section 4.1.4) and taking into account all potential explanatory variables available, the twelve strongest predictors of Q_6 are $\{x_j \, j = 27, 26, 4, 9, 10, 12, 14, 15, 16, 17, 18, 19\}$. The sample size obtained in this case is 1307 observations, which contain all valid data ranging from 1961 to 1993 at annual basis and for each spatial unit.

In order to obtain higher Pearson correlation coefficients, the three variables representing the share of land cover within a spatial unit have been evaluated as follows: x_{17} and x_{19} have been evaluated within the domain $\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$, whereas x_{18} is within $\mathcal{L}_i \equiv \Omega_i$. In other words, the former are estimated within the buffer zones of the stream network, while the latter is within the whole basin.

The functional relationships to be established between the potential predictors and the explained variable are similar to those represented by (4.4), (4.5) and (4.6). In this case, however, the subscripts take the values $l = 6$ and $j' = 27$. In addition to that it should be said that the model to be found should fulfil the constraints stated in (3.2).

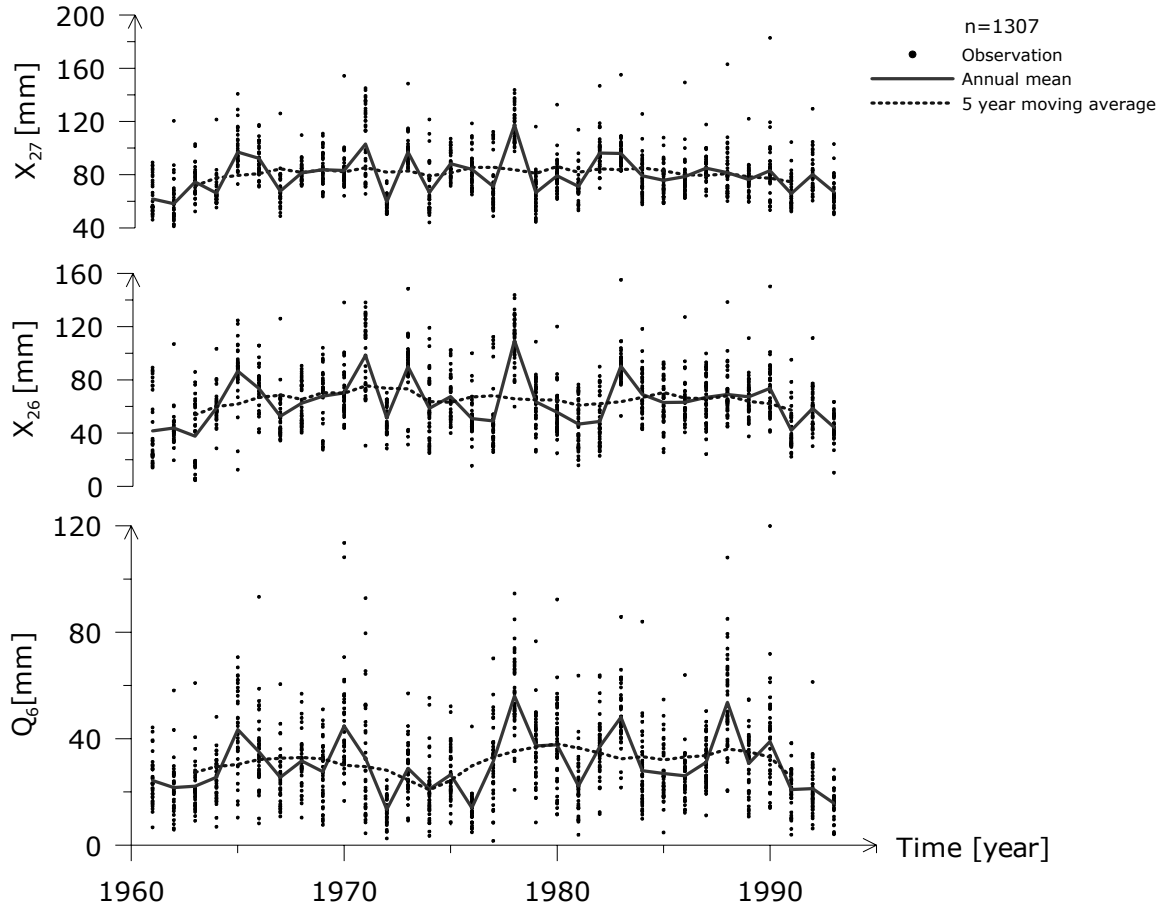


Figure 4.16 Comparison of time series showing the variability of the explained variable (Q_6) and two climatic factors (x_{26}) and (x_{27}). Each observation is represented by a point during the period from 1.11.1960 to 31.10.1993. The annual mean is depicted by a continuous line. The trend of these series is illustrated by a 5-year moving average represented by a continuous dotted line.

As a result of applying the method proposed in Section 4.1.3 a set of the best models has been selected and illustrated in Table 4.20. This table shows that multi-linear potential models of type MLP1 are more suitable and robust than those with functional forms of type MLP2 and POT, because both the estimators and the Jackknife statistics are always the smallest among the subset of the most reliable models. It is noteworthy to state that among the best models, three variables are always present, namely: x_4, x_{26} and x_{27}. This result agrees with the highly correlated relationships among the predictors and the explained variable shown in Figure 4.16. According to the results illustrated in Table 4.20 the most robust model is No. 3662.

Table 4.20 Sample of the best models for cumulative specific discharge of a yearly peak (1 = a variable is included in the model, otherwise it is omitted). Values of the optimum estimators (minimum) with $\varphi = 2$ and $\varphi = 1$ are presented, as well as the results for the cross validation and the Akaike's information criterion. The most robust models are highlighted with the symbol ✖. All values are dimensionless since the optimisation has been carried out in the interval $(0,1]$.

Model	x_4	x_9	x_{10}	x_{12}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{26}	x_{27}	$\varphi = 2$				$\varphi = 1$		Obs.
													Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																			
3825	1		1	1	1	1				1	1	1	8.743	10.9	5560.4	8.917	79.48	9.051	
3835	1		1	1	1	1	1	1		1	1	1	8.719	11.6	5561.1	8.934	79.20	9.090	
3807	1		1	1		1	1	1	1	1	1	1	8.728	12.9	5562.4	8.946	79.23	9.084	
Multilinear-potential models: MLP1																			
3662	1			1			1	1	1		1	1	8.473	9.5	5530.4	8.6018	77.75	8.765	✖
3614	1					1	1	1	1		1	1	8.483	10.9	5531.7	8.6205	77.88	8.757	
3661	1			1			1		1	1	1	1	8.473	9.4	5530.3	8.5972	77.79	8.782	
Multilinear-potential models: MLP2																			
3733	1		1			1			1	1	1	1	8.873	7.5	5585.2	9.0410	79.58	9.114	
3734	1		1			1		1	1		1	1	8.872	7.5	5585.2	9.0400	79.54	9.130	
3717	1		1						1	1	1	1	8.909	10.4	5588.0	9.0638	79.89	9.137	

The error term of the selected model is not homoscedastic as was initially expected. This means that a correction has to be made before the simulation test is applied. The best results have been obtained by introducing a weight that is inversely proportional to x_{27} (i.e. $\varphi_w = 1.0$). The results of the Monte Carlo simulation aimed at determining the level of significance of each variable are shown in Table 4.21.

Table 4.21 Results of the permutation test for model No. 3662 using R=500. The tabulated figures are the Monte Carlo p-values as fractions. Heteroscedasticity has been removed using the estimator described in (4.20) with $\varphi = 2$ and $\varphi_w = 1.0$.

Model	Type	x_4	x_{12}	x_{16}	x_{17}	x_{18}	x_{26}	x_{27}
3662	MLP1	$\simeq 0$	0.004	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$

The results of the simulation shown in Table 4.21 indicate that all variables constituting model No. 3662 are certainly not independent from the explained variable Q_6 at a level of significance even less than 1%. The quality measures estimated for this model are shown in Table 4.22.

Table 4.22 Quality measures for the selected robust model with $\varphi = 2$ and $\varphi_w = 1.0$.

Model	Type	E_1 [mm]	E_2 [mm^2]	E_3 [mm]	E_4 [–]	E_5 [mm]	E_6 [–]	E_7 [–]
3662	MLP1	0.00	223.1	14.9	0.48	11.8	0.38	0.75

As shown in Table 4.22, the selected model has a bias about zero. The differences between RRMSE and RMAE and between RMSE and MAE as well as their magnitude are a good indication of the uncertainty present in the data, which cannot be explained by the model. In fact, it is able to explain 56.1% of the total variance or in other words, it has a coefficient of correlation of about 0.75. Such a

result is satisfactory considering that the model is composed of seven predictors and eight parameters. The optimized coefficients are shown in Table 4.23.

Table 4.23 Optimized parameters (with $\varphi = 2$ and $\varphi_w = 1.0$) for model No. 3662 after removing heteroscedasticity.

Model	β_0	β_4	β_{12}	β_{16}	β_{17}	β_{18}	β_{26}	β_{J*}	β_{27}
3662	188.11	3.7250	0.0105	-0.0834	-0.0852	0.4167	-0.7521	0.0085	1.7998

These coefficients show that variables representing a trimmed slope (x_4), the mean elevation (x_{12}), the share of impervious cover (x_{18}), and the specific precipitation index of a catchment (x_{27}) have a direct relationship with the explained variable. In other words, the higher they are, the bigger the cumulative specific discharge of a yearly peak (Q_6). On the contrary, the remaining predictors have an inverse relationship. It is interesting to note the opposite relationship of those variables representing the share of land cover within a basin. Forest cover will reduce the cumulative volume of a flood event whereas impervious cover will do the opposite.

4.4 Specific Volume and Total Duration of High Flows

According to the correlation matrix shown in Table 4.24, it has been found that the specific volume of high flows (Q_7) is highly correlated with the total duration of high flows (Q_9) in winter, and so are the correspondent variables in summer Q_8 and Q_{10}. Because of that, it would be sufficient to search for explanatory variables for any of them and for both seasons. The variables that will be used in the following analysis are Q_9 and Q_{10}.

These variables, whose positive skewed distributions (skewness of about to 1.3 and 2.7 respectively) are depicted in Figure 4.17, are correlated in various degrees with the following subsets of observables, which can be considered as potential explanatory variables. For instance, in winter the subset is composed of $\{x_j \quad j = 24,30,41,1,4,9,10,12,16,17,18,19\}$, whereas in summer it is composed of $\{x_j \quad j = 25,31,40,1,4,9,10,12,16,17,18,19\}$.

Table 4.24 Correlation matrix $[\mathbf{R}]$ among explained variables Q_7, Q_8, Q_9, and Q_{10}. The sample size is equal to 976.

	Q_7	Q_9	Q_8	Q_{10}
Q_7	1	0.871		
Q_9	0.871	1		
Q_8			1	0.890
Q_{10}			0.890	1

For the evaluation of the land cover variables the following criteria have been used: for winter, variables x_{17} and x_{18} have been evaluated within the buffer zone of the streams ($\mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i$), whereas x_{19} has been evaluated within the whole catchment ($\mathcal{L}_i \equiv \Omega_i$). For summer, x_{17} and x_{19} are calculated within the buffer zones, whereas x_{18} is calculated for the entire spatial unit. By using these criteria, the highest correlation coefficients have been obtained.

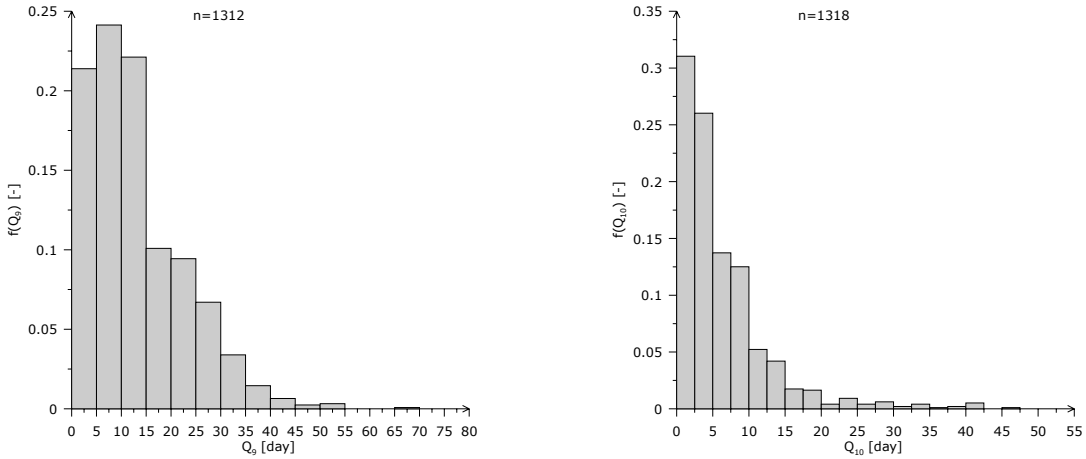


Figure 4.17 Histograms depicting the empiric PDFs for both total duration of high flows in winter (left panel) and summer (right panel) considering all spatial units during the period from 1.11.1960 to 31.10.1993.

Having these subsets of plausible explanatory variables, the proposed method (Section 4.1.3) was applied and the results shown in Table 4.25 have been obtained. Results obtained for winter and summer indicate that the total duration of high flows have a very strong correlation with the macroclimatic situation represented by the variables x_{30} and x_{41} in winter and x_{31} and x_{40} in summer. By a careful inspection of Table 4.25, it can also be noticed that such predictors mostly govern the occurrence of peak flows which equalled or exceeded 5% of the time.

Independent of the functional form employed, the best models for either winter or summer always contain variables x_{40} and x_{41}. Furthermore, the inclusion of almost all variables only reduced the total explained variance by a modest 1.3% in winter and by 1.6% in summer (e.g. models MLP2 in summer).

However, a multi-linear potential model in summer (MLP2 - 3076) having two climatic variables and an additional one representing land cover got the highest ranking because it is the most robust model according to the cross validation statistics. A characteristic of the best models in summer is the absence of morphological variables, or, if they are included, their contribution is negligible. A similar situation occurs with the best model in winter (POT - 3074).

Tests of significance conducted according to the method proposed do not indicate that the variables included in the best models are independent from the explained variable at a 5% level of significance. Results of the simulations are shown in Table 4.26. The quality measures and the optimized parameters are presented in Tables 4.27 and 4.28.

Based on these results it can be stated that the variable total duration of high flows in both winter and summer is mainly governed by the macroclimatic conditions. Morphological variables play an irrelevant role in this case but land cover variables have been found to be statistically dependent and significant although their contribution to the total explained variance is quite small. In other words, this is a case where very small or even "*zero correlation does not imply independence*" (Casti, 1990). On the contrary, independence always implies zero correlation (Deutsch, 2001).

Table 4.25 Sample of the best models for total duration of high flows in winter and summer (1 = a variable is included in the model, otherwise it is omitted). Values of the optimum estimators (minimum) with $\varphi = 2$ and $\varphi = 1$ are presented, as well as the results for the cross validation and the Akaike's information criterion. The most robust models are highlighted with the symbol ✖. All values are dimensionless since the optimisation has been carried out in the interval $(0,1]$.

Winter

Model	x_1	x_4	x_9	x_{10}	x_{12}	x_{16}	x_{17}	x_{18}	x_{19}	x_{24}	x_{30}	x_{41}	$\varphi = 2$				$\varphi = 1$		Obs.
													Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																			
3074							1				1	1	3.091	10.76	2919.1	3.21	38.46	3.34	✖
3974	1	1					1		1	1	1	1	3.055	4.53	2912.9	3.24	38.07	3.37	
4055	1	1	1		1		1	1	1	1	1	1	3.051	9.06	2917.4	3.26	38.03	3.39	
Multilinear-potential models: MLP1																			
3080						1					1	1	3.131	2.36	2936.8	3.21	38.86	3.35	
3769		1		1	1	1		1		1	1	1	3.115	6.26	2940.7	3.29	38.48	3.37	
3656			1			1				1	1	1	3.126	4.19	2938.6	3.28	38.60	3.38	
Multilinear-potential models: MLP2																			
3073								1			1	1	3.160	27.3	2947.9	3.27	38.61	3.38	
3074							1				1	1	3.154	24.8	2945.4	3.26	38.61	3.39	
4055	1	1	1		1		1	1	1	1	1	1	3.081	10.7	2931.3	3.29	38.10	3.44	

Summer

Model	x_1	x_4	x_9	x_{10}	x_{12}	x_{16}	x_{17}	x_{18}	x_{19}	x_{25}	x_{31}	x_{40}	$\varphi = 2$				$\varphi = 1$		Obs.
													Φ	C_{p^*}	AIC	θ	Φ	θ	
Potential models: POT																			
2048												1	2.536	24.2	1625.0	2.57	31.64	2.77	
2565								1	1	1		1	2.489	12.7	1613.7	2.57	30.85	2.78	
4031	1	1		1	1	1	1	1	1	1	1	1	2.447	11.0	1611.9	2.62	30.47	2.93	
Multilinear-potential models: MLP1																			
2564								1		1		1	2.305	5.14	1543.3	2.36	29.61	2.47	
2565								1	1	1		1	2.295	3.14	1541.3	2.36	29.41	2.48	
2563							1		1	1		1	2.297	3.89	1542.0	2.36	29.46	2.48	
Multilinear-potential models: MLP2																			
3076								1			1	1	2.291	14.48	1537.6	2.34	30.02	2.50	✖
2052								1				1	2.328	27.47	1550.4	2.37	30.07	2.54	
4093	1	1	1	1	1	1		1	1	1	1	1	2.254	15.61	1538.7	2.40	29.19	2.61	

In this respect, the proposed method is much more robust than the standard inference tests of independence based on the normal distribution theory, in which zero correlation implies independence. Paraphrasing what has been clearly stated by Blyth (1996) and Shaw (1997), among others, the standard linear correlation methods cannot capture the non-linear dependencies existing between time series of n given variables. As a corollary, it can be stated that if the normality assumption does not hold, as is the case here (e.g. see Figure 4.17), the standard inference theory can lead to deceptive conclusions.

Furthermore, a consequence of what has been found by these simulations can be also stated in a probabilistic context. For instance, the likelihood of their joint occurrence of the total duration of high flows in winter, the mean temperature in January, the occurrence of a certain type of circulation pattern, and the fraction of the buffer zones of streams covered with forest is not equal to the product of the likelihood of each event occurring independently from each other.

Table 4.26 Quality measures for the selected robust models with $\varphi = 2$.

Model	Type	Season	E_1 [day]	E_2 [day^2]	E_3 [day]	E_4 [–]	E_5 [day]	E_6 [–]	E_7 [–]
3074	POT	Winter	-0.09	11.15	3.34	0.25	2.21	0.17	0.94
3076	MLP2	Summer	0.00	5.28	2.30	0.34	1.59	0.24	0.94

Table 4.27 Results of the permutation test for models No. 3074 and No. 3076 for winter and summer respectively. The tabulated figures are the Monte Carlo p-values as fractions using R=500.

Model	Type	Season	x_{17}	x_{18}	x_{30}	x_{31}	x_{40}	x_{41}
3074	POT	Winter	$\simeq 0$	-	$\simeq 0$	-	-	$\simeq 0$
3076	MLP2	Summer	-	$\simeq 0$	-	0.024	$\simeq 0$	-

Table 4.28 Optimized parameters (with $\varphi = 2$) for models No. 3074 and No. 3076 for winter and summer respectively.

Model	Type	Season	β_0	β_{17}	β_{18}	β_{J^*}	β_{30}	β_{31}	β_{40}	β_{41}
3074	POT	Winter	1.4185	0.0589	-	-	-0.1155	-	-	0.9509
3076	MLP2	Summer	0.8890	-	0.1048	3.3653	-	-0.5023	1.1430	-

4.5 Frequency of High Flows

Based on the previous analyses, it has been shown that land cover variables are related to many runoff characteristics at a mesoscale level (e.g. peak flow) during both winter and summer. Besides that, and since those relationships have statistically significant variables, it can be expected that a change of one of them, for instance the share of impervious areas within a basin, will have an impact sooner or later on the maximum peak flow, for example, or on the total annual discharge. In other words, land cover variables have been related with the magnitudes of the observables. However, up to here, nothing has been said about the factors that govern the probability of occurrence of high flows in a given catchment during winter or summer.

In order to address this issue, it has been investigated by means of the maximum likelihood method which theoretical distribution function fits the data best. In this study the available information, i.e. Q_{11} and Q_{12} (which stand for the absolute frequency of high flows during winter and summer respectively) will be used. After several trials, the best fits obtained for the EDF (empirical

distribution function) of these variables (see Figure 4.18) are the Poisson and the Weibull distribution functions, whose probability and density functions are

$$\Pr(Q_{i11}^t = \mathrm{k} \Big| \mu) = \frac{e^{-\mu}\mu^{\mathrm{k}}}{\mathrm{k}!} \qquad \mathrm{k} = 0,1,2,\ldots \tag{4.22}$$

$$f(Q_{i12}^t \Big| \mathrm{a},\mathrm{b}) = \left(\frac{\mathrm{a}}{\mathrm{b}}\right)\left(\frac{Q_{i12}^t}{\mathrm{b}}\right)^{\mathrm{a}-1} e^{-\left(Q_{i12}^t/\mathrm{b}\right)^{\mathrm{a}}} \qquad Q_{i12}^t, \mathrm{a}, \mathrm{b} > 0 \tag{4.23}$$

for both winter and summer correspondingly. The MLEs (maximum likelihood estimates) of the parameters μ, a, and b are $\hat{\mu} = 3.952[1/year]$, $\hat{\mathrm{a}} = 0.820$, and $\hat{\mathrm{b}} = 1.918$ respectively. In case of the summer frequencies, a continuity correction has to be made because a continuous distribution has been used to estimate discrete data. Comparing the EDFs and the fitted ones shown in Figure 4.18, it seems that the theoretical models fit the data reasonably well although some differences exist. For instance, the Poisson distribution tends to under-allocate probability for smaller values of Q_{11}, whereas the opposite occurs for higher ones. In summer appears the opposite if the Weibull distribution is used. In order to assess the goodness of the fits a χ^2 test is indispensable. It shows that the null hypothesis (i.e. that the data were drawn from the fitted distribution) for both the Poisson (winter) and the Weibull (summer) distributions cannot be rejected because their p-values are 0.206 and 0.254 respectively.

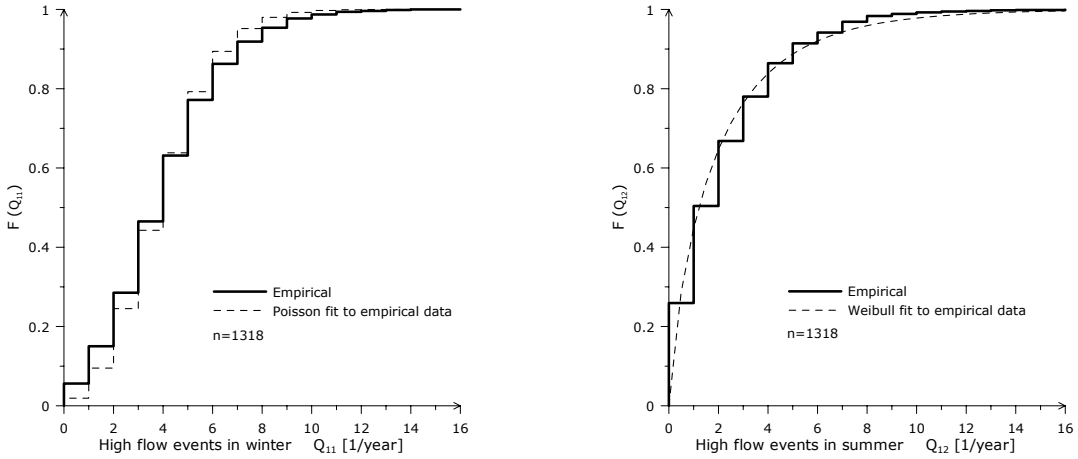


Figure 4.18 Empirical and fitted CDFs for both frequency of high flow events in winter (left panel) and summer (right panel) considering all spatial units during the period from 1.11.1960 to 31.10.1993.

Having done this, the previously mentioned issue can be re-stated based on the GLM (Generalized Linear Models theory) (Gilchrist, 1984; Clark, 1994; Davison and Hinkley, 1997; Lindsey, 1999). It is worth mentioning that this method has been used to estimate probabilities or occurrence frequencies of a given event; e.g. Stahl and Demuth (1999) have used a logit model to fit binary data, and Davisson and Hinkley (1997) have estimated counts of a discrete variable using a log-linear model. The method employed here is based on GLM but with some modifications suitable for the present case. It is as follows.

A generalized linear (or non-linear) model can be used to relate the parameters of the PDF of a given variable Q_{il}^t (the l^{th} characteristic of the runoff process for the i^{th} spatial unit at time t) with a

number of predictors or observables. In other words, μ and b should be related to a number of predictors or observables $\left(x_{i1}^t, x_{i2}^t, \ldots, x_{iJ}^t\right)$.

The structure of a generalized model can be written using three elements:

1. The deterministic element or the *predictor*, which is a suitable function of the explanatory variables x_j; for instance, a multi-linear (ML), a potential (POT), or a multi-linear-potential (MLP) relationships whose explicit equations are

$$\eta_{il}^t = \beta_0 + \sum_j \left(x_{ij}^t\right)^{\beta_j}, \tag{4.24}$$

$$\eta_{il}^t = \beta_0 \prod_j \left(x_{ij}^t\right)^{\beta_j}, \tag{4.25}$$

and

$$\eta_{il}^t = \beta_0 + \sum_{j\in\mathbf{U}} \beta_j x_{ij}^t + \beta_{J^*} \prod_{\substack{j \\ j\notin\mathbf{U}}} \left(x_{ij}^t\right)^{\beta_j} \tag{4.26}$$

respectively.

Where

$$\begin{aligned}
\mathbf{U} &= \left\{x_j \; j = 17,18,19\right\} \\
l &= 11,12 \\
i &= 1,\ldots,46 \\
t &= 1961,\ldots,1993 \\
j &\in \{1,\ldots,J\} \\
J^* &= J+1 \\
J &= 41 \\
\beta_0, \beta_j, \beta_{J^*} &= \text{coefficients to be optimised.}
\end{aligned}$$

2. The distributional element, which indicates that the variance of the response is an explicit function of the mean μ for each observation, i.e. $\mathrm{var}(Q_{il}^t) = \kappa V(u_{il}^t)$.

   Where, $V(\bullet)$ is the *variance function* and κ is the *dispersion parameter*.

   For example, for the Poisson distribution

$$\begin{aligned}
Q_{il}^t \sim \mathrm{Poisson}(\mu_{il}^t) \quad \text{with} \quad E\left[Q_{il}^t\right] &= \mu_{il}^t \quad \forall i,t \\
\mathrm{var}\left[Q_{il}^t\right] &= \kappa\mu_{il}^t \\
\kappa &= 1 \\
l &= 11
\end{aligned}, \tag{4.27}$$

and for the Weibull distribution

$$\begin{aligned} Q_{il}^t \sim \text{Weibull}(\text{a}, \text{b}_{il}^t) \quad \text{with} \quad E\left[Q_{il}^t\right] &= \mu_{il}^t = \text{b}_{il}^t \Gamma\left(1+\text{a}^{-1}\right) \quad \forall i,t \\ \text{var}\left[Q_{il}^t\right] &= \kappa\left(\mu_{il}^t\right)^2 \\ \kappa &= \Gamma\left(1+2\text{a}^{-1}\right) - \Gamma^2\left(1+\text{a}^{-1}\right) \\ l &= 12 \end{aligned}, \tag{4.28}$$

where $\Gamma(\cdot)$ is the gamma function.

3. Finally, the last element of the model is the monotone and differentiable *link function* $g(\cdot)$, which establishes a "link" between the predictor and the mean so that $g(\mu_{il}^t) = \eta_{il}^t$.

In the present case, three link functions are to be tested:

Name	Link Function
Identity	$\mu_{il}^t = \eta_{il}^t$
Logit	$\mu_{il}^t = \dfrac{K}{1+\exp(\eta_{il}^t)} \quad K>0$
Log	$\mu_{il}^t = \exp(\eta_{il}^t)$

In the logit model, K is a case specific constant denoting an asymptotic behaviour of the data.

The estimation of the parameters $\boldsymbol{\beta}$ is to be carried out by maximizing the log-likelihood function $\ell(\cdot)$, whose general form for a variable Q_{il}^t exhibiting a PDF $f(Q_{il}^t \left| \text{a}, \text{b}, \ldots, \mathbf{x}_i^t, \boldsymbol{\beta}\right.)$ given a set of explanatory variables $\left(x_{i1}^t, x_{i2}^t, \ldots, x_{iJ}^t\right)$, and provided that all observations are independent is written as

$$\ell(\boldsymbol{\beta}) = \log \prod_{i,t} f(Q_{il}^t \left| \text{a}, \text{b}, \ldots \mathbf{x}_i^t, \boldsymbol{\beta}\right.). \tag{4.29}$$

Once the three elements of a given model have been defined, the maximum likelihood estimators (MLEs) of its parameters $\boldsymbol{\beta}$ can be found by maximizing

$$\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}), \tag{4.30}$$

and the goodness of the fit can be assessed either by the *deviance*

$$\text{D} = 2\kappa\left\{\ell(Q) - \ell(\hat{\boldsymbol{\beta}})\right\}, \tag{4.31}$$

or by the Akaike's Information Criterion AIC

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2p^*. \tag{4.32}$$

In (4.31) κ can be estimated by

$$\hat{\kappa} = \frac{1}{n_0 - J - 1} \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{\left(Q_{il}^t - \hat{u}_{il}^t\right)^2}{V(\hat{u}_{il}^t)}. \tag{4.33}$$

and $\ell(Q)$ is the log-likelihood of the saturated model which is nothing else than a model where $Q_{il}^t = \hat{u}_{il}^t \ \forall\, i,l,t$. The term p^* in (4.32) is the number of parameters used in a given model that contains j input variables.

For the selection of variables and other relevant quality measures, as well as for the significance tests, the employed method is the same as before, with the only difference that the likelihood $-2\ell(\hat{\boldsymbol{\beta}})$ will be used instead of the objective function Φ (see Section 3.3.2-4, Section 3.3.6-7, and Section 4.1.3).

Tables 4.29 to 4.31 summarized the results obtained by applying the previous methodology to the available data.

Table 4.29 The best models obtained for the frequency of high flows in winter and summer (1 = a variable is included in the model, otherwise it is omitted). The estimated deviance, as well as the results for the cross validation statistic and the Akaike's information criterion, is presented. The most robust models are highlighted with the symbol ✖. All values are dimensionless.

Winter: $Q_{i11}^t \sim \text{Poisson}(\mu_{i11}^t)$

Model No.	x_4	x_9	x_{10}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{21}	x_{32}	x_{41}	Predictor	Link	κ	AIC	θ	Obs.
2653			1		1	1	1		1	1		1	ML	log	0.656	4545.8	3405.8	
2651			1		1	1		1	1	1		1	ML	logit	0.601	4458.5	2934.4	
3933	1		1		1	1	1		1	1	1	1	POT	identity	0.746	4327.3	2626.4	✖

Summer: $Q_{i12}^t \sim \text{Weibull}(\text{a}, \text{b}_{i12}^t)$

Model No.	x_7	x_8	x_{12}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{25}	x_{31}	x_{40}	Predictor	Link	κ	AIC	θ	Obs.
4015	1	1		1		1	1	1	1	1	1	1	ML	log	0.909	1902.0	231813	
2821	1							1	1	1		1	POT	identity	3.299	1367.0	2060.9	
3052	1	1	1	1		1		1		1		1	MLP	identity	2.981	1372.2	1378.4	✖

Table 4.30 Parameter estimates and results of the permutation test (the Monte Carlo p-values with R=500) obtained for the selected models for winter and summer respectively.

Winter (POT model No. 3933)

Parameter	β_0	β_4	β_{10}	β_{15}	β_{16}	β_{17}	β_{19}	β_{21}	β_{32}	β_{41}
Estimates	0.0240	-0.6373	0.5561	0.2248	-0.0046	-0.1310	-0.1663	0.7614	0.0583	0.1602
p-value	-	$\simeq 0$	$\simeq 0$	0.015	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	0.018	$\simeq 0$

Summer (MLP model No. 3052)

Parameter	β_0	β_{18}	β_{J*}	β_7	β_8	β_{12}	β_{14}	β_{16}	β_{25}	β_{40}
Estimates	0.1003	0.0115	4.9230	-0.8247	-1.1360	-0.2890	0.4863	-0.0063	1.0979	0.4540
p-value	-	0.032	-	0.010	0.004	0.002	0.064	0.024	$\simeq 0$	$\simeq 0$

Table 4.31 Additional quality measures for the selected robust models.

Model No.	Type	Season	E_1 [year^{-1}]	E_2 [year^{-2}]	E_3 [year^{-1}]	E_4 [–]	E_5 [year^{-1}]	E_6 [–]	E_7 [–]
3933	POT	Winter	0.00	2.16	1.47	0.36	1.16	0.29	0.77
3052	MLP	Summer	0.16	1.13	1.06	0.49	0.76	0.35	0.87

The selected models (see Tables above) have statistically significant variables, in other words, the null hypothesis (see Section 3.3.7) of independence can be rejected in favour of the alternative hypothesis (i.e. predictors are certainly not independent of the explained variable) at the 5% level of significance, with the exception of variable x_{14} in model No. 3052 in summer. It is worth noting that the selected models have one or more land cover variable(s) as predictor(s).

Based on the model structure and on the evidence contained in the samples, the following remarks can be stated.

For winter, the frequency of occurrence of high flows is, as expected, largely dependent on the meteorological conditions, specially the total precipitation x_{21}; the wetter a given year is, the more likely a flood event would arise. The same direct relationship applies to the maximum temperature in January x_{32} and the composed indicator of wet circulation patterns x_{41}, share of north-facing slopes x_{10}, and average field capacity x_{15}. Inversely related are the trimmed mean slopes x_4 and the shares of forest and permeable areas (such as grasslands) in the buffer zones of the stream network x_{17} and x_{19} respectively.

During summer, the model shows that variables with a direct relationship are the meteorological ones, i.e. mean precipitation x_{25} and the composed index for wet circulation patterns x_{40}, the share of saturated areas x_{14} and the share of impervious areas within a catchment x_{18}. Inversely related appear to be the mean slope near the stream network x_7, drainage density x_8 , mean elevation x_{12}, and the share of karstic formation x_{16} within a given basin.

As stated by equations (4.27) and (4.28) the variance of the i^{th} response at time t is a function of its mean μ_i^t, which is, in turn, a function of a set of predictors $\{x_j\}_i^t$. Figure 4.19 illustrates this fact for the MLP model No. 3052 for summer as an example. Although it is not shown here, the proposed model for winter also exhibits similar features.

The plot in Figure 4.19 also shows the way in which the selected MLP model (No. 3052) for summer has been able to cope with the heteroscedasticity present in the sample ($n_0 = 1196$).

Concerning the frequency of high flows in summer Q_{12} plotted in the ordinates, this Figure depicts also that the expectation of the observed values is quite close to the expectation of the calculated ones at different levels of the predictor x_{40}. Hence, considering these facts, it can be said that the proposed model (which has a Pearson correlation coefficient between the observed and calculated values of about $r = 0.87$) is fitting the observed data quite well, even though some mismatches occur at higher levels of the predictor. These shortcomings of the model can be attributed to the lack of enough observations at those levels.

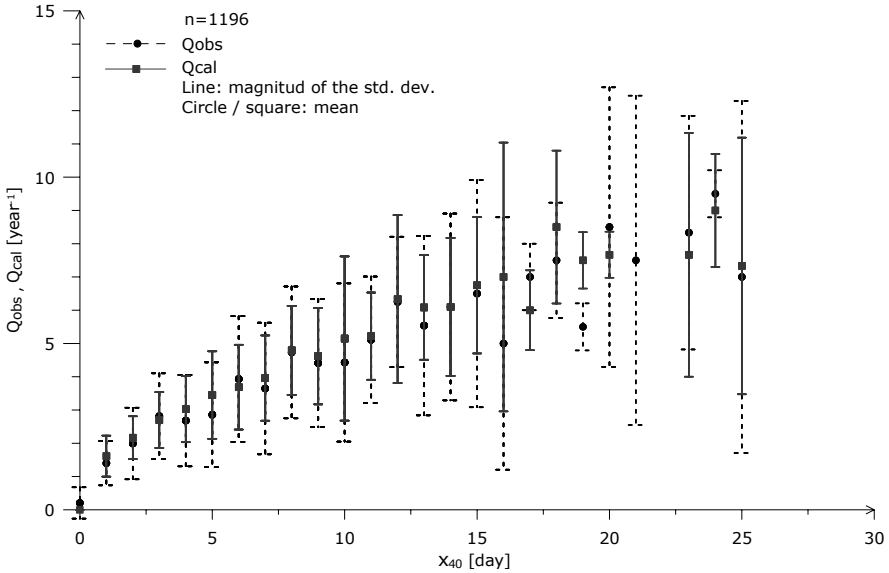


Figure 4.19 Plot showing the variation of the dispersion of the explained variable Q_{12} (observed and calculated by model No. 3052) as a function of the predictor x_{40}. Both continued and dashed lines represent the magnitude of the standard deviation whereas dots and rectangles represent the mean values at each level of predictor.

Chapter 5

Modelling Characteristics of Low-Flows with Time-Dependent Data

5.1 Introduction

An accurate analysis of low-flow regimes occurring in a given stream or river is of crucial importance in watershed management because of the following reasons. Firstly, low-flows will constrain the design of key infrastructure facilities such as water supply and irrigation systems, river navigation networks, and hydroelectric power plants. Secondly, they will indicate to the water-manager the maximum levels of BOD[1] and the maximum allowable concentrations of other pollutants (e.g. Hg, Pb, Zn, P, N, Rn) that should not be reached in a given stream so that its ecosystem will not be jeopardized or damaged during a drought period.

In general, longer low-flow periods will increase investment costs of a given infrastructure facility in a non-linear way. Additionally, an erroneous estimation of such regimes will cause substantial economic losses for a region since the water shortage will hamper production processes.

In order to better understand this phenomenon, it is necessary to determine the most likely period of the year when it may occur. In other words, this means that the temporal distribution of discharge and precipitation should be determined within a given domain (i.e. a basin) for different time intervals during a water year, say months. By knowing these two observables and assuming that the annual change of underground storage is insignificant, the basic form of the water budget for a given spatial unit Ω_i during a given time interval t can be determined as

$$\left\langle P_i^t \right\rangle - \left\langle Q_i^t \right\rangle - \left\langle \mathcal{V}_i^t \right\rangle - \cancelto{\approx 0}{\left\langle \Delta S_i^t \right\rangle} = 0\,, \tag{5.1}$$

where the variables P, Q, $\mathcal{V}$, and ΔS stand for precipitation, discharge, evapotranspiration, and change of underground storage. The operator $\langle\cdot\rangle$ represents the integral of a given variable over the spatial domain Ω_i and/or during the time interval t (e.g. one month). This equation must hold everywhere because it represents the principle of conservation of mass within the system. The results of (5.1) can then be averaged in order to have an unbiased estimator for each variable at a given time

[1] Biochemical Oxygen Demand (BOD) refers to the amount of oxygen that would be consumed if all the organics in one litre of water were oxidized by bacteria and protozoa (ReVelle and ReVelle, 1988).

interval (e.g. for January). The result of such a procedure for the Study Area is depicted in Figure 5.1. This graphical presentation shows that the most probable low-flow spells would take place during summer (M, J, J, A, S, O), in which the evapotranspiration will increase because of higher air temperature; which in turn will reduce the river discharge although the precipitation has increased within its basin.

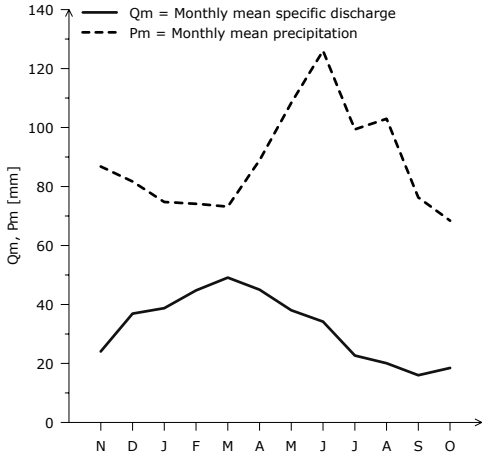


Figure 5.1 Annual water balance of the Study Area. Each value is computed over the period 1961 to 1993.

Because of this fact, the present study will only consider low-flow spells that happen during summer. Having defined the time span for the study of low-flows, the following question can be formulated in connection with the general aim of the present research: are the land cover changes that have taken place within the study area influencing in some way the probabilities of occurrence and/or the total duration of low-flow events?

In order to answer this question, the available information will be first described.

Table 5.1 Correlation matrix $[\mathbf{R}]$ among explained variables Q_{13}, Q_{14}, Q_{15}, and Q_{16} and some climatic explanatory variables in summer. The sample size is equal to 860.

	Q_{13}	Q_{14}	Q_{15}	Q_{16}	x_{25}	x_{31}	x_{37}	x_{38}
Q_{13}	1							
Q_{14}	0.771	1						
Q_{15}	0.291	0.345	1					
Q_{16}	0.633	0.711	0.633	1				
x_{25}	-0.335	-0.587	-0.113	-0.251	1			
x_{31}	0.103	0.246	0.036	0.130	-0.312	1		
x_{37}	0.143	0.252	0.060	0.142	-0.297	0.846	1.	
x_{38}	0.664	0.849	0.297	0.587	-0.569	0.228	0.241	1.

Symmetric

5.2 Description of Time-Dependent Variables

The time series depicted in Figure 5.2 for catchments No. 11 and No. 13 as well as the correlation matrix shown in Table 5.1 point out the same fact: the explained variables $\{Q_l \ \forall\, l = 13,\ldots,16\}$ are mutually correlated.

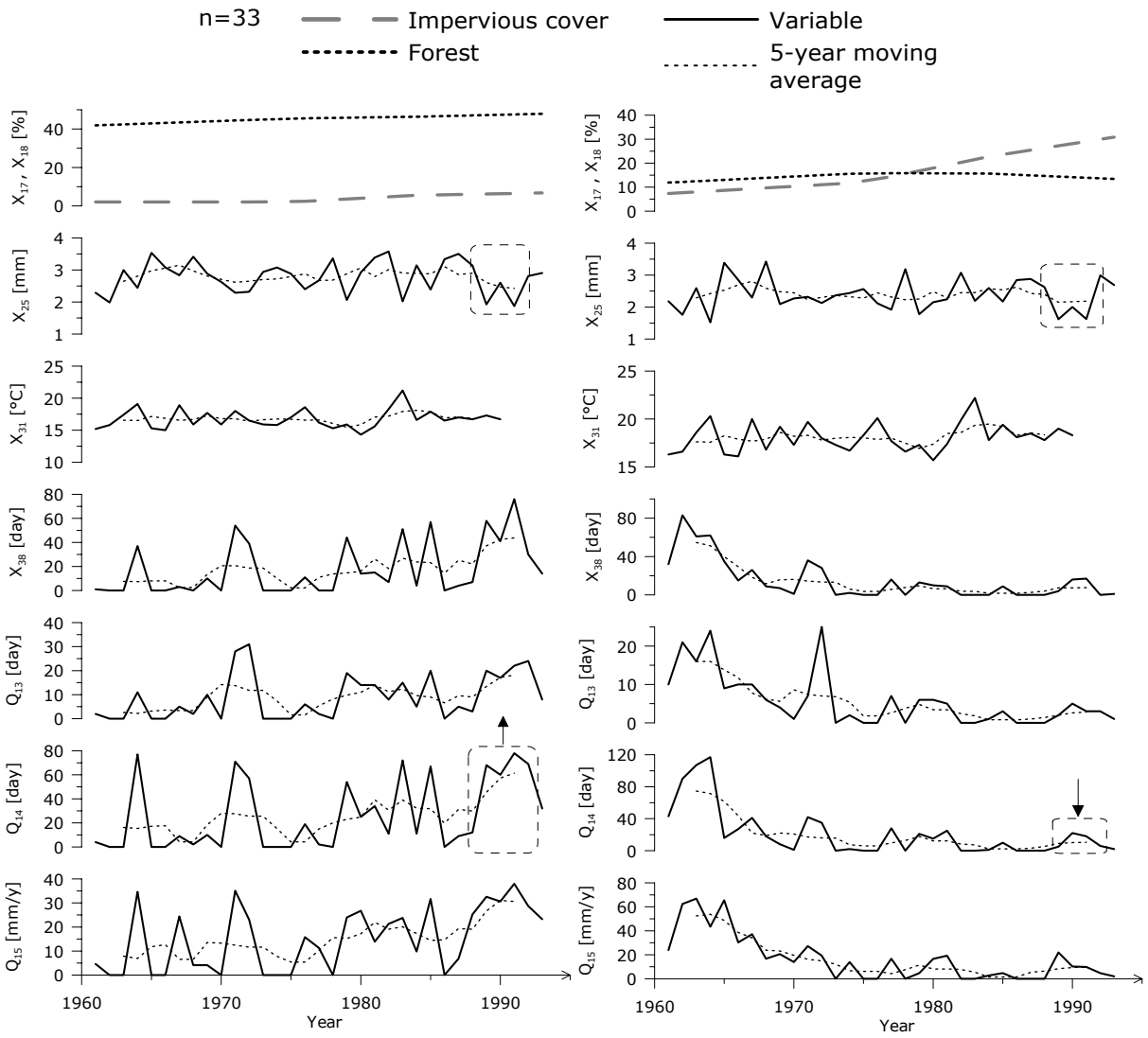


Figure 5.2 Time series showing the trends (by means of a 5-year running average) and the actual values for variables Q_{13}, Q_{14}, and Q_{15} as well as for some explanatory variables including land cover for two catchments of approximately the same size ($A \cong 125\,[\mathrm{km}^2]$). On the left panel is catchment No. 11 with growing shares of forest and impervious cover; whereas on the right panel is catchment No. 13 which has endured a steady land use transition from grassland (permeable land cover) to settlement (impervious land cover) and a steady decline of forest since the mid 70s. The data shown here correspond to the period from 1961 to 1993.

Hence, it would be sufficient to model one of them in order to give an answer to the previous question. An evident selection will be the total drought duration Q_{14} since it exhibits the higher correlations not only with all potential predictors but also with the rest of the explanatory variables.

The following remarks can be stated based on the time series shown in Figure 5.2.

- Firstly, that both the daily mean precipitation (for summer) x_{25} and the daily mean air temperatures in July x_{31} do not reveal any significant trend;
- Secondly, that the composed variable x_{38}, which accounts for days with dry circulation patterns and a decreasing antecedent precipitation index exhibits a completely different behaviour

depending on the shares of land cover while almost constant seasonal-mean precipitation and temperature have been observed;

- Thirdly, that the variable x_{38} is highly correlated with total drought durations (Q_{14}), and;
- Finally, that the land cover variables seem to have played a significant role in the duration of low-flows at mesoscale level, especially the shares of forest and/or impervious areas. As is shown in Figure 5.2, a combination of a rapid growth of impervious cover accompanied by a decline of forest may have led to a rapid shrink of the total drought durations; conversely, slightly growing shares of forest and impervious areas may have led to an increase of total drought durations. In other words, the explained variable Q_{14} has been attenuated by land cover variables. The rectangles with dashed line shown in Figure 5.2 illustrate this fact, i.e. the same climatic phenomenon (a dry year) may produce different outcomes depending on the land cover situation within the catchment, as well as on its morphology.

The distribution of the explained variable Q_{14} is positively skewed (skewness = 1.45) with its mode and median occurring at 0 and 6 [day] respectively. Hence, three positively skewed theoretical distributions from the exponential family, specifically the exponential, the gamma, and the Weibull distributions, were fitted to the observations using the maximum likelihood method. The χ^2 test statistic obtained for each fit was 1.2, 2.5, and 1.05 respectively. Based on this statistical test, whose $p\text{-value} = 0.31$, it is possible to assume that the data can be modelled using a Weibull distribution with a shape factor $\mathrm{a} = 1.035$ and a scale factor $\mathrm{b} = 29.708$. The sample also indicates that the explained variable is heteroscedastic with regard to one of its predictors, namely x_{38}. Additional statistics of the explained variable can be found in the Appendix 4.

5.3 Total Drought Duration

Using the procedures described in Chapter 4 it was found that the most significant variables to explain Q_{14} are $\{x_j \;\; j = 1,7,8,9,10,13,16,17,18,19,31,38\}$. In this case, forest and permeable cover have been evaluated at basin level whereas impervious cover has been evaluated within the floodplains and buffer zones of the stream network, i.e. $\{x_j \;\; \forall j = 17,19 \;\wedge\; \mathcal{L}_i \equiv \Omega_i \;\; \forall i = 1,\ldots,46\}$ and $\{x_j \;\; \forall j = 18 \;\wedge\; \mathcal{L}_i \equiv \mathcal{B}_i \subset \Omega_i \;\; \forall i = 1,\ldots,46\}$.

Using these twelve variables and a sample with 752 observations, all possible combinations of predictors have been calculated using the following model with three variants (predictors). Explicitly it can be written as

$$
\begin{aligned}
&Q_{i14}^t \sim \mathrm{Weibull}(\mathrm{a}, \mathrm{b}_i^t) \quad Q_{i14}^t > 0 \quad \forall i,t \\
&E\left(Q_{i14}^t\right) = \begin{cases} 0 & x_{i38}^t = 0 \\ \mu_i^t & x_{i38}^t > 0 \end{cases} \\
&\mathrm{var}\left(Q_{i14}^t\right) = \begin{cases} 0 & x_{i38}^t = 0 \\ \kappa\left(\mu_i^t\right)^2 & x_{i38}^t > 0 \end{cases}
\end{aligned}
\tag{5.2}
$$

where

$$\begin{aligned}\mu_i^t &= \mathrm{b}_i^t \Gamma\left(1+\mathrm{a}^{-1}\right)\\ \kappa &= \Gamma\left(1+2\mathrm{a}^{-1}\right)-\Gamma^2\left(1+\mathrm{a}^{-1}\right)\end{aligned}. \tag{5.3}$$

Three predictors that are functions of the explanatory variables will be used. They will be called (POT), and multi-linear potential 1 and 2 (MLP1, MLP2) respectively. They can be written as follows

$$\eta_i^t = \beta_0 \prod_j \left(x_{ij}^t\right)^{\beta_j}, \tag{5.4}$$

$$\eta_i^t = \beta_0 + \sum_{j\in \mathbf{G}\cup\mathbf{U}} \beta_j x_{ij}^t + \beta_{J^*} \prod_{j\in\mathbf{M}} \beta_j \left(x_{ij}^t\right)^{\beta_j}, \tag{5.5}$$

and

$$\eta_i^t = \beta_0 + \sum_{j\in\mathbf{U}} \beta_j x_{ij}^t + \beta_{J^*} \prod_{\substack{j\\ j\notin\mathbf{U}}} \left(x_{ij}^t\right)^{\beta_j} \tag{5.6}$$

respectively. In all cases the link function will be the identity one, so that

$$\eta_i^t = \mu_i^t \quad \forall i,t\,. \tag{5.7}$$

Where

$$\begin{aligned}\mathbf{U} &= \left\{x_j \;\; j = 17,18,19\right\}\\ \mathbf{G} &= \left\{x_j \;\; j = 1,7,8,9,10,13,16\right\}\\ \mathbf{M} &= \left\{x_j \;\; j = 31,38\right\}\\ \beta_0,\beta_j,\beta_{J^*} &= \text{coefficients to be optimised.}\end{aligned}$$

Table 5.2 shows the summarised results of applying the proposed method (see Chapters 3 and 4) to the present dataset.

Table 5.2 Robust models for total drought duration in summer (1 = a variable is included in the model, otherwise it is omitted). The estimated deviance as well as results for the cross validation statistic and the Akaike's information criterion is presented. The most robust models are highlighted with the symbol ✖ . All values are dimensionless.

Summer: $Q_{i14}^t \sim \mathrm{Weibull}(\mathrm{a},\mathrm{b}_i^t)$

Model No.	x_1	x_7	x_8	x_9	x_{10}	x_{13}	x_{16}	x_{17}	x_{18}	x_{19}	x_{31}	x_{38}	Predictor	Link	κ	AIC	θ	Obs.
3502		1	1		1		1	1		1	1	1	POT	identity	0.611	6293.9	20.42	
3149				1			1	1	1		1	1	MLP1	identity	0.811	6217.8	24.16	
2964	1	1	1			1		1				1	MLP2	identity	0.794	5958.8	7.18	✖

Table 5.3 Parameter estimates and results of the permutation test (the Monte Carlo p-values with R=500) obtained for the selected model MLP2 No. 2964.

Parameter	β_0	β_{17}	β_{J^*}	β_1	β_7	β_8	β_{13}	β_{38}
Estimates	-0.269	0.071	14.658	-0.075	-0.711	-2.126	0.236	0.869
p-value	-	0.045	-	$\simeq 0$	$\simeq 0$	$\simeq 0$	0.016	$\simeq 0$

Table 5.4 Additional quality measures for the selected robust model.

Model No.	Type	Season	E_1 [day]	E_2 [day^2]	E_3 [day]	E_4 [–]	E_5 [day]	E_6 [–]	E_7 [–]
2964	MLP2	Summer	0.00	175.83	13.26	0.45	9.26	0.31	0.86

During the process of selection of predictors and for a given model type, it has been observed that some variables appear always or very often as elements of the subset of the best models. This fact is illustrated in Table 5.2 where variables x_{17} and x_{38} have been always present. The selected model whose variables are all significant at 5% not only indicates that the total drought duration within a catchment primarily depends on the macroclimatic conditions represented by the variable x_{38}, but also that the morphology of the catchment and the land cover will play an important role. This evidence also provides valuable support to the remarks presented above in Section 5.2. Hence, variables such as mean terrain slope in the buffer zones of streams x_7, drainage density x_8, and share of forest cover x_{17} should be taken into account when watershed management plans are carried out.

The behaviour of the water system concerning the total drought duration appears to have complex and non-linear relationships with the observables or predictors. The following reasons help to corroborate this statement. On the one hand, the variable x_{38} has a nonlinear relationship with the explained variable (see Table 5.2), which not only depends on the macrocirculation patterns but also on the antecedent precipitation index. The latter, which is an indicator of the soil moisture, is, in turn, directly related to the share of forest within a catchment and inversely related to the share of impervious areas. On the other hand, the share of forest appears as a linear predictor of the explained variable, too (see Table 5.2). Such a complex relationship makes the analysis of low flows more complicated to model.

Fortunately, using the proposed method, a model composed of six predictors out of twelve potential ones has been found, i.e. model MLP2 No. 2964. It has not only a correlation coefficient of 0.86 between the observed and calculated total drought duration, but it also exhibits the smallest Jackknife statistic (7.18) compared with other potential robust models (see Table 5.2). In addition to that, the model's output largely supports the presumption that the explained variable has been drawn from a Weibull distribution, as can be seen in the Q-Q plot of Figure 5.3, although deviations are accounted at the right tail of the distribution.

A Q-Q plot is a scatterplot, in which '*each coordinate pair consists of a data value and a corresponding theoretical estimate for that data value derived from the empirical cumulative probability estimate*' (Wilks, 1995). The empirical cumulative probability estimate for the i^{th} smallest data value will be assumed to be equal to $p(x_{(i)}) = i/(n_0 + 1) \approx \Pr\{X \leq x_{(i)}\}$, where n_0 is the sample

size. Hence, the i^{th} coordinate pair of the Q-Q plot in the present case is given by $[x_{(i)}, F^{-1}(p(x_{(i)}))]$, where $F^{-1}()$ represents the fitted inverse Weibull CDF with parameters a and b given above.

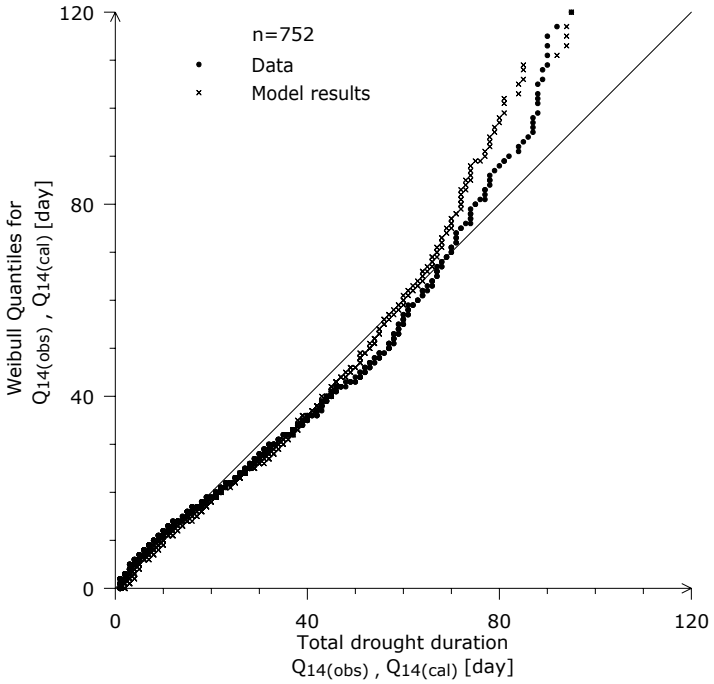


Figure 5.3 Q-Q plot showing the fit of a Weibull (a, b) distribution to both the observed and the calculated total drought duration that have occurred in the case Study Area during the period of 1961 to 1993. The calculated values are the output of model type MLP2 No. 2964. A perfect fit would have all points falling on the 1:1 line.

The Q-Q plot depicted in Figure 5.3 shows two facts. Firstly, it illustrates how the fitted Weibull distribution has been able to reproduce the empirical distribution of the data up to values of a total drought duration of about 88 [day], which corresponds to the 97$^{\text{th}}$ percentile. In other words, the fit works satisfactorily with the exception of the right tail, which exhibits larger differences because the Weibull distribution allocates too much probability to the few observations with values greater than 88 [day], which are too few in the given sample. This is why the Weibull quintiles are above the 1:1 line. Secondly, this plot depicts clearly how closely the selected model has been able to reproduce the empirical distribution function of the data up to about 75 [day].

Chapter 6

An Integrated Approach to Assess the Impacts of Climatic and Land Use/Cover Changes on the Hydrological Cycle at the Mesoscale Level

6.1 Introduction

As was stated before, the present state of land use in a given spatial unit or region is the outcome of two complex and interacting dynamic systems, namely: the anthropogenic activities and 'natural' subsystems as depicted in Figures 1.1 and 1.2. Hence, land use changes cannot be understood completely without an 'insight' into the innumerable relationships among all possible driving forces that may cause a transformation from a land use/cover into another one.

It is, however, unrealistic to pursue a model that attempts to find all possible links (most of them non-linear) between all processes (e.g. weather conditions, soil type distribution, hydrological regime) and all actors involved (e.g. individuals, firms, government acting according to a legal framework) because of the tremendous size of such a model, which, in turn, would make the analysis so complicated and inefficient that the whole modelling effort would become worthless. In physics, for instance, this sort of determinism has been abandoned a long time ago, especially for the analysis of macroscopic multi-particle systems. A good example is the Metropolis algorithm (Metropolis et al. 1953), which simulates the evolution of a system in a heat bath towards thermal equilibrium. Moreover, in the present case, this amount of information will never be collected because of both economic reasons and data protection laws.

It is possible, nevertheless, to simplify the system's complexity to some extent. For instance, Allen (1978-1997) and Peréz-Trejo (1996) introduced a spatial dynamic modelling framework, which describes the average behaviour of an individual or a firm by a system of interacting differential equations that govern the structural changes of the system with regard to population and employment growth (in various sectors), as well as their spatial distribution. In this case, the system's self-organization is assured by the existence of bifurcation points (a critical point at which the system will branch into completely different paths or possible future states), which in a way preserve the adaptability and creativity of the system according to Allen (1997).

This approach has, however, some shortcomings. Firstly, the fact that most dynamical systems have a strange attractor in some region of the parameter values describing the system (Casti, 1990). Such an

attractor is characterized by instability in all motions, deterministic randomness, and sensitivity to initial conditions. These characteristics imply that a small nudge in one of its variables will take the system to a completely different course. Secondly, the mathematical structure of the system is deterministic; hence, the inherent uncertainty of various subsystems has not been taken into account.

The intrinsic randomness of the system is evident, for instance, when one tries to model the weather conditions in a given area or the housing location choice of an individual. Because of this important characteristic of the system, only one fact is certain: a perfect prediction of a future state based on the present state is impossible.

Based on the previous considerations the following questions may be stated. How can a dynamic system be formalized in order to avoid the aforementioned difficulties and take into account its intrinsic randomness? Moreover, which kind of answers should be expected from this type of model? The consensus found in the reviewed literature points out that this sort of system can be modelled using a mathematical construct named a stochastic process. The technique used for solving such systems is called a stochastic or a Monte Carlo simulation (Hammersley and Handscomb 1964, Ripley 1987, Haldorsen and Damsieth 1990).

This technique employs batches of artificial data (i.e. realizations) generated for every random variable of the model resampled from their corresponding probability distributions. The numerous solutions of the system would allow determining the PDFs of the output variables, from which decisions can be taken in a probabilistic way.

These ideas and their application to assess impacts of land use/cover change on the hydrological cycle in a given basin will be illustrated with the following stochastic simulation model.

6.2 Model Structure

The model presented here consists of four modules (see Figure 6.1):

1. The land use/cover change model,
2. The scenario definition,
3. The stochastic simulation, and
4. The statistical inference.

It should be noted that the model structure is general and, hence, it can be applied anywhere. However, the calibration of the land use/cover change model, as well as, modules two and four depend on local conditions that have to be analysed for each particular case. Therefore, they will be discussed extensively in Section 6.3 which deals with the model implementation in the Special Study Area.

6.2.1 A Simple Land Use/Cover-Change Model

The land use/cover change (LUCC) model described below has been chosen because it has a simple structure and can be implemented easily. Furthermore, it assumes that there exists a one to one relationship between each land use and land cover class employed. Despite its simplicity, this model still captures key components that characterize the complexity of the real phenomenon as will be

shown in the next section. This model is based on works carried out by Bell (1974), Turner (1987), Flamm and Turner (1994), and, Muller and Middleton (1994). This model can be further improved by considering the spatial variability of the land cover transitions (e.g. by using semivariograms) as has been proposed recently by Brown (2002). In this respect, more research is still needed.

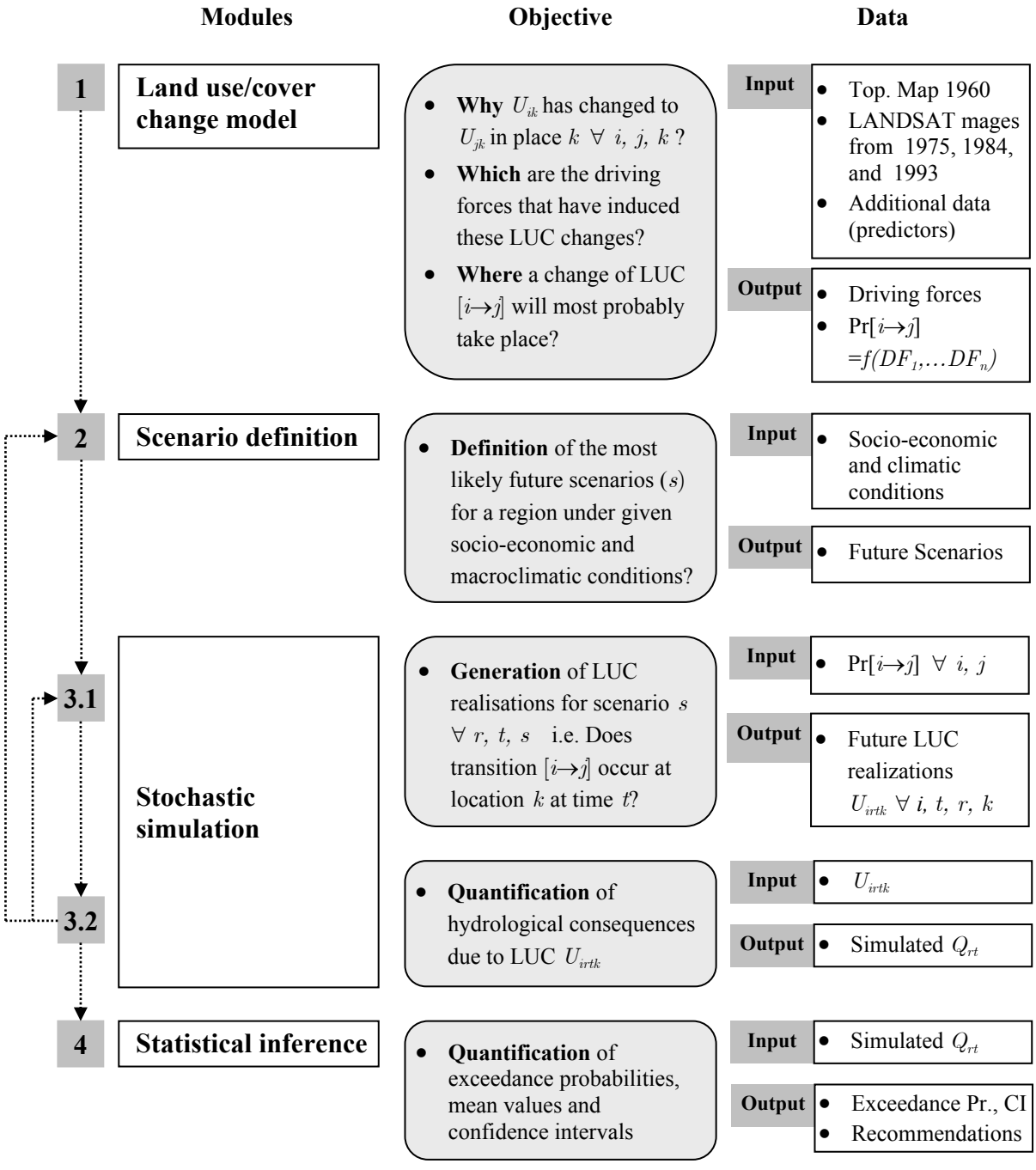


Figure 6.1 Model structure showing the main objectives, required inputs, and outputs for each module.

Let the pair $(\mathbf{Z}, \boldsymbol{F})$ be a stochastic process resembling the land use/cover transformations to be endured by the system during a time span T. Let $\mathbf{Z} = \{(i,j) : 1 \le i,j \le N\})$ denote the $N \times N$ integer lattice covering a given spatial unit Ω ; then $\mathbf{U}^t = \{U_{ij}^t\}, (i,j) \in \mathbf{Z}$ denotes the land use/cover of spatial unit Ω at time t, where $t = 1,\ldots,T$. Let $\mathbf{S} = \{s_q : q = 0,1,\ldots,u\} = \{0,1,\ldots,u\}$ be a finite state space

denoting $u+1$ mutually exclusive land use/cover classes so that $U_{ij}^{t} \in \mathbf{S}\ \forall\, i,j,t$. Finally, let $\boldsymbol{\mathcal{F}} = \{\boldsymbol{\mathcal{F}}_{ij}\}, (i,j) \in \mathbf{Z}$ be a neighbourhood system where $\boldsymbol{\mathcal{F}}_{ij} \subseteq \mathbf{Z}$ denotes the neighbours of (i,j).

Then, the system can be defined as a Markov random field[1] over $(\mathbf{Z}, \boldsymbol{\mathcal{F}})$ if for every (i,j) and every s_q (see Geman and Geman, 1984)

$$\begin{aligned}\Pr(U_{ij}^{t} = s_{q'} \,\big|\, U_{kl}^{t-1} = s_{q_{kl}}^{t-1},\ \forall (k,l) \in \mathbf{Z},\ t = 1,\ldots,t-1) \\ = \Pr\left(U_{ij}^{t} = s_{q'} \,\middle|\, U_{kl}^{t-1} = s_{q}^{t-1}, \left((k,l) \in \boldsymbol{\mathcal{F}}_{ij} \vee (k,l) = (i,j)\right)\right), \\ = (\pi_{qq'})_{ij}^{t}.\end{aligned} \tag{6.1}$$

where $(\pi_{qq'})_{ij}^{t}$ is the probability that the outcome of the t-th transition in cell (i,j) will be $s_{q'}$, given that the outcome of the $t-1$-th transition was s_q, and $s_q, s_{q'} \in \mathbf{S}$. In other words, the system has no memory; the selection of the new state (t) for a given cell (i,j) depends only on the current state $(t-1)$ of this cell and its neighbours and not on prior states. Since historical records support this condition, the system is fully determined by the transition-probability matrix $\boldsymbol{\Pi}(\pi_{qq'})_{ij}^{t}$ given by

$$(\pi_{qq'})_{ij}^{t} = \begin{cases} 0 & q = 0 \vee q' = 0 \\ p_{ij}^{t}(q,q') & \forall q \neq q'\ \ q,q' \geq 1 \\ 1 - \sum_{\substack{l=0 \\ l \neq q}}^{u} p_{ij}^{t}(q,l) & q = q' \end{cases}. \tag{6.2}$$

Turner (1987), Brown (2000, 2002), among others, have pointed out that the transition probability $p_{ij}^{t}(q,q')$ depends on local and time specific conditions. Several empirical studies also have shown that the transition probability is related with socio-economic factors, land use policies, and morphological characteristics of the terrain (Bell, 1974; Flamm and Turner, 1994; Berry, 1995; Brown et al. 2000, 2002).

In the present case, this probability will be determined by (based on Berry et al. 1995)

$$p_{ij}^{t}(q,q') = w_0(q,q')\, w_P(q,q')\, w_{ij}^{t}(q,q') \frac{\exp\left(\beta_0(q,q') + \sum_{k=1}^{K} \beta_k(q,q') x_k(i,j)\right)}{1 + \sum_{\substack{l=1 \\ l \neq q}}^{u} \exp\left(\beta_0(q,l) + \sum_{k=1}^{K} \beta_k(q,l) x_k(i,j)\right)} \tag{6.3}$$

$$\forall q \neq q'\ \ q,q' \geq 1\ ,$$

where

$w_0(q,q')$ = Calibration and scaling parameters to be determined with past information.

$w_P(q,q')$ = Control parameters denoting both the political willingness and the society's level of awareness with regard to environmental impacts and sustainability. The

---

[1] A Markov random field (MRF) is a stochastic process regarded as a generalization of the usual Markov chain (Cross and Jain, 1983). A Markov chain is a sequence of trials, where the outcome of each trial depends only on the outcome of the previous one (Feller, 1950 quoted by van Laarhoven and Aarts, 1992).

set of parameters will be scenario specific and will be of key importance during the simulation. In general, they are values greater or equal to zero. Zero means that a transition is not possible and the greater the value, the greater the willingness to promote such transformation.

$w_{ij}^{t}(q,q')$ = Location and time specific factor indicating the likelihood that a given cell will be transformed to another land use/cover type based on the neighbourhood conditions.

K = Number of exogenous variables regarded as driving forces behind a land use/cover change.

l,k,i,j = Indexes.

$x_k(i,j)$ = Time independent driving force k, with $(i,j) \in \mathbf{Z}$.

$\beta_0(q,q')$ = Intercept for a LUCC from $q \rightarrow q'$.

$\beta_k(q,q')$ = Coefficient estimate for driving force k related with a LUCC from $q \rightarrow q'$.

This probability assumes that the driving forces will be constant or quasi-constant during the simulation time, and it takes into account the fact that landscape changes do not occur randomly in space but in patches or clusters (Brown, 2002). In other words, if a given cell is surrounded by cells belonging to a distinct land use/cover class it is more likely that a land use/cover change occurs here rather than in another one that is surrounded by the same land use/cover class.

The variable $w_{ij}^{t}(q,q')$ has been estimated as

$$w_{ij}^{t}(q,q') = \frac{\left|\left\{(i,j) : U_{ij}^{t-1} = s_{q'} \wedge (i,j) \in \boldsymbol{\mathcal{F}}_{ij}\right\}\right|}{n_c + 1}, \tag{6.4}$$

where $|\{\cdot\}|$ represent the cardinality of the set composed of all neighbours of the cell (i,j) having a land use/cover type q' at the $t-1$-th transition. n_c denotes the number of neighbours of a given cell and c an integer denoting the neighbourhood configuration. In the present case $c = 2$, which means that a given cell has eight neighbours (i.e. $n_c = 8$). The neighbourhood is determined as in Geman and Geman (1984)

$$\boldsymbol{\mathcal{F}}_{ij} = \left\{(k,l) \in \mathbf{Z} : 0 < (k-i)^2 + (l-j)^2 \leq c\right\}. \tag{6.5}$$

6.2.2 Stochastic Simulation

The purpose of the stochastic simulation is to determine how severely a change in land use/cover would affect the hydrological system of a given basin Ω provided specific scenario conditions. The impacts on the hydrological system will be quantified by those empiric models calibrated in Chapters 4 and 5. The variables used by these models will be obtained as follows. The land use/cover variables are obtained as realizations of the LUCC model proposed before; the morphological variables are invariants for the period of the simulation; and the climatic variables will be drawn from their multivariate joint distribution. The resampling procedure, however, has to be done sequentially since the climatic variables are mutually dependent. The procedure is as follows. Firstly, a variable assumed

to be independent (i.e. either x_{24}, x_{25}, or x_{27}, for winter, summer and annual respectively) has to be drawn from their respective EDF. For a subsequent variable, however, the distribution from which it has to be drawn will be modified by the value of the primary variable. This modified distribution is the conditional distribution as defined by the Bayes theorem. The conditional distribution for a secondary variable x_k can be formally written as

$$F_{X_k|X_i,\forall i\neq k}(x_k\,|x_i,\forall i\neq k)=\Pr(X_k\leq x_k\,|X_i=x_i,\forall i\neq k)\,. \tag{6.6}$$

The proposed simulation is carried out by the subsequent algorithm.

Algorithm 6

1. For $r=1,\ldots,R$, where R denotes the total number of realizations.
2. For $t=1,\ldots,T$, where T denotes the total number of years in each realization r.
3. For all $(i,j)\in\mathbf{Z}$.
   a. Estimate $p_{ij}^{t}(q,q')$ as in (6.3).
   b. Generate a random number $\varpi\sim\mathrm{unif}[0,1)$.
   c. If $\sum_{l=1}^{q'-1}p_{ij}^{t}(q,l)<\varpi\leq\sum_{l=1}^{q'}p_{ij}^{t}(q,l)$ accept transition from $q\rightarrow q'$.
4. Estimate land use/cover shares $\mathbf{U}_{\Omega r}^{t^*}$ for the spatial unit Ω.
5. Resample (with replacement) the independent variables $x_i,\ i=24,25,27$ from their respective EDFs.
6. Resample (with replacement) the remaining secondary climatic variables $\mathbf{M}_{\Omega r}^{t^*}$ from their respective conditional distribution functions (6.6).
7. Scale up all climatic variables according to the scenario conditions. Check additional constraints. In case they are not fulfilled return to step 6.
8. Estimate $Q_{\Omega kr}^{t^*}=f\left(\mathbf{G}_{\Omega}^{t},\mathbf{U}_{\Omega r}^{t^*},\mathbf{M}_{\Omega r}^{t^*},\hat{\boldsymbol{\beta}}_k\right)\ \forall k=(2,3,4,5,6,9,10,11,12,14)$. With $\hat{\boldsymbol{\beta}}_k$ and $f(\bullet)$ according to Chapters 4 and 5.
9. Repeat step 2. T times.
10. Estimate the long-term mean for each realization $\overline{Q}_{kr}^{*}=E\left[Q_{\Omega kr}^{t^*}\right]\ \forall k,r\ \ t=1,\ldots,T$.
11. Repeat step 1. R times.
12. Estimate means and variances for each runoff characteristic at each time interval t, $\overline{Q}_{kt}^{*}=E\left[Q_{\Omega kr}^{t^*}\right]\ \ \forall t,k\ \ r=1,\ldots,R$, and

    $$\mathrm{var}(Q_{kt}^{*})=E\left[\left(Q_{\Omega kr}^{t^*}-\overline{Q}_{kt}^{*}\right)^2\right]\ \ \forall t,k\ \ r=1,\ldots,R\,.$$

13. Estimate $Q_k^{*}=E\left[Q_{kr}^{*t}\right]\ \forall k,\ t=1,\ldots,T\ \ r=1,\ldots,R$.
14. For each k, estimate from the simulated-EDF for the long term means $\left\{(Q_{kr}^{*}):r=1,\ldots,R\right\}$ the exceedance probabilities α_k with respect to historical records as

    $$\alpha_k=1-\frac{r}{R+1}\,,\quad Q_{k(r-1)}^{*}\leq E\left[Q_{\Omega k}\right]<Q_{k(r)}^{*}$$

15. For each k, estimate 95% confidence intervals based on $\left\{(Q_{kr}^{*}):r=1,\ldots,R\right\}$.

6.3 Model Implementation

6.3.1 Special Study Area

The proposed simulation model will be applied in the spatial unit No. 13, which is located upstream of the station Denkendorf-Sägewerk (E3 526 300, N5 397 100) in the river Körsch. Its area is about 126.3 km² and because of its vicinity to Stuttgart it has endured a rapid land use and cover change in the past four decades. Figure 6.2 shows the location of the Special Study Area as well as the main transportation network and main settlements in the region.

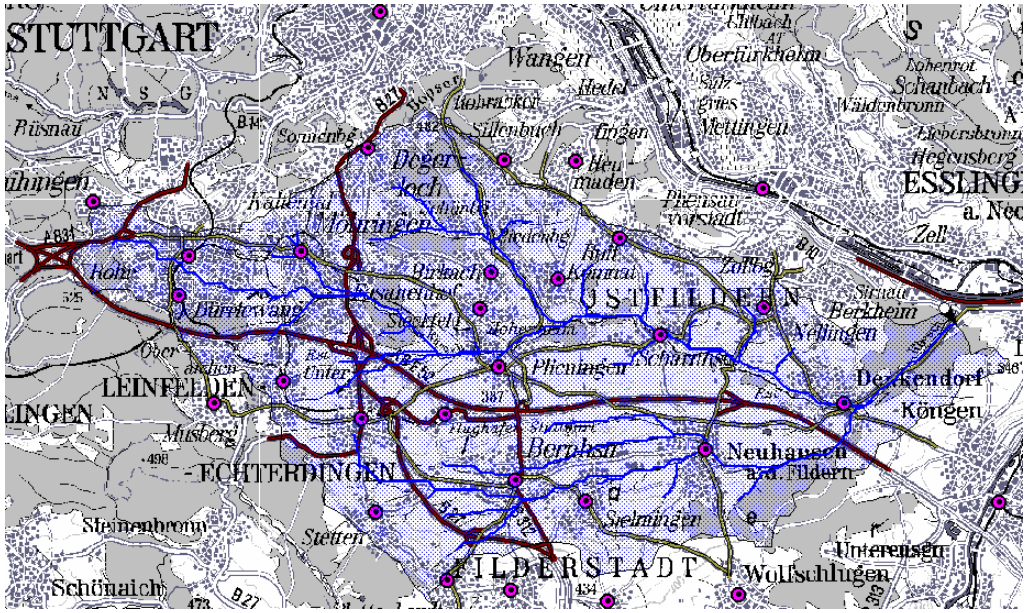


Figure 6.2 Special Study Area for the land use and cover change simulation model.

Stuttgart is the capital of the State of Baden-Württemberg as well as the state's Central Business District, and the main cultural and industrial hub of the Greater Stuttgart Region, which is composed of the following counties (*Landkreis*) Böblingen, Esslingen, Göppingen, Ludwigsburg, and Rems-Murr, as well as the independent municipality (*Stadtkreis*) Stuttgart. This region is considered a densely populated area (BBR, 2000), with a gross density in 2001 of about 717 inh/km² (SLA).

Stuttgart municipality provides an oversupply of jobs since its activity rate (i.e. $Total\,Employment : Total\,Population$) is about $3:5$, whereas the region has an average of about $1.8:5$ (in 2001, SLA). This large difference in the employment distribution, as well as the variety of services offered in the central city, constitute the main driving forces for daily commuting to Stuttgart.

6.3.2 Calibration and Validation of the LUCC Model

The LUCC model proposed before will use four (i.e. $u = 3$) mutually exclusive land use/cover classes, namely

Table 6.1 Land use/cover categories.

q	**Description**
0	restricted or unused land
1	forest
2	impervious cover
3	permeable area

The procedure to calibrate the model comprises the following steps. The first step is the definition of the potential predictors. According to (6.3), the probability $p_{ij}^{t}(q,q')$ has to be explained by exogenous variables called driving forces. In the present case, six potential predictors have been conceived as proximate sources of a land use/cover change. They denote accessibility to main transportation axes, jobs, amenities (located in towns and settlements) as well as morphological variables. In this model, it is assumed that such variables remain unchanged during the simulation period. A summary of these variables is shown Table 6.2. All variables have been defined as lattices $x_k(i,j) : (i,j) \in \mathbf{Z}$.

Having done this, the available land use/cover images acquired in 1975 and 1993 respectively (see Section 2.6.1) were used to determine all sites where land use/cover transitions ($q \rightarrow q'$, $\forall q,q' > 0$) have taken place during this period. As a result, it was found that the number of cells with transition $(2,1)$ is negligible (i.e. 0.08%). This implies that the probability of a transition from impervious cover to forest can be taken as approximately equal to zero. Moreover, it is assumed that it will remain constant during the simulation period for all cells in the Special Study Area.

Table 6.2 Potential predictors of land use/cover change.

k	**variable**	**Description**	**Units**	**Source**
1	$x_1(i,j)$	Distance to main highways	[m]	Digitized 1:50 000 topographic maps
2	$x_2(i,j)$	Distance to towns and settlements with metro or railway connection	[m]	Digitized 1:50 000 topographic maps
3	$x_3(i,j)$	Distance to streams	[m]	From DEM, 30×30m
4	$x_4(i,j)$	Elevation	[m]	DEM, 30×30m
5	$x_5(i,j)$	Slope	[°]	From DEM, 30×30m
6	$x_6(i,j)$	Aspect relative to south	[°]	From DEM, 30×30m

Then, five independent random samples, one for each possible land use/cover transition (q,q'), were obtained according to the following criteria. Each sample should include a binary indicator variable $y_{qq'}(i,j)$ and the corresponding values of the exogenous variables $x_k(i,j)$, i.e. each sample is composed of the following information $\left\{\left(y_{qq'}(i,j), x_1(i,j),\ldots,x_6(i,j)\right) : (i,j) \in \mathbf{Z}\right\}_{qq'}$. Additionally,

each sample should have an equal number of observations for each category of the binary indicator, and a total of $\mathrm{nobs} = 2000$. The binary indicator variable denotes the probability of occurrence of a land use/cover change. If it has occurred it takes the value 1, if not it takes the value 0. More formally

$$y_{qq'}(i,j) = \begin{cases} 1 & \text{if } U_{ij}^{t1} = q' \wedge U_{ij}^{t0} = q \wedge q \neq q' \\ 0 & \text{if } U_{ij}^{t1} = U_{ij}^{t0} = q \end{cases}, \tag{6.7}$$

where $t0 = 1975$, and $t1 = 1993$.

The calibration of the parameters needed for (6.3) will be carried out for each transition probability independently. The explained variable is the binary indicator whereas the explanatory variables are the driving forces x_k. A model for a transition probability (q, q') assumes that the observations of the binary indicator provided by the corresponding sample are realizations of a Bernoulli distribution. The expectation of this variable is therefore

$$E\left[y_{qq'}(i,j)\right] = \Pr(q,q') = \frac{e^{\eta(x_k)}}{1+e^{\eta(x_k)}}, \tag{6.8}$$

where $\eta(x_k)$ is the linear predictor (see Chapter 4) defined as

$$\eta(x_k) = \beta_0(q,q') + \sum_{k=1}^{K} \beta_k(q,q') x_k(i,j). \tag{6.9}$$

Upon this basis, the best models where obtained by applying the method described in Chapter 3 to select the best model given K predictors. The coefficients were fitted by the maximum likelihood method (Chapter 4). The results for the most robust models are shown in Table 6.3. All variables are significant at the 5% level.

Table 6.3 Fitted model coefficients for each transition probability.

Land use/cover transition		$\hat{\beta}_0(q,q')$	$\hat{\beta}_1(q,q')$	$\hat{\beta}_2(q,q')$	$\hat{\beta}_4(q,q')$	$\hat{\beta}_5(q,q')$
From (q)	**To** (q')					
1	2	5.966E -01	7.030E-03	-9.173E-02	-1.259E-03	-1.768E-03
1	3	5.561E+00	-9.179E-03	-8.639E-02	-5.745E-04	-7.529E-04
2	3	3.027E+00	-1.018E-02	-8.011E-02	-8.123E-05	9.876E-04
3	1	-3.168E+00	4.267E-03	1.798E-01	-2.508E-04	5.638E-04
3	2	-3.678E+00	1.227E-02	3.160E-02	-9.757E-04	-5.455E-04

In order to validate the model, a land use/cover map from 1984 (see Section 2.6.1) has been used as a starting condition. Then, using the parameters shown in Table 6.3 and corresponding scaling parameters, the model was run for an interval of 9 years with an $R = 100$. As a result, one hundred realizations for the land use/cover state in 1993 were obtained and compared with the observed land cover map from 1993 using the standard error matrix. On average, the realizations have shown that the model has an overall accuracy of 85%.

6.4 Development Scenarios for the Special Study Area

The proposed LUCC model as well as the hydrological models found before (see Chapters 4 and 5) will be coupled during the simulations under the framework conditions of a given scenario for the Special Study Area (see Figure 6.2).

Scenarios are "*neither predictions nor forecasts of future conditions. Rather they describe alternative plausible futures that conform to sets of circumstances or constraints within which they occur*" (Hammond, 1996). The purpose of scenarios "*is to illuminate uncertainty, as they help in determining the possible ramifications of an issue along one or more plausible (but indeterminate) paths"* (Fisher, 1996). Scenarios to be conceived for this study will have a dynamic character because they "*not only look into consistent future situations, but* [also] *include the consideration of feasible development paths*" (Treuner, 1995). This character of a given scenario will be accomplished in this study by using a stochastic simulation, which will deliver a number of "images of possible futures" given a common starting situation. According to Treuner (1995), scenarios must be envisaged and elaborated taking into account three fundamental issues: 1) future social values; 2) interpretation of a region's external conditions; and 3) assumptions (explicit or implicit) as to the mechanisms of causes and effects of changing patterns.

It should be noted that the third point has been already carefully analysed in the context of the present study. For instance, cause-effect relationships have been found between many runoff characteristics and the shares of the land cover, morphological, and climatic variables for a given basin. Besides that, the land cover state of a given basin at a point in time has been related with exogenous variables governing land use/cover change. The fundamental hypothesis in this case is that these models fitted with past information will still be valid in the future. The remaining two issues are to be discussed below.

6.4.1 Socio-economic Scenarios

In order to simplify the analysis and taking into account the actual socio-economic and political situation in Germany, only two future paths with regard to socio-economic factors and attitudes have been conceived for the Special Study Area. They have been termed as Scenarios S1 and S2. These scenarios have some common features to ease comparison. For instance, the population in the region will slightly decline at about 0.1% per year (according to an external forecast for the administrative units covering the Special Study Area i.e. Stadtkreis Stuttgart and Landkreis Esslingen, SLA, 2002). Furthermore, the GDP per capita of Baden-Württemberg will grow at an average rate of about 2.3% per year (SLA). However, these scenarios will have characteristic conditions with regard to the driving forces and the society attitudes that promote land use/cover changes, namely:

Scenario S1

The keyword for this scenario is status quo. The storyline of this scenario describes a future state of the Special Study Area in which its development can be explained as an extrapolation of past trends. This scenario assumes that the steady growth of income per capita combined with an excellent provision of road transportation network and stable taxation for fossil fuels (0.65 €/l) will keep the relationship between car-ownership and the demand for residential floor space tightly correlated

(r^2=0.88 from 1974 to 1997, as can be seen in Figure 2.2). These indicators will continue to grow at 2.6% and 1.5% per year respectively. In addition to that, the rent for housing in Stuttgart and its surroundings will soar due to the region's high level of centrality.

The implication of these assumptions is that although the population settled in the region is quasi constant, the demand for larger apartments and detached houses with large gardens located in villages and settlements with good road accessibility will grow rapidly. As a result, new housing areas will appear everywhere in the outskirts of Stuttgart, accompanied by large shopping malls with huge parking places whilst floor space downtown will be swiftly taken by branches of the service sector.

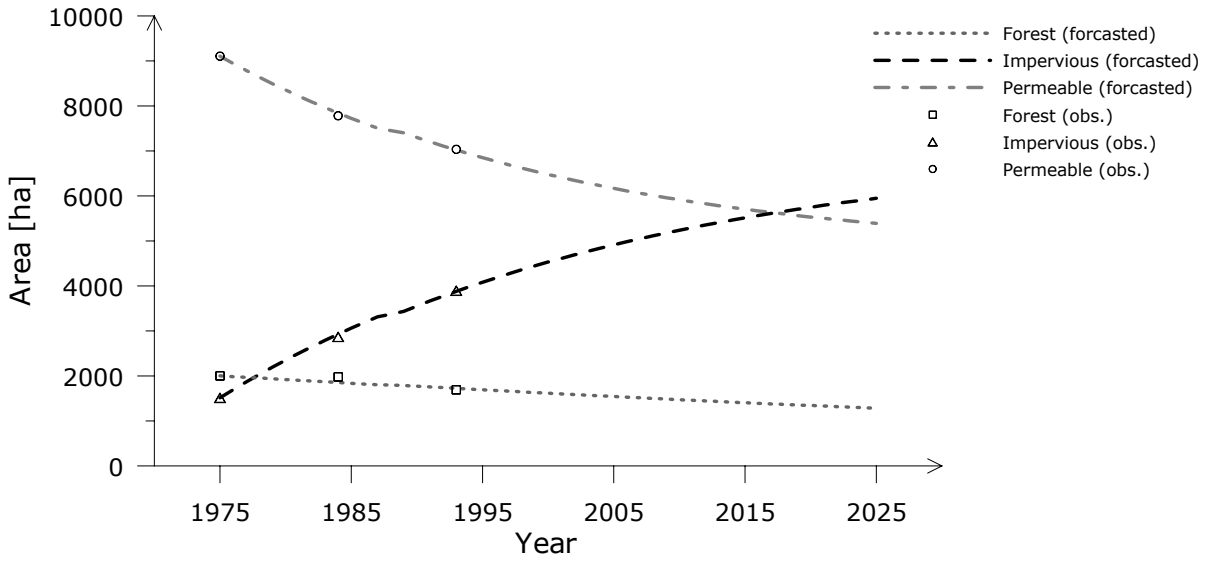


Figure 6.3 Land use/cover forecast based on Scenario S1 conditions for Special Study Area as a whole (total area 126.3 km²). The observation points are derived from the LANDSAT scenes for 1975, 1984 and 1993. The forecasted period is up to 2025.

The consequences of these developments for the overall balance of the land use/cover in the Special Study Area can be seen in Figure 6.3. The forecast has been done based on a Markov chain whose transition matrix adjusted to fit the observations is

$$\mathbf{U}^{t+1} = \mathbf{U}^{t^{\mathrm{T}}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.984 & 0.016 & 0 \\ 0 & 0 & 0.986 & 0.014 \\ 0 & 0.002 & 0.018 & 0.980 \end{bmatrix}, \tag{6.10}$$

where $\mathbf{U}^t$ is a 4-dimention vector denoting the area for each land use/cover category for the whole Special Study Area in time t. Applying this procedure, only the last three land use/cover categories ($q = 1,\ldots,3$) will endure transformations, for instance, forest will decrease slightly, impervious areas will grow rapidly, and permeable areas will decrease continuously. Restricted areas are preserved. This scenario describes a fast urban sprawl in the Special Study Area. Using this forecast the LUCC model will be scaled up so that the land use/cover categories forest, impervious, and permeable cover will reach in average 1280, 5950, 5390 ha respectively by the end of 2025.

Scenario S2

The keyword for this Scenario is local sustainability. The storyline of this scenario differs from the previous one in several topics. Firstly, the public opinion, in general, and the political decision-making bodies, in particular, will finally become aware that a rapid urban sprawl represents a threat to the environment, which, in turn, may contribute to increased flooding and drought hazards in the region. Consequently, tougher land use by-laws and higher property tax regulations will be adopted. As a result, the demand for floor space per capita will be reduced significantly.

Secondly, the "Eco Tax" (tax on fuel that makes commuting more expensive) will be strengthened. Tax exceptions will be introduced for smaller and pollution-free cars, whereas higher taxes will be imposed on vehicles with standard combustion engines. These regulations, along with a sufficient frequency and capacity offered by almost pollution-free mass transportation systems, will slow down the growth rate of the car-ownership ratio. As a result, the demand for space required for new roads and parking places will be reduced dramatically. Because of the new legislation, the growth rate of impervious areas, as can be seen in Figure 6.4, will slowdown from 1.3% per year of the "status quo" Scenario to 0.4% per year in Scenario S2. The scenario denotes a consolidation of the urban fabric of the Special Study Area.

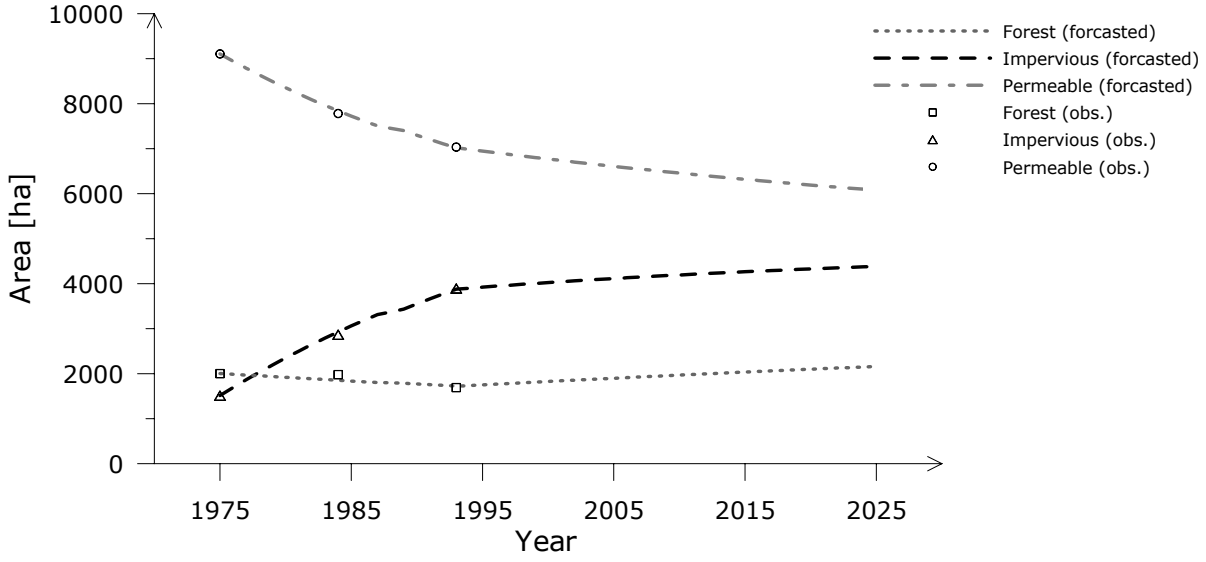


Figure 6.4 Land use/cover forecast based on Scenario S2 conditions for the Special Study Area as a whole.

Thirdly, the decrease of forest observed in the period 1975 to 1993 has been taken by the public opinion as a loss of German "identity". Therefore, land use/cover compensation rules stated in the EIA (Environmental Impact Assessment) by-laws will be strengthened, and wherever possible reforestation projects will be initiated. At the end of the simulation period (i.e. 2025) the land use/cover categories forest, impervious, and permeable cover will have an average of 2160, 4390, and 6075 ha respectively.

6.4.2 Macroclimatic Scenarios

Why are macroclimatic scenarios needed during these simulations? Before this question is answered, another question must be asked: Is climate really changing? The answer is unequivocally yes (Karl and Trenberth 1999, Houghton et al. 2001, Zwiers 2002). Currently, there is plenty of empirical evidence that the Earth's surface mean temperature has endured a very rapid increase during the last 100 years that "counters a millennial-scale cooling trend, which is consistent with long-term astronomical forcing" (Mann et al. 1999) (i.e. the gravitational driving force "which is thought to have driven long-term temperatures downward since the mid-Holocene at a rate within the range of -0.01° to -0.04°C/century" [see Berger,1988 in Mann et al. 1999]). As an example, Figure 6.5 depicts the reconstruction of the temperature anomalies[2] for the past millennium in the Northern Hemisphere carried out by Mann et al. (1999). Based on proxy data (i.e. paleoclimatic) and instrumental records from many studies (Hansen and Lebedeff 1988, Jones, 1994, Vinnikov et al. 1990, Mann et al. 1999) the IPCC (Houghton et al. 2001) has concluded that the average surface temperature in the Northern Hemisphere has increased by 0.6 ± 0.2°C during the 20th century.

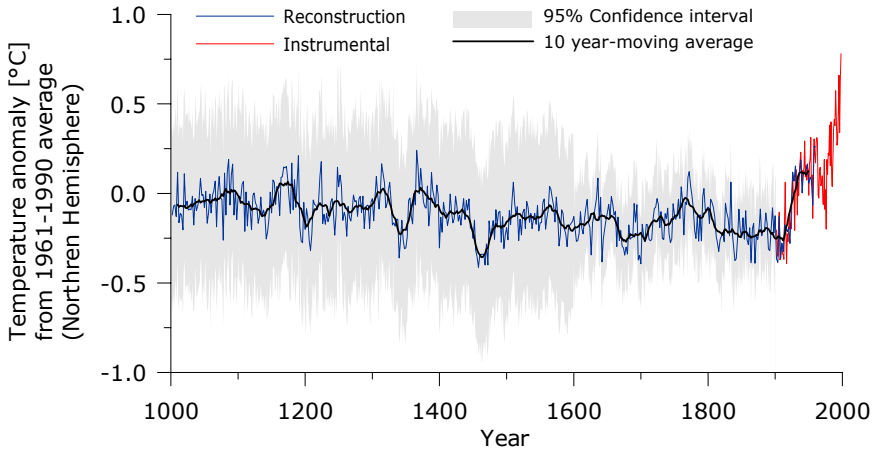


Figure 6.5 Reconstruction of the Northern Hemisphere average temperature anomaly for the past millennium according to Mann et al. 1999(2). Data from the period (1000 to 1902) is reconstructed from tree rings, ice cores, varved sediments, and corals [Mann et al. (1) 1999]. Data from period (1902-1998) is from instrumental measurements. The grey region represents the 95% confidence range. The moving average shows the decreasing trend up to 1900.

Since climate is changing, the weather and its meteorological indicators used in this study at mesoscale will certainly change in the future. However, to estimate how big these changes would be in a given place using General Circulation Models (GCM) is rather complicated because of the extremely high uncertainty involved in future estimates. The uncertainty of the system does not come only from the complexity of the system[3] itself but also from future actions of human beings, especially with

---

[2] The air temperature anomaly is defined as the difference between the temperature in a given year and the average from period (1961-1990), which is roughly 15°C for the Northern Hemisphere (IPCC, 2000).

[3] The intrinsic uncertainty of the state-of-the-art GCM models is caused by the complexity of the iterations among the components of the climatic system, i.e. the atmosphere, the hydrosphere, and the biosphere. At the moment, even using the best supercomputers available, the system of equations can be solved for a spatial resolution of about H: 250 km, V: 1 km. Hence, results of GCM cannot be used directly for climatic inferences at local level (Karl and Trenberth 1999). Estimates at local level are then obtained by statistical downscaling techniques (von Storch et al. 1999).

regard to both the amount of emissions of greenhouse gases into the atmosphere and the magnitude of land use/cover changes. Greenhouse gases (e.g. CH_4, N_2O, CO_2, SO_2) are directly linked with climatic disruptions of the last century. Figure 6.6 shows for example the strong correlation between the temperature anomalies of the Northern Hemisphere and the atmospheric concentration of CO_2 (r = 0.72).

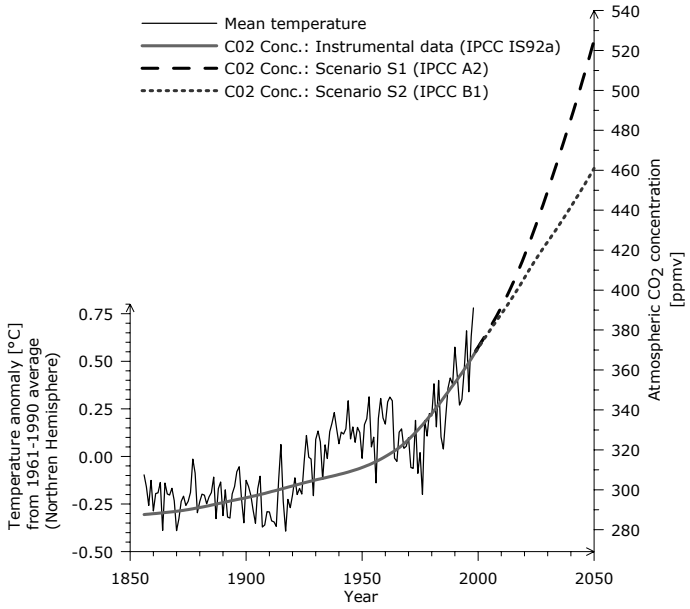


Figure 6.6 Relationship between the atmospheric CO_2 concentrations and the temperature anomalies in the Northern Hemisphere up to 1998. Additionally, this Figure depicts the emission conditions adopted in this study for scenario S1 and S2, which correspond to the IPCC emissions scenarios A2 and B1 respectively. [Data: temperature anomalies from Mann et al. 1999; CO_2 concentrations from the standard IPCC CO_2 concentration history dataset (Enting et al. 1994); IPCC scenario concentrations obtained from IPCC Data Centre].

Based on these facts, the answer to the first question is now straightforward. Macroclimatic scenarios are needed in order to deal with the uncertainty of climate in the future. They will provide the framework for the climatic conditions for a future world under given hypothesised socio-economic and emission scenarios. In order to simplify the analysis, this study only conceives two extreme macroclimatic scenarios called Scenario C1 and C2 respectively.

Scenario C1

The keyword for this scenario is pessimistic, and describes the worst-case situation. The storyline for this emission scenario corresponds to the "A2" scenario described by Jordan et al. (2000) and McCarthy (2001). It envisages a heterogeneous future world with a continuously growing population. Emphasis is given to local, short-term solutions instead of long-term, globally-oriented, and sustainable ones. Free market, consumerism, and increase of income per capita are pursued all over the world. The promotion of clean and resource-efficient technologies will be very limited, and the main source of energy will still be fossil fuels. Global inequality will grow. Under these conditions, it is hypothesized that the atmospheric CO_2 concentration will reach about 525 ppmv by the year 2050.

Put differently, this scenario assumes that Nature is very resilient to human stress; that global warming is a natural process in which anthropogenic activities do not play a significant role; and that sustainability is a rather expensive and unachievable goal.

GCM simulations (CGCM1: Boer et al., 2000; and HadCM2: Johns, 1996) under this emission scenario suggest that 30-year mean climate changes at regional levels will be very likely to happen in the future. For Germany in particular, the most expected climatic disruptions in the future are summarized below (for details see also Table 6.4).

Precipitation will increase in winter due to an intensified hydrological cycle but will decrease in summer because of an increased evapotranspiration. Furthermore, the intensity and frequency of extreme precipitation events in summer will likely increase, mainly because of changes in atmospheric moisture, thunderstorm activity, and large-scale storm activity. (Hennessy et al., 1997; McGuffie et al., 1999). In other words, the return period of extreme events will be shortened. Consequently, magnitude and frequency of high flows will most likely increase. It is also very likely that low-flow periods or droughts will increase due to greater evaporation (Gregory et al., 1997). Mean temperature will very likely increase in both seasons. The frequency of minimum and maximum temperatures will also change, i.e. fewer cold and frost days in winter, and much dryer and hotter days in summer (Houghton et al. 2001). In other words, weather patterns in this future world will become more intense and erratic.

Scenario C2

The keyword for this scenario is optimistic. The storyline of this emission scenario corresponds to scenario "B1" described by Jordan et al. (2000) and McCarthy (2001). It describes a convergent future world with a global population stabilizing in mid-century. "Global Sustainability" is the motto of all governments on Earth, which implement global solutions for economic and environmental issues. Most of the energy demand will be covered by renewable energy sources (e.g. biomass, solar, hydroelectric, tidal power, eolic, and geothermic). Promotion of clean and resource efficient technologies will be a key element of the decision-making process. As a result, CO_2 emissions as well as other greenhouse gasses will decrease after 2050, and the atmospheric CO_2 concentration will reach about 550 ppmv only by the year 2100.

The GCM (CGCM1: Boer et al., 2000; and HadCM2, Johns 1996) fed with these conditions predict that the climatic changes in Germany will be much less severe than those in scenario C1; in fact, the difference between these scenarios in growth rate per decade for both mean precipitation and temperature is in relation of 3:1 approximately. The mean temperature increase, for instance, by the year 2020 will remain under the 95% confidence interval of the natural variability (McCarthy, 2001), but the change of mean precipitation in winter will certainly exceed the natural variability of the last century (about 0.1% per decade, New et al., 2000). The detailed information obtained from these simulations is shown in Table 6.4.

6.4.3 Assembling the Development Scenarios

The assessment of the future state of a complex system requires a starting point situation and a reference framework from which the system will evolve into the future, i.e. a development scenario. In the present case, the starting point is the state of the catchment of the river Körsch in 1993. The

development scenarios will be assembled by combining one socio-economic scenario (S1, S2) with one macroclimatic scenario (C1, C2) at a time. As a result, four development scenarios are obtained, which are called C1S1, C1S2, C2S1, and C2S2 respectively. The specific conditions for each of them are shown Table 6.4.

Table 6.4 Composition of the development scenarios.

Variable			**Development Scenario**			
Description	**Name**	**Class / Season /Cat.**	**C1S1**	**C2S1**	**C1S2**	**C2S2**
Land-cover			Change [% / year]			
	x_{17}	Forest	-0.9	-0.9	+0.7	+0.7
	x_{18}	Impervious cover	+1.3	+1.3	+0.4	+0.4
	x_{19}	Permeable cover	-0.8	-0.8	-0.5	-0.5
Mean precipitation[II] $\{x : F(x) < 0.9\}$			Change [% / decade]			
	x_{24}	Winter	+4.1	+1.6	+4.1	+1.6
	x_{25}	Summer	-2.7	-1.0	-2.7	-1.0
Low precipitation[II] $\{x : F(x) \leq 0.1\}$			Probability of occurrence[I] $\Pr(X \leq x)$ [-]			
	x_{24}	Winter	*	*	*	*
	x_{25}	Summer	0	*	0	*
High precipitation[II] $\{x : F(x) \geq 0.9\}$			Change in probability and magnitude [% / decade]			
	x_{24}	Winter	*	*	*	*
	x_{25}	Summer	+4.0	*	+4.0	*
Mean temperature[II] $\{x : F(x) < 0.9\}$			Change [% / decade]			
	x_{30}	Winter	+2.1	+0.8	+2.1	+0.8
	x_{31}	Summer	+2.9	+1.1	+2.9	+1.1
Low temperature[II] $\{x : F(x) \leq 0.1\}$			Probability of occurrence[I] $\Pr(X \leq x)$ [-]			
	x_{30}	Winter	0	*	0	*
	x_{31}	Summer	0	*	0	*
High temperature[II] $\{x : F(x) \geq 0.9\}$			Change in probability and magnitude [% / decade]			
	x_{30}	Winter	*	*	*	*
	x_{31}	Summer	+10.0	*	+10.0	*
Annual precipitation[III]			Change [% / decade]			
	x_{21}	Winter	+3.9	+1.6	+3.9	+1.6
	x_{22}	Summer	-2.7	-1.0	-2.7	-1.0
Maximum API[III]			Change [% / decade]			
	x_{27}	Annual	+1.4	+0.5	+1.4	+0.5
	x_{28}	Winter	+3.9	+1.5	+3.9	+1.5
	x_{29}	Summer	0.0	0.0	0.0	0.0
Maximum Temperature[III]			Change [% / decade]			
	x_{32}	Winter	+2.0	+0.7	+2.0	+0.7
	x_{33}	Summer	+1.6	+0.6	+1.6	+0.6
ATI at annual peak[III] discharge			Change [% / decade]			
	x_{34}	Annual	-0.2	-0.1	-0.2	-0.1
Duration of a given category of Circulation Patterns[III, IV]			Change [% / decade]			
	x_{41}	Winter / Wet	+6.9	+2.7	+6.9	+2.7
	x_{40}	Summer / Wet	-8.9	-3.5	-8.9	-3.5
	x_{38}	Summer / Dry	+9.0	+3.3	+9.0	+3.3

Notes: * Denotes that there will be no significant change in magnitude or that the probability of occurrence will remain equal to that of the reference period 1961-1993.

I Based on the PDF of the variable during the reference period.

II Based on GCM simulations carried out by IPCC (CGCM1: Boer et al., 2000; and HadCM2: Johns, 1996) under a given emission scenario.

III Based on potential relationships between a given variable and mean precipitation and mean temperature at catchment level. Winter, summer, and annual relationships are

$$x_i^t = \beta_0 \left(x_{24}^t\right)^{\beta_1} \left(x_{30}^t\right)^{\beta_2} + \varepsilon_i^t$$

$$x_i^t = \beta_0 \left(x_{25}^t\right)^{\beta_3} \left(x_{31}^t\right)^{\beta_4} + \varepsilon_i^t$$

$$x_i^t = \beta_0 \left(x_{24}^t\right)^{\beta_1} \left(x_{30}^t\right)^{\beta_2} \left(x_{25}^t\right)^{\beta_3} \left(x_{31}^t\right)^{\beta_4} + \varepsilon_i^t$$

respectively. This formulation has the advantage that can be easily transformed into an incremental equation. For example, for winter the incremental equation is

$$\Delta x_i^t = \beta_1 \Delta x_{24}^t + \beta_2 \Delta x_{30}^t .$$

The parameters β_i were found empirically and all are significant at the 5% level. The following Table shows these coefficients for each variable:

Variable	β_1	β_2	β_3	β_4
x_{21}	0.956	0.026		
x_{22}			0.862	-0.131
x_{27}	0.290	0.017	0.475	0.508
x_{28}	0.945	0.048		
x_{29}			0.808	0.765
x_{32}	0.125	0.703		
x_{33}			-0.042	0.516
x_{34}	-0.032	0.004	0.020	0.009
x_{38}			-2.492	0.677
x_{40}			2.796	-0.603
x_{41}	1.483	0.361		

IV This fact is also supported by the excellent agreement found between observed and downscaled monthly precipitation at catchment level (Bardossy and Caspary, 1999) using the CPs.

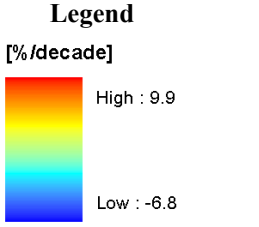
6.5 Simulation Results

A total number of 2500 realizations have been carried out for each development scenario (simulation time $\sim$ 7.5h on an 800 MHz workstation). Based on the simulation results, the following summary has been prepared to show how the conditions of each scenario have influenced the runoff characteristics of the Special Study Area, however, a more detailed analysis of results and conclusions will be presented in Chapter 8.

Firstly, Table 6.5 shows the average growth rate in percent per decade for each simulated variable and for each scenario, taking as reference year the beginning of the simulation, i.e. 1994. This information is presented in both a tabular and a visual way to ease the comparison between different scenarios and types of impact measured by the simulated variables. Based on this optical aid, it can be clearly seen that the hydrological system of the studied catchment will endure the greatest disruptions under the C1S1 scenario conditions, and conversely, the least ones under scenario C2S2. The other two scenarios, i.e. C1S2 and C2S1, are in-between the previous two.

Table 6.5 Average percent change per decade for each simulated variable taken 1994 as reference year. The colours indicate the magnitude and the sign of the simulated changes (see legend below); e.g., red represents the highest positive change whereas dark blue does the opposite.

Variable Description	Symbol	Development Scenario				Development Scenario			
		C1S1	C1S2	C2S1	C2S2	C1S1	C1S2	C2S1	C2S2
Total discharge in winter	Q_2	6.9	5.4	3.7	2.4				
Total discharge in summer	Q_3	-2.6	-6.8	0.4	-4.1				
Specific peak in winter	Q_4	8.8	5.4	5.4	2.5				
Specific peak in summer	Q_5	-3.7	-1.6	0.1	-0.6				
Specific volume of the annual peak	Q_6	9.9	3.2	8.0	2.2				
Total duration of high flows in winter	Q_9	5.6	6.2	2.3	2.7				
Total duration of high flows in summer	Q_{10}	-1.9	-4.5	1.8	-1.1				
Frequency of high flows in winter	Q_{11}	7.1	3.5	4.4	1.3				
Frequency of high flows in summer	Q_{12}	-2.8	-2.6	-1.2	-1.8				
Total drought duration in summer	Q_{14}	8.4	8.0	3.7	3.8				



Each average growth rate has been estimated based on the simulated mean values for 1994 and 2025. This approach yields satisfactory approximation of the decadal growth rates since the simulated mean value grows continuously during the simulation period as can be seen in Figure 6.7.

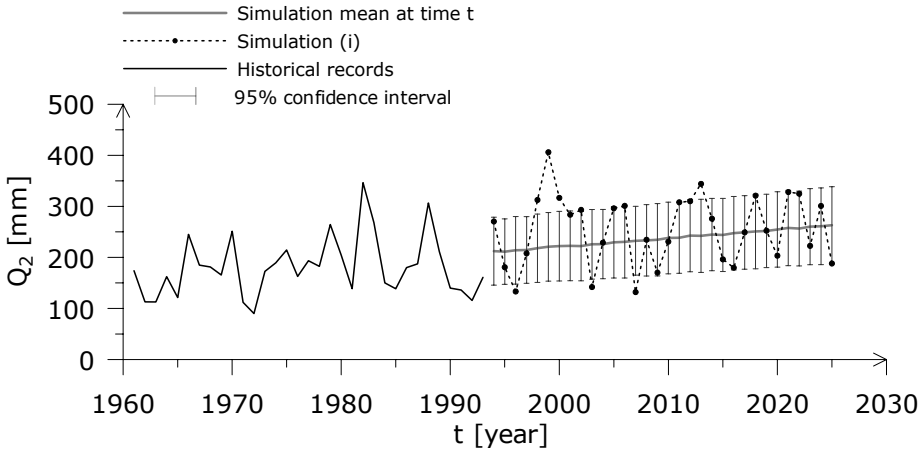


Figure 6.7 Historical records for total winter discharge (Q_2) in the Special Study Area from 1961 to 1993. The dotted line on the right depicts one of the realizations of this variable for the period 1994 to 2025 under C1S1 scenario conditions. The continuous line denotes the simulated mean value at a give time t, which has a positive trend in this case (i.e. 6.9% per decade, see Table 6.5).

From the simulations, it is also possible to estimate the likelihood that the long-term mean of a given variable will be exceeded during the period 1994-2025 under a given scenario. The summary of these exceedance probabilities and the long-term means (i.e. from 1961 to 1993) for each variable and scenario are presented in Table 6.6. If the probability is greater than 0.95, this means that it is very likely that the past mean of a given variable will be surpassed in the future, or in other words, that the expectation of a variable will increase over time. On the contrary, a value less than 0.05 will mean that the past mean of a variable will be hardly reached (in fact it will only occur 5% of the time at this probability level), thus, a decreasing tendency of the expectation of such a variable is very likely foreseeable.

Table 6.6 Probability that the long-term mean for a given variable will be exceeded under certain scenario conditions.

Variable Description	Symbol	Long-term mean	Unit	Development Scenario			
				C1S1	C1S2	C2S1	C2S2
Total discharge in winter	Q_2	181.2	[mm]	1.000	1.000	1.000	1.000
Total discharge in summer	Q_3	153.4	[mm]	0.332	0.020	0.238	0.020
Specific peak in winter	Q_4	7.1	[mm]	0.218	0.037	0.052	0.002
Specific peak in summer	Q_5	9.1	[mm]	0.335	0.335	0.231	0.229
Specific volume of the annual peak	Q_6	25.8	[mm]	1.000	0.992	1.000	0.973
Total duration of high flows in winter	Q_9	10.0	[day]	0.732	0.775	0.580	0.649
Total duration of high flows in summer	Q_{10}	8.1	[day]	1.000	0.995	0.998	0.984
Frequency of high flows in winter	Q_{11}	4.3	[-]	0.920	0.712	0.786	0.460
Frequency of high flows in summer	Q_{12}	4.9	[-]	0.265	0.219	0.071	0.060
Total drought duration in summer	Q_{14}	21.2	[day]	0.437	0.423	0.703	0.716

The long-term means for both simulated and observed values for each runoff characteristic can also be plotted in order to visualize the effects of a given development scenario on a given runoff characteristic. In the present case the deviation from the mean of the historical records (1961-1993) expressed in percent has been found appropriate for this purpose. In addition to the magnitude of the deviation, which is shown in Figure 6.8 by a dot, it is also very important to know the degree of

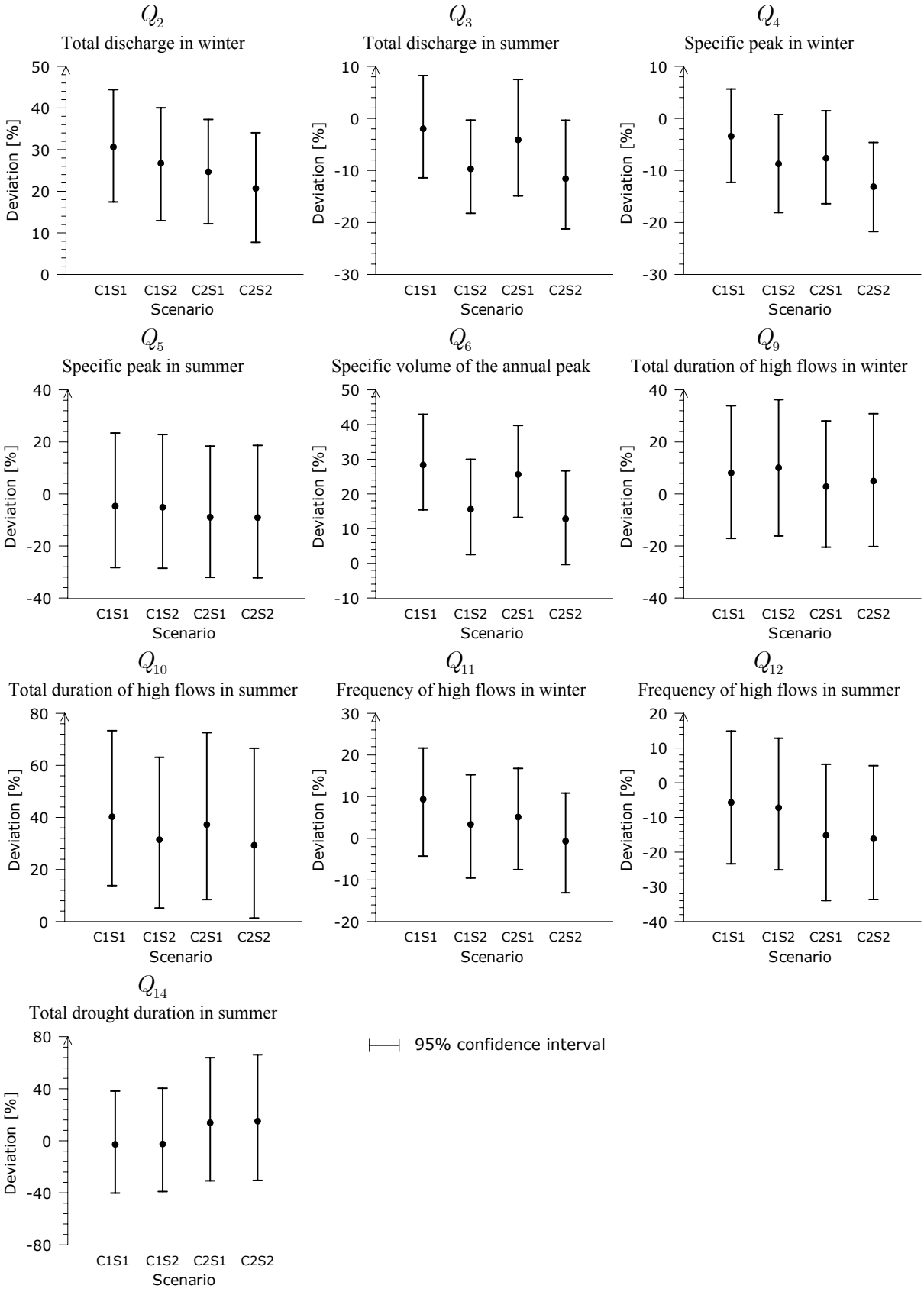


Figure 6.8 Deviations in percent of the mean of the simulated variables with respect to the respective historical mean (i.e. observations during 1961-1993) under given scenario conditions. The mean value and its 95% confidence interval are represented here with a dot and a bar respectively.

dispersion of this indicator. This has been achieved by plotting the 95% confidence interval of the simulated mean with respect to the reference period.

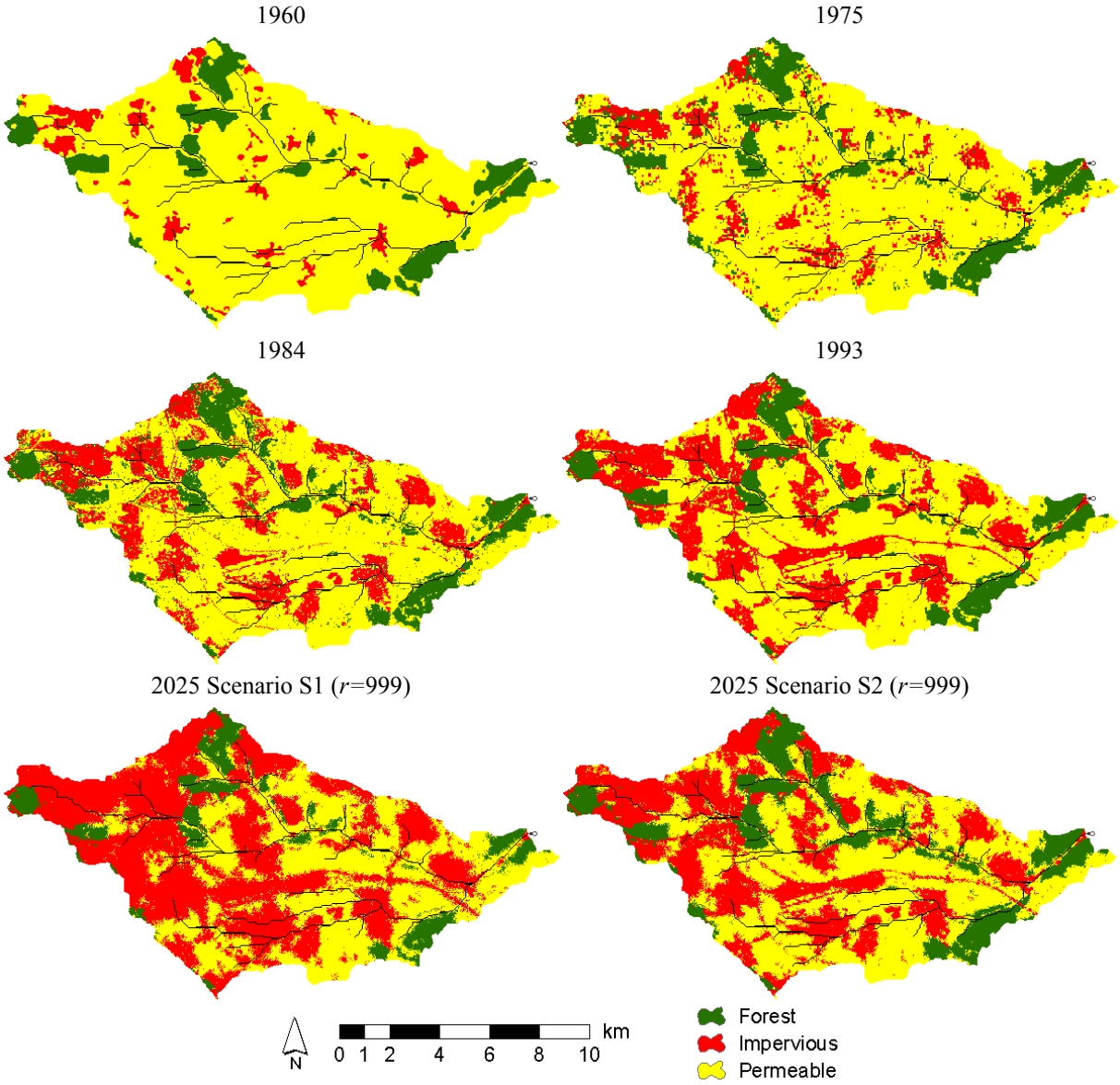


Figure 6.9 Time series of land cover in the Special Study Area from 1960 to 1993. Additionally, random realizations of the land use/cover for the year 2025 under two different scenario conditions.

As said above, 2500 land use/cover realizations were conducted for each scenario. The spatial domain of the catchment was divided into cells of 30×30 m with a total extension of 745 cells in west-east direction and 393 cells in north-south direction. The state of each land use/cover category for each cell has been reckoned during the simulation period based on the model proposed before. The land use/cover balance in the basin is estimated after each LUCC simulation has been finished. A sample of such results can be seen in Figure 6.10. These simulated values are subsequently employed for the evaluation of the runoff characteristics at the correspondent point in time. The results of the LUCC simulation, however, were not kept in order to speed up the simulation. The model, nevertheless, can deliver one of such realizations at a certain point in time, as can be seen in Figure 6.9 for socio-economic scenarios S1 and S2 respectively.

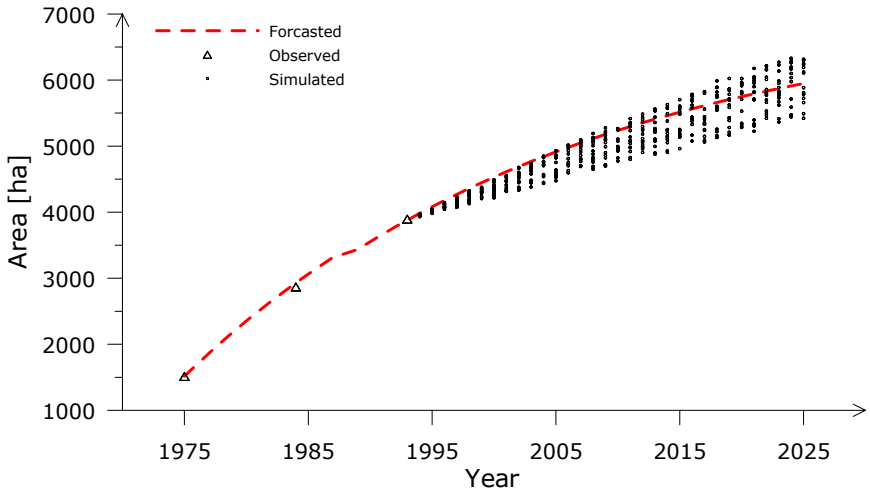


Figure 6.10 Sample from the land use/cover simulations showing the evolution of impervious cover in the Special Study Area based on socio-economic scenario S1. The forecasted trend and the observations have been depicted as a reference.

Additionally, the probability that the land use/cover state of a given cell will be transformed to a different state during the time span of the simulation can be estimated from the simulation results. Figure 6.11 shows, for instance, the spatial distribution of the probability that the land use/cover in a given location of the basin will be transformed to impervious cover by the end of 2025. In this Figure, for instance, the red colour indicates that the probability that a cell would be transformed to impervious cover (e.g. road, urban settlement) is greater than 0.9. This situation, as shown in this Figure, tends to occur mainly in the fringes of existing settlements were available land with particular morphological and accessibility conditions remains still under other usages.

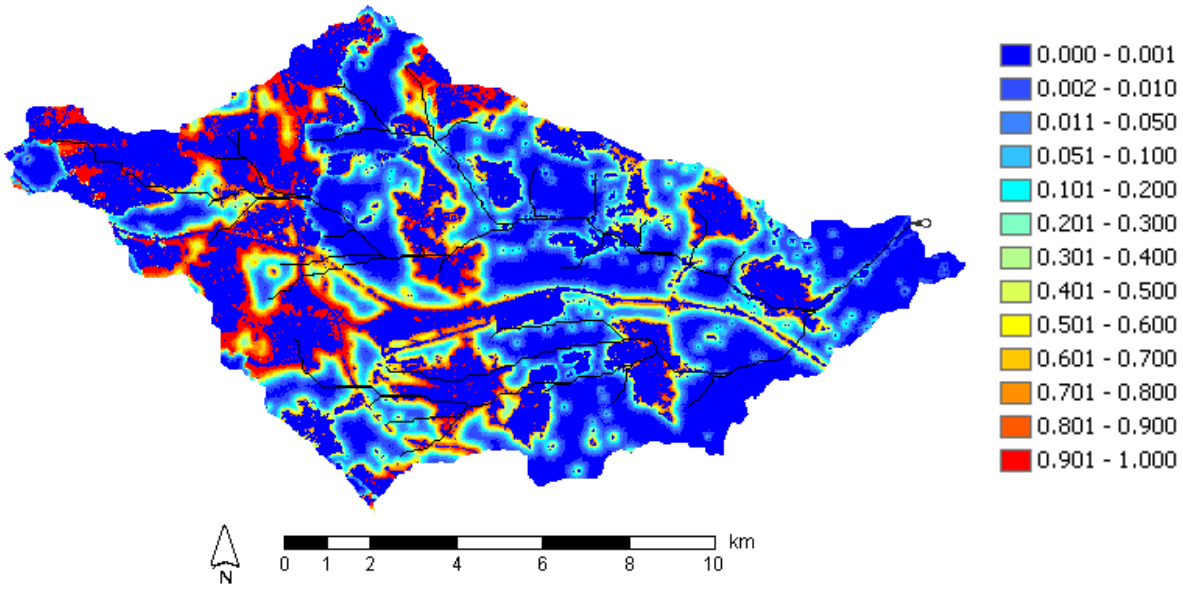


Figure 6.11 Probability that the land use/cover of a given location will be transformed to impervious cover up to the year 2025 based on the socio-economic scenario S1. (The sample size for each cell is 2500).

Chapter 7

Sensitivity Analysis

7.1 Introduction

Once a model has been selected, calibrated and validated, it is of crucial importance to study how changes in variables, parameters, and model structure would affect the behaviour of the model output. Such a study is generally known as sensitivity analysis (Gilchrist, 1984). As to the model user, the sensitivity analysis will provide him/her all required information and insight about the model performance and its limitations, which, in turn, will contribute to reduce the risk of an inappropriate application of the model.

It should be noted that the sensitivity with regard to model structure was already considered during the model selection (see Chapter 4 and Chapter 5). Therefore, the present chapter will go through the remaining issues, namely: 1) sensitivity of model parameters to a given variable, 2) model sensitivity to a given parameter, and, 3) sensitivity of the significance probability ($p-\text{values}$) as to the number of replicate simulations R.

7.2 Sensitivity of Parameters to Catchment Size

One of the major concerns in the present study is to investigate the effects of the spatial scale at which the model is optimised with regard to the model parameters and its overall performance. In other words, it would be necessary to answer the question: are the model- parameters invariant with regard to the spatial scale? In this case, the spatial scale is represented by the catchment size x_1, which ranges from 4.5 to 4002.0 km^2.

In order to illustrate the procedure presented below, the model No. 3733 fitted for the annual specific discharge in winter (see Section 4.1.4) will be used as an example. In this case, the model estimates can be written as

$$\hat{Q}_{i2}^t = f\left(x_{i7}^t, x_{i8}^t, x_{i11}^t, x_{i15}^t, x_{i17}^t, x_{i19}^t, x_{i21}^t, \hat{\boldsymbol{\beta}}\right) \quad i = 1,\ldots,46 \quad t = 1961,\ldots,1993\,, \tag{7.1}$$

using the vector $\hat{\boldsymbol{\beta}}$ as in Section 4.1.4, Table 4.5.

Algorithm 7

1. For all $a = 10, 25, 50, 100, 200, 250, 500, 1000, 2000, 3500, 4500$, where a is a threshold for the variable x_1 given in [km²].
   a. Build a sample $\mathcal{D}_a$ of size n_{0a} so that $x_{i1} < a \quad \forall i = 1, \ldots, 46$.
   b. Use $\mathcal{D}_a$ to estimate $\hat{\boldsymbol{\beta}}_a$ for the model $Q_{i2}^t = f\left(x_{i7}^t, x_{i8}^t, x_{i11}^t, x_{i15}^t, x_{i17}^t, x_{i19}^t, x_{i21}^t, \hat{\boldsymbol{\beta}}_a\right) + \varepsilon_i^t$ so that $\Phi_a \rightarrow \min!$
   c. Estimate the Akaike's Information Criterion $\mathrm{AIC_a}$ for the previous model.
2. Repeat step 1. if needed.
3. Plot n_{0a}, $\mathrm{AIC_a}$, and $\hat{\boldsymbol{\beta}}_a$ versus a.

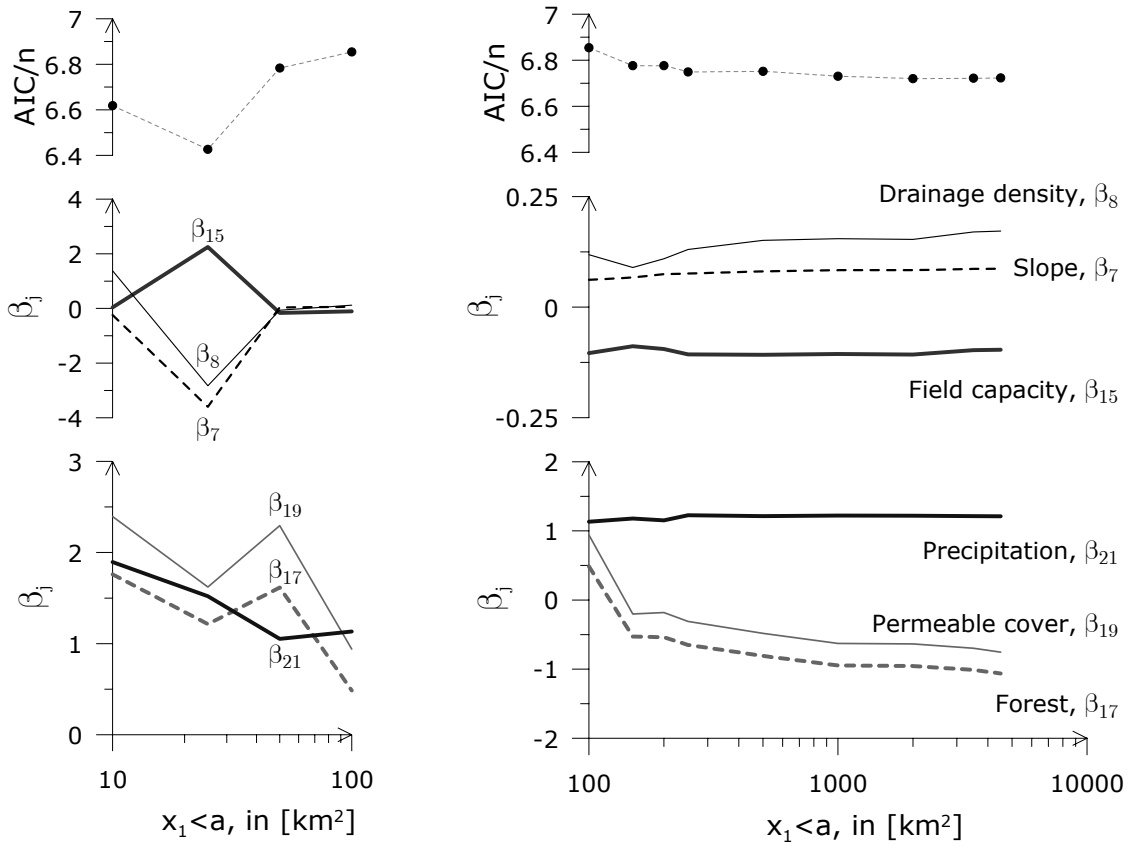


Figure 7.1 Parameter sensitivity to catchment size for the multi-linear potential model (No. 3733) selected for the annual specific discharge in winter Q_2. Samples are from the period from 1961 to 1993.

Based on the results of the previous Algorithm, which are depicted in Figure 7.1, the following remarks can be formulated.

1. Since AIC is proportional to the sample size, the ratio AIC_a / n_{0a} can be used to compare the results obtained by the previous Algorithm with regard to the quality of the model with respect to the amount of information provided. As can be seen in the Figure above, this indicator reaches a peak at around 100 km² and then decreases slowly. Based on this finding, it can be inferred that the amount of information for those samples with spatial units whose area is less than $a\,(\sim 100\,\mathrm{km}^2)$ is not as complete as for those derived with thresholds greater than or equal to $a\,(\sim 100\,\mathrm{km}^2)$.

2. Parameters for the morphological variables (i.e. x_7, x_8, x_{11}, and x_{15}) exhibit an irregular behaviour (as to their sign and magnitude) when the samples used for the model-calibration have spatial units with an area less than $a\,(\sim 100\,\mathrm{km}^2)$. This can be regarded as a direct consequence of what has been mentioned above since the data here perhaps reflects a case specific situation unwanted for a model supposed to describe the phenomenon at a much broader scale. However, values for the analyzed parameters tend to stabilize for threshold values a greater than 100 km² (see right panel of Figure 7.1). As this example has shown, parameters of morphological variables in a model conceived to explain the specific discharge for a given catchment can be considered as scale invariant if $a > 100\,\mathrm{km}^2$.

3. The parameter β_{21}, which is linked with the climatological variable total precipitation, has a downward trend within the interval $a \in [10, 50]\,[\mathrm{km}^2]$ and becomes asymptotic when it reaches a magnitude of about 1.2 (see left panel of Figure 7.1). Thus, it can be stated that this parameter is scale invariant for values of $a > 50\,\mathrm{km}^2$. Additionally, it should be noticed that its order of magnitude is several times greater than that of the morphological variables. Such a fact just points out how important this variable is with regard to discharge predictions at a mesoscale level.

4. Finally, those parameters associated with land cover variables such as forest and permeable cover (β_{17} and β_{19}) exhibit in general a downward tendency, keeping an almost constant relationship between them. Because of this fact, it can be inferred that these variables have a complex relationship to the water system, which depends greatly on the scale at which the analysis is carried out. Consequently, their corresponding parameters appear to be scale dependent as shown in the Figure 7.1.

7.3 Model Sensitivity to a Given Parameter

In many cases, it would be desirable to know how changes of a parameter (e.g. due to errors of estimation caused by data quality) would influence the behaviour of the model output. In other words, to assess how the uncertainty of one parameter can influence the model results (Mein and Brown, 1978).

A simple procedure to assess the percentage rate of change in the expected output $\hat{Q}_{il}^t$ per unit of percentage change in the parameter β_j, frequently referred as relative sensitivity, is presented below.

Let

$$\hat{Q}_{il}^t = f\left(x_{i1}^t, \ldots, x_{iJ}^t, \hat{\boldsymbol{\beta}}\right) \qquad i = 1, \ldots, 46 \quad j = 1, \ldots, J \quad t = 1961, \ldots, 1993\,, \tag{7.2}$$

be a general model for a given runoff characteristic l, which depends on J predictors x_j, and where $f(\bullet)$ and $\hat{\boldsymbol{\beta}}$ are a known functional form and a vector of estimated parameters respectively. Based on these definitions, the rate of change of $\hat{Q}_{il}^t$ with β_j or simply the absolute sensitivity coefficient c_{ij}^t, can be computed as (McCuen, 1973, Leavesley et al. 1983, Gilchrist, 1984)

$$c_{ij}^t = \frac{\partial \hat{Q}_{il}^t}{\partial \beta_j}\Big|_{\hat{\boldsymbol{\beta}}}\,, \tag{7.3}$$

where the partial derivative is evaluated at $\hat{\boldsymbol{\beta}}$. Absolute sensitivities, however, have the serious disadvantage that the values estimated for two different parameters cannot be directly compared because their values largely depend on the magnitudes of each parameter respectively. Therefore, dimensionless-relative sensitivities are preferred in practice. The relative sensitivity of the model output, e_{ij}^t, with respect to the parameter $\hat{\beta}_j$ can be written as

$$e_{ij}^t = c_{ij}^t \frac{\hat{\beta}_j}{E\left[\hat{Q}_{il}^t\right]} = \left(\frac{\partial \hat{Q}_{il}^t}{E\left[\hat{Q}_{il}^t\right]}\right)\left(\frac{\partial \beta_j}{\hat{\beta}_j}\right)^{-1}, \tag{7.4}$$

where $E\left[\hat{Q}_{il}^t\right]$ is the expectation of the output given by $E\left[\hat{Q}_{il}^t\right] = f\left(x_{i1}^t,\ldots,x_{iJ}^t,\hat{\boldsymbol{\beta}}\right)$.

Figure 7.2 illustrates for a specific case how the two factors shown in parenthesis in (7.4) are related to each other, considering three different parameters. Based on this Figure, it can be concluded that the most sensitive parameter in this case is β_{22}, which is associated with the variable cumulative precipitation in summer, and the least β_{17}, which is associated with the share of forest of a given basin. These results show that the system is highly sensitive to precipitation and much less sensitive to land cover or slopes. These results are not surprising because the system is mainly governed by climatic variables, and only modulated by the morphology and the land cover of a given basin. However, it should be noted that the magnitude of the relative sensitivity of the parameter β_7 associated with the mean slope in the buffer zones of the stream network is quite similar to that of β_{17}. This result suggests that the sensitivity of the model to a change of the parameter value for land cover is as important as that corresponding to mean slopes. Nevertheless, the sign of the changes of the output will be the opposite because these variables (x_{17} and x_7) associated with these parameters have an inverse and a direct relationship with the model output respectively as can be seen in the Figure below.

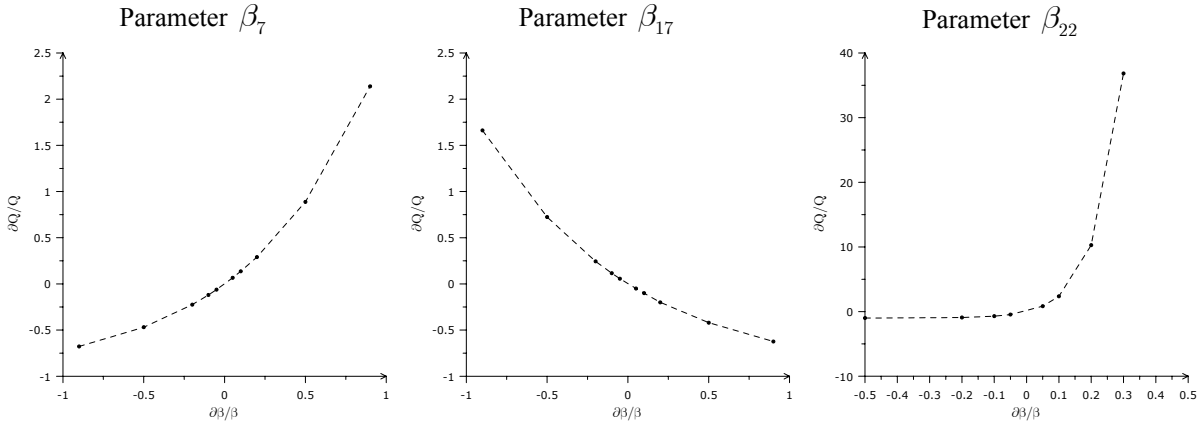


Figure 7.2 Relationship between $\partial \hat{Q}_{il}^t \big/ E\left[\hat{Q}_{il}^t\right]$ and $\partial \beta_j \big/ \hat{\beta}_j$ for model No. 3965 obtained for the specific discharge in summer (Q_3). The relative sensitivities for each parameter at a given level can be obtained as the quotient between an ordinate and its corresponding abscise. The dots represent the geometric mean of the relative changes taking into account all observations is the sample.

7.4 Convergence of the Monte Carlo Simulations

A randomization test is to be performed in order to assess whether an explained variable Q is either statistically independent (H_0) or dependent on a given variable x_j under a joint distribution of J predictors. As mentioned in Section 3.3.7, the estimator Φ measures the level of interdependence between Q and x_j (ceteris paribus), and the $p-$value indicates whether to accept or reject the null hypothesis in favour of the alternative one at a given label of significance α, say 5%. However, this procedure can be executed only if one knows in advance how many replicates of the statistical test have to be carried out in order to have a conclusive result, which, in turn, leads to take the right decision.

Of course, the more replicates the better, but a large value (say $R > 10\,000$) still constitutes a great hindrance at the actual state of development of desktop computers, namely a dramatic increase of computing time. This side effect would then make this procedure too time consuming to be applied for practical purposes. Therefore, it would be advantageous to establish a certain minimum number of simulations required to guarantee a stable result.

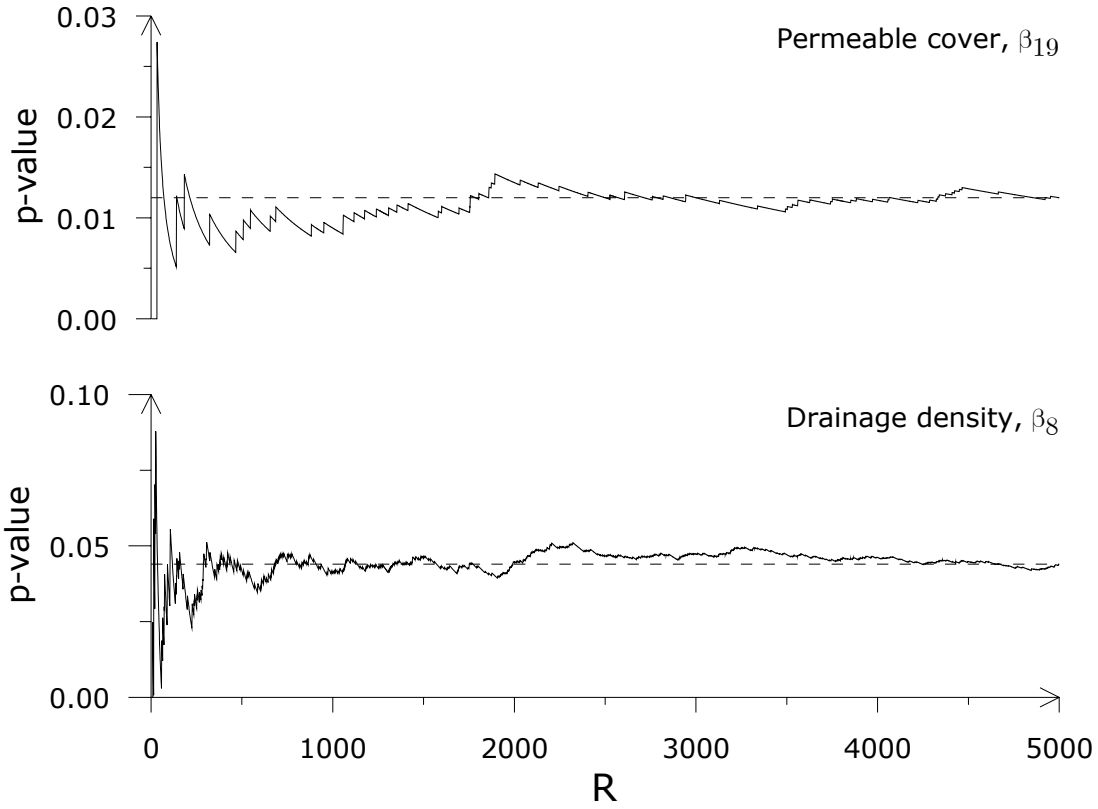


Figure 7.3 Sensitivity of the $p-$value with respect to the number of replicate simulations. Panel on top shows the results for variable x_{19}; and panel down depicts the results for variable x_8. The simulations are carried out for model No. 3733 fitted for the annual specific discharge in winter.

As shown in Figure 7.3, the $p-$value obtained for two variables x_{19} and x_8 that are part of the model No. 3733 calibrated for the annual specific discharge in winter [see (7.1)] tend to converge to a certain limit, which is the value that would be obtained if R would tend to infinite. As a rule of thumb, it is suggested that reasonable estimates can be obtained when R varies between 100 and 1000

(Davison and Hinkley 1997). The convergence of the $p-$value is commonly achieved when $R \simeq 500$ as actually it happened with the variable x_8 shown in the Figure below. However, there are cases in which convergence only occurs for values of $R > 1000$ as it is the case of the variable x_{19}. Since this is not known in advance, a good recommendation would be to continue with the simulations if the obtained $p-$value is too close (about 10%) to the level of significance decided in advance for the test of independence.

In the present case, since the $p-$value obtained for the variable x_{19} is much less than 5%, the simulation could have been stopped at $R = 500$.

A great advantage of this test with relation to the parametric tests is that in this case the PDF of the test statistic (e.g. the estimator) is not assumed but rather built up from the simulation results as shown in Figure 7.4. The $p-$value estimated by the Algorithm 4 (see Section 3.3.7) is the area on the left tail of the empirical distribution function of Φ that is less than or equal to the value of the estimator given the original sample, for the model 3733, this value is $\Phi = 0.9342$. As seen in the Figure below, the EDF is not symmetrical and skewed to the left.

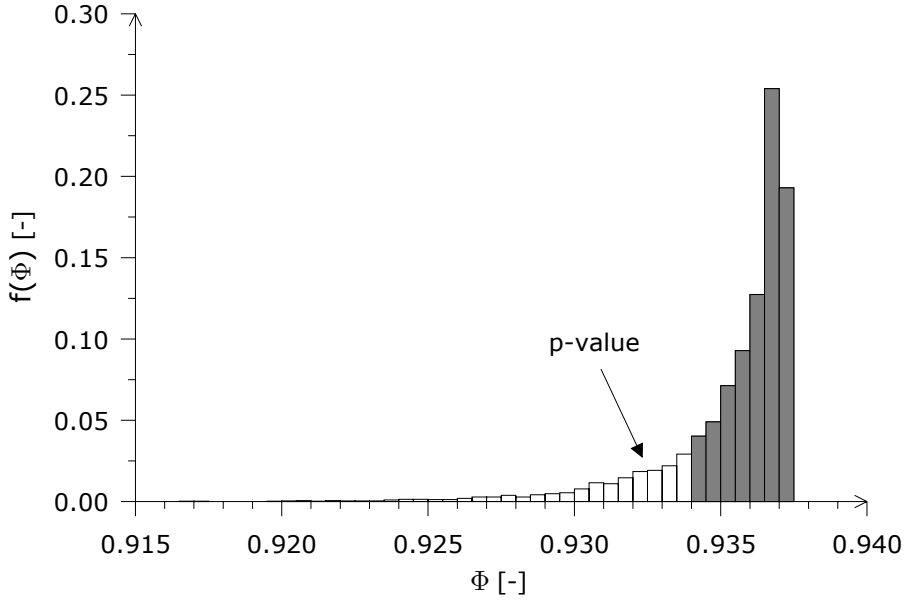


Figure 7.4 Histogram of $R = 5000$ Monte Carlo replicates of the estimator Φ for the model No. 3733 when the variable x_8 has been tested for independence. The unshaded area in the left tail correspond to the $p-$value.

Chapter 8

Discussion and Conclusions

8.1 Discussion

In this section, some remarks as to the methodology and results presented in previous chapters are to be set forth and discussed.

First, it should be stated that the methodology employed in this study is general and can be applied anywhere if the required information is available. The results obtained, however, are specific for the Upper Neckar catchment.

Concerning the methodology employed in chapter three to five to select parsimonious models for ten runoff characteristics having just the minimum number of variables and parameters has been proved convenient with regard to providing insight into the functioning of the system and easing their applicability in more complex simulation models. Those variables that constitute a given model were selected taking into account two conditions. Firstly, the number of variables in a model should be as close as possible to the dimensionality of the system (i.e. around 7); and secondly, in each model there should be at least one variable directly linked to the morphology, to the climate factors, and to the land cover state of the basin respectively. By doing so, these models not only have saved considerable computing time in the subsequent simulations carried out in Chapter 6, but also have been able to react to changes in macroclimate and land cover as can be appreciated in Table 6.5. In other words, these minimalist models have been capable of detecting many of the entangled relationships between the predictors (which are very often non-linear) contained in their unknown joint distribution function.

Additional advantages of the selection procedure are twofold. The risk of over-parameterization as well as the possible multicollinearity among predictors was considerably minimised. A direct consequence of the latter is, for instance, the significant reduction of the confidence intervals of all model's parameters.

The inclusion of statistically significant variables in a model is also of key importance with regard to finding "good" but "simple" models among the numerous possibilities given by a set of predictors. The main reason for this is that a non-significant variable will increase the total variance, but will not contribute to explain it better. In other words, it will only add noise to the system and deteriorate the explanatory power of other significant predictors. In this respect, the randomisation test employed has proved to be an indispensable analytical tool compared with any conventional parametric statistical tests. In this case, considering the fact that the multivariate joint distribution function of the predictors is

unknown, the latter would have provided misleading decision results as to which variable is to be in or out of a model. This would have undoubtedly occurred since all parametric tests are based on assumptions with regard to the distribution functions of the variables.

The use of the Jackknife statistic during the cross-validation of the best models has tremendously facilitated the task of the selection of the "best" model. Additionally it was of essential importance in the present study since it allows estimating at the same time the level of predictability and the robustness of a model in presence of data that contain outliers. One important advantage of this statistic is that it can be used always regardless of the estimator employed.

It has to be mentioned that all of these statistical techniques are very effective for modelling complex systems but they require substantial computing resources, which will increase non-linearly with respect to the amount of data employed.

As said above, the integration of two realms of the system, namely the hydrological behaviour of a catchment and the state of the land cover at a given point in time has been achieved in this study because of the simplicity of the hydrological models employed describing many runoff characteristics and of the character of the Land Use/Cover Change (LUCC) model, which in spite of its simple formulation, has reached an 85% level of predictability in its validation phase (see 6.3.2). This result has been achieved mainly because of the spatially distributed character of the LUCC model. Recalling its basic formulation, it makes the transition probability from one land cover type to another depend on external driving forces that vary over space but not in time. In this case, the driving forces behind land cover change have been considered static just to keep the model as simple as possible under the existing constraints of time and resources assigned for this research. This shortcoming should be improved in future versions, because some of the external driving forces behind land cover change may vary over time, for instance accessibility to towns and settlements with access to railway connection. The LUCC model, however, can be modified to accept dynamic driving forces based on the same formulation presented here. Another aspect of the LUCC model that should be improved in future versions is its link with other factors that induce a land use change. The land use of a given location depends among other factors on its accessibility, the existing land use regulations, and the location of residents and job places. Under specific conditions the state of these variables are such that they can induce a land use change. Then, eventually, a land cover change will follow with all its consequences to the environment. This coupling between land use and land cover models is still a challenge for future versions of this kind of simulation model.

Secondly, as to the results obtained in this study the subsequent remarks can be formulated, which, as was already said, are relevant for the Upper Neckar catchment in general and for the Körsch catchment in particular.

Based on the relationships thoroughly discussed in Chapters 4 and 5, which relate several runoff indicators for basins exhibiting various sizes and morphological characteristics, having different percentages of land cover shares and at different points in time, it can be inferred that land cover variables are certainly having an effect on the hydrological cycle at mesoscale basins. In all cases that have been analysed, the subset of the best performing models always contains one or more of these

variables. The test of independence has shown that such variables are certainly not independent of the explained variables at the 5% level, and in some cases at even less than that significance level.

The magnitude of the effects of land use/cover change on the hydrological cycle of a mesoscale basin, of course, cannot be compared with those triggered by sudden meteorological changes. It should be borne in mind that the hydrological system is driven by the weather and that the morphological characteristics of the basin and its land use/cover can only modulate the response of the system. However, the effects of land use/cover change are cumulative and may cause long-lasting consequences. For instance, they will further a continuous increase of the total discharge in winter and will induce long periods of drought in summer (see Table 6.5). Both effects will have enormous consequences for the environment and the economy of the region.

More specifically, the effects of land use/cover change on the hydrological system of a mesoscale basin under two extreme climatic scenarios can be summarised below.

The total discharge in winter, Q_2, will increase at about 6.9% per decade in the worst-case scenario C1S1, i.e. a very rapid urban sprawl (about 1.3% per year) accompanied by a continuous increase of mean air surface temperature caused by global warming. If the growth rate of impervious cover will decline to a modest 0.4% per year and climate will continue with the actual warming tendency the growth rate may be about 5.4% per decade (i.e. development scenario C1S2). These growth rates may decrease only if the global community really will minimize the amount of emissions of greenhouse gases to the atmosphere. If this becomes true, then these figures will be reduced to 3.7 and 2.4% per decade for development scenarios C2S1 and C2S2 respectively. The 95% confidence intervals shown in Figure 6.8 indicate that the average total discharge in winter up to 2025 would be 17% to 44% bigger that that of the base period (1961-1990) for scenario C1S1. In the most favourable scenario (C2S2) these figures will be as low as 7% and 34% respectively.

The total discharge in summer, Q_3, will in general tend to decrease because of higher temperatures and corresponding increasing evapotranspiration. This variable, for instance, will suffer a decrement of 6.8% per decade in the development scenario C1S2 (which contemplates an increase in forested areas). In development scenarios C1S1 and C2S1, however, it may endure increments as to the reference period.

Specific peaks in winter, Q_4, will tend to increase in all scenarios. However, the largest deviation from the historical mean corresponds to the development scenario C1S1 (i.e. urban sprawl accompanied with future climatic conditions exhibiting hotter winters). Land use/cover change plays a very important role in this runoff indicator. The difference in percent between the socio-economic scenarios S1 and S2 is about 3% per decade regardless of the macroclimatic settings.

Specific peaks in summer, Q_5, will tend to decrease in all scenarios with the exception of scenario C2S1. In the latter, summers will not be much hotter as during the reference period but an increase of impervious cover will reduce the concentration time of surface runoff, which, in turn, will tend to increase peak flows at the rate of 0.1% per decade. However, the confidence intervals estimated for each scenario show that this variable might have a large fluctuation around the mean of the base period. Summer peaks are often local phenomena caused by convective precipitation.

The specific volume of the annual peak event, Q_6, is the runoff characteristic that is mostly affected by land cover changes simulated in the Special Study Area. The difference between the growth rates of this variable under socio-economic scenarios S1 (i.e. urban sprawl) and S2 (i.e. densification) may range from 5.8% to 6.7% per decade, depending on whether the future macroclimate conditions will be either moderate or exacerbated respectively (i.e. climate conditions of scenario C2 or C1 correspondingly). Furthermore, if impervious cover will follow the actual trend, and mean temperature will increase at the rate of 0.4°C per decade (i.e. a grow of 2.4% per decade in average) due to global warming (C1S1), then, this variable will grow at about 9.9% per decade. This implies that if scenario C1S1 would become true, the future volume of the annual peak flow could be between 15% and 43% greater than during the reference period with a 95% level of confidence. In other words, more intensive floods can certainly be expected downstream of the Special Study Area.

Total duration of high flows in winter, Q_9, will be higher on average in the densification scenario (S2) than in the urban sprawl scenario (S1) assuming constant climatic conditions. In other words, discharges that occur less than five percent of the time will persist during longer periods. The reason for that stems from the fact that the former scenario promotes an increase of forest and restricts the development of impervious cover at a low growth rate. The latter, though, does just the opposite. A higher amount of forest would induce lower rates of evapotranspiration in winter, which, in turn, would tend to increase surface runoff. Additionally, it should be noted that even in the most favourable case (i.e. C2), the mean temperature in winter would grow at about 0.8% per decade based on actual trends for the Northern Hemisphere. This increase of heat in the system will induce a faster melting of the snowpack, which, in turn, will also contribute to increase the surface runoff and its persistence at higher discharge levels.

Total duration of high flows in summer, Q_{10}, will do just the opposite of its counterpart in winter, i.e. they will tend to decline in general, mainly because mean temperature in summer will increase. The growth rate in the densification scenario is even smaller because forest will grow under this scenario, which implies higher rates of evapotranspiration and, hence, lower surface runoff. The exception is the scenario C2S1 (i.e. moderate climate and urban sprawl) where this variable will tend to grow at just 0.1% per decade. The main reasons are the moderate temperatures of the climatic scenario and the large reduction of forest combined with a large expansion of impervious areas promoted by the socio-economic scenario S1. Less forest implies less evapotranspiration whereas more impervious areas imply shorter concentration time and drastically decreased infiltration capacity of the basin. All put together they have made this variable to grow at 1.8% per decade.

Frequency of high flows will grow in winter, Q_{11}, and conversely, will decline in summer , Q_{12}. Moreover, the urban sprawl scenario (S1) will exhibit the larger growth rates under the same climatic conditions. In other words, if the actual trends in land cover continue, regardless of the macroclimatic conditions, it is very likely (95% certainty) that the frequency of high flows in winter will be greater than that of the reference period. Although the average tendency of this variable is to decline in summer, sudden increases as compared to the reference period can be expected. The reasons for these developments in general are closely related to those already stated for variables Q_9 and Q_{10}, but they

are inversely related. In short, if a combination of factors cause the total duration of high flows to persist at higher values during longer times then the frequency of such flows will tend to decrease.

Finally, the total drought duration in summer, Q_{14}, will tend to increase faster in climatic scenario C1 than in C2. Land use/cover changes will have an impact on this variable but in a lesser degree as compared with those originated by a macroclimatic change.

8.2 Conclusions

The present study was based on three general objectives (see Section 1.4) aimed at investigating the impacts of climatic and land cover/use changes in a mesoscale catchment. Considering these objectives and the results that have been achieved and documented in previous chapters, it is possible to draw the following conclusions.

1. The key element in the analytical part of this study was the use of temporal and spatially distributed data available for 46 gauging stations at the Upper Neckar Catchment from 1961 to 1993. Based on this vast amount of information and with the help of sophisticated optimisation algorithms and nonparametric statistical techniques, it was possible to search and validate "very good" models that describe the state of the system at any point in time and for each spatial unit (i.e. a basin). These numerical relationships have allowed discriminating between the effects of climatic and land cover variability at mesoscale level. The quantifications of the magnitude of a given land cover change is straightforward.

2. Calibrated models for several runoff characteristics in winter and summer have shown that land cover variables are statistically significant (5% level) components of the water cycle at the mesoscale. However, the performance of models in winter is better than that in summer.

3. The integration of these hydrological models with a simple stochastic land use/cover change model has proved to be feasible and enlightening. Although the land use/cover model used in this study is quite simple, the results show that it is a promising planning tool since it allows testing the effects of several land use/cover and climatic scenarios on the hydrological cycle.

4. Further research, however, is still needed in order to improve the land use/cover change model so that it includes other time dependent factors which induce land use/cover changes.

5. Further steps should be carried out in order to promote the use and development of integrated planning tools as logical and systematic constructs that support planners' actions dealing with the complexities of natural systems and their entangled relationships with anthropogenic activities. By doing so, a step towards sustainability will be realized.

References

Abadie, J. and Carpentier, J. (1969): Generalization of the Wolfe Reduced Gradient Method to the Case of Nonlinear Constraints", in Optimization, R. Fletcher (ed.). London: Academic Press, 37-47.

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E., and Rasmussen, J. (1986): An introduction to the European Hydrological System- Systeme Hydrologique European, 'she', 2: structure of a physically-based distributed modelling system. Journal of Hydrology, 87:45-59.

Abdulla, F.A. and Lettenmaier, D.P. (1997): Development of regional parameter estimation equations for a macroscale hydrologic model. Elsevier: Journal of Hydrology 197:230-257.

Ahmed, S. and de Marsily, G. (1987): Comparation of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. Water Resources Research, 23(9):1717-1737.

Akaike, H. (1973): A new look at the statistical model identification. IEEE Transactions on Automatic Control. 19:716-723.

Allen, P.M. (1997): Cities and regions as self-organizing systems: models of complexity. Amsterdam: Gordon and Breach Science, 1997.

Anderson, M.G. and Burt, T.P. (1978): The role of topography in controlling throughflow generation. Earth Surface Processes, 3:331-344.

Avisar, R. and Verstraete, M.M. (1990): The representation of continental surface processes in atmospheric models. Reviews of Geophysics 28(1):35-52.

Ayros, E. (2000): Regionalisierung extremer Abflüsse auf der Grundlage statistischer Verfahren. Institut für Wasserbau, Universität Stuttgart. Helf 101.

Bárdossy, A. (1993): Stochastische Modelle zur Beschreibung der raum-zeitlichen Variabilität des Niederschlages. Heft 44, Institut für Hydrologie und Wasserwirtschaft der Universität Karlsruhe.

Bárdossy, A. (1997): Introduction to Geostatistics. Compendium, Institute for Hydraulic Engineering and Water Resource Management, University of Stuttgart.

Bárdossy, A. (1999): On CC-HYDRO. Impact of climate change on river basin hydrology under diffrenet climatic conditions. Final report ENV4-CT95-0133. Nachtnebel, P. (Editor).

Bárdossy, A., Bogardi, I., Kelly, W.E. (1990): "Kriging with imprecise (fuzzy) variograms", Mathematical Geology, v. 22(1):63-94.

Bárdossy, A., and Caspary, H.J. (1999): The Stochastic Downscaling Model "CP-precipitation and temperature" on ACCORD: Atmospheric Circulation Classification and Regional Downscaling, Reid P.A., Jones P.D., Davies T.D. (eds.). CRU, UK.
Online: http://www.cru.uea.ac.uk/cru/projects/accord/

Bárdossy, A. and Plate, E.J., (1992): Space-time model for daily rainfall using atmospheric circulation patterns. Water Resources Research, 28:1247-1259.

Bárdossy A., Samaniego, L. (2002): Fuzzy Rule-Based Classification of Remotely Sensed Imagery. IEEE Transactions on Geoscience and Remote Sensing. Vol.40, No.2, Feb. 2002.

Bedient, P.B. and Huber, W.C. (1992): Hydrology and Floodplain Analysis. Massachusetts: Addison-Wesley Publishing Company, 692 p.

Bell, E.J. (1974): Markov analysis of land use change-Application of stochastic processes to remotely sensed data, Socio-Economic Planning Sciences, 8(6):311–316.

Berenson, M.L. et al. (1983): Intermediate statistical methods and applications: a computer package approach. Englewood Cliffs, NJ : Prentice-Hall,. - XVIII, 579 p.

Bergström, S. and Forsman, A. (1973): Development of a conceptual deterministic rainfall-runoff model. Nordic Hydrology, Vol. 4(3):147-170.

Bergström, S. (1995): The HBV model. In: Computer models of watershed hydrology. Vijay P. Singh (Ed.). Colorado: Water Resources Publications, XIV, 1130 p.

Berry, M.W., Flamm, R.O., Hazen, B.C., and MacIntyre, R.L. (1995): The land-use change analysis system (LUCAS) for evaluating landscape management decisions. IEEE Computational Science and Engineering.

Beven, K.L. and Kirkby, M.J. (1979): A physically-based variable contribution area model of basin hydrology. Hydrological Sciences Bulletin, 24:43-69.

Beven, K.J. (1986): Hillslope Runoff Processes and Flood Frequency Characteristics. In A. D. Abrahams (ed.) Hillslope Processes, Boston: Allen and Unwin. 187-202.

Beven, K.J. (1989): Changing ideas in hydrology – the case of physically-based models. Journal of Hydrology 105:157-172.

Blyth, S. (1996): Out of line, Risk, 9(10):82-84.

Boer, G.J., Flato, G., Reader, M.C., and Ramsden, D. (2000): A transient climate change simulation with greenhouse gas and aerosol forcing: experimental design and comparison with the instrumental record for the 20th century. Clim. Dyn., 16:405-426.

Bosch, J.M. and Hewlett, J.A. (1982): A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. Journal of Hydrology, (55):3-23.

Bonell, M. (1993): Process in the understanding of runoff generation dynamics in forest. Journal of Hydrology 150:217-275. Elsevier Science Publishers.

Brown, D.G, Pijanowski, B.C., and Duh, J.D. (2000): Modelling the relationships between land-use and land-cover on private lands in the Upper Midwest, Journal of Environmental Management, 59:247–263.

Brown, D.G., Goovaerts, P., Burnicki, A. and Meng-Ying Li (2002): Stochastic Simulation of Land-Cover Change Using Geostatistics and Generalized Additive Models. Photogrammetric Engineering and Remote Sensing. MI 48109-1115 Vol. 68(10):1051–1061.

Bronstert, A. (1995): User Manual for the HILFLOW-3D Catchment Modelling System. (Physically Based and Distributed Modelling of Runoff Generation and Soil Moisture Dynamics for Micro-Catchments), Cooperative Research Centre for Catchment Hydrology. Working Document 95/4.

Bundesamt für Bauwesen und Raumordnung (2000): Raumordnungsbericht 2000, Band 7, Bonn.

Burrough, P.A. (1986): Principles of Geographical Information Systems for Land Resources Assessment. Oxford University Press, New York, 50 p.

Cada, G.F. and Hunsaker, C.T. (1990): Cumulative impacts of hydropower development: Reaching a watershed in impact assessment. The Environmental Professional 12(1):2-8.

Calder, I.R. (1993): Hydrologic effects of land-use change. In: Maidment D.R. (ed). Handbook of Hydrology, New York: McGraw-Hill. 13.1-13.50.

Canty, A.J. (1998): An S-Plus Library for Resampling Methods. Online: http://statwww.epfl.ch/davison/BMA/library.html.

Carlston, C.W. (1963): Drainage density and stream flow. United States Geological Survey. Professional Paper, 422-C.

Carlston, C.W. (1966): The effect of climate on drainage density and stream flow. Bulletin of the International Association of Scientific Hydrology, 11(3):62-69.

Casti, J. (1984): On the Theory of models and the modelling of natural phenomena in Recent Developments in Spatial Data Analysis, Methodology, Measurement, Models, Bahrenberg, G., Fischer, M., and Nijkamp, P. (eds.). Aldershot: Gower Publishing Co. Ltd.

Casti, J. (1990): Searching for Certainty: What Scientists Can Know About the Future. William Morrow and Company, Inc.: New York.

Chatfield, C. (1989): The analysis of time series: an introduction. London: Chapman and Hall, XII, 241 p.

Chilès, J.P. and Delfiner, P. (1999): Geostatistics: modeling spatial uncertainty. New York: Wiley - XI, 695 p.

Chow, V.T. (ed.) (1964): Handbook of Applied Hydrology: a compendium of water resources technology. New York: McGraw-Hill.

Clarke, R.T. (1994): Statistical modelling in hydrology. Chichester : Wiley,. - XII, 412 p.

Cross, G., and Jain, A. (1983): Markov Random Field Texture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-5, No.1, Jan. 1983.

Daniel, C. and Wood, F.S. (1980): Fitting equations to data: computer analysis of multifactor data, 2. ed.. New York: Wiley - XVIII, 458 p.

Davison, A.C., Hinkley, D.V. (1997): Bootstrap methods and their application. Cambridge: Cambridge University Press.

Dawdy, D.R., Lichty, R.W., and Bergmann, J.M. (1972): A rainfall-runoff simulation model for estimation of flood peaks for small drainage basins, United States Geological Survey Professional Paper 506-B, 28 p.

Demuth, S. and Külls, C. (1997): Probability Analysis and regional aspects of droughts. IAHS Publication. 240:97-104.

Deutsch, C. (2001): Principles of Monte Carlo Simulations. Centre for Computational Geostatistics, University of Alberta. Online: http://www.ualberta.ca/~cdeutsch.

Deutscher Wetterdienst (1960-1995): Meteorologisch-statistische Jahrbücher von 1960 bis 1995 für verschiedene Klima- und Niederschlagsstationen. Offenbach.

Dickert, T.G., and Tuttle, A.E. (1985): Cumulative impact assessment in environmental planning: A coastal wetland watershed example. Environmental Impact Assessment Review 5:37-64.

Dingman, S.L. (1994): Physical Hydrology. New York: Macmillan Publishing Company, 575 p.

Dooge, J.G. (1988): Hydrology in Perspective. Hydrological Sciences Journal 33:61-85.

Dooge, J.G. (1992): Hydrologic models and climate change. Journal of Geophysical Research. 97(D3):2677-2686.

Dracup, J.A., and Kahya, E. (1994): The Relationships Between U.S. Streamflow and La Niña Events, Water Resources Research, 30(7):2133-2141.

Draper, N.R. and Smith, H. (1981): Applied regression analysis (2. ed.). New York: Wiley. XIV, 709 p.

Duckstein, L., Bardossy, A., and Bogardi, A. (1993): Linkage Between the Occurrence of Daily Atmospheric Circulation Patterns and Floods: an Arizona Case Study. J. Hydrology., 143:413-428.

Duckstein, L., Plate, E.J. (Eds.) (1987): Engineering Reliability and Risk in Water Resources. Dordrecht: Martinus Nijhoff Publishers. NATO ASI Series.

Dudewicz, E.J. (1992): The Generalized Bootstrap. In Bootstrapping and related techniques (proceedings of an international conference, held in Trier, FRG, June 4 - 8, 1990): Joeckel, K.-H (Ed.). Berlin; Heidelberg: Springer. VIII, 245 p. (Lecture notes in economics and mathematical systems; 376 p.).

Eeles, C.W. and Blackie, J.R. (1993): Land-use changes in the Balquhidder catchments simulated by a daily streamflow model. Journal of Hydrology, 145:315-336.

Efron, B. (1982): The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, Pa.: Society for Industrial and Applied Mathematics, VII, 92 p. (Regional conference series in applied mathematics; 38).

Enting, I.E., Wigley, T.M.L., and Heimann. M. (1994): Future Emissions and Concentrations of Carbon Dioxide: Key Ocean/Atmosphere/Land Analyses. CSIRO Division of Atmospheric Research Technical Paper No. 31.

Eurostat (1996): GDP per head in the European' Union's Richest and Poorest Regions.

Ezekiel, M. (1930): Methods of Correlation Analysis, Wiley, New York.

Fisher, R.W. (1996): Future Energy Use. Future Research Quarterly, 31:43-47.

Flamm, R.O. and Turner, M.G. (1994): Alternative model formulations for a stochastic simulation of landscape change. Landscape Ecology, 9(1):37-46.

Foley, J.A., Levis, S., Prentice, I.C. et al (1998): Coupling dynamic models of climate and vegetation, Global Change Biology, 4:561-579.

Forrester, J.W. (1969): Urban dynamics. Cambridge: M.I.T. Press,. - XIII, 285 p.

Fowler, H.W. and Fowler, F.G. (eds.) (1995): The Concise Oxford Dictionary of current English First (9. ed). Oxford: Clarendon Press.

Frank, E.C., and Lee, R. (1966): Potential solar beam irradiation on slopes. U.S. Department of Agriculture, Forest Service Research Paper, RM-18, 116 p.

Freeze, R.A. (1972): The role subsurface flow in generating surface runoff. Water Resources Research. 8(5):1271-1283.

Gentleman, W.M. (1974): Basic procedures for large, sparse or weighted linear least squares problems. Applied Statistics, 23:448-454.

Geyer, O.F. and Gwinner, M.P. (1991): Geologie von Baden-Württemberg. Stuttgart.

Gilchrist, W. (1984): Statistical modelling. Chichester: Wiley.

Good, P. (2000): Permutation tests: a practical guide to resampling methods for testing hypotheses (2. ed.). New York: Springer, XVI, 270 p.

Gregory, J.M., Mitchell, J.F.B., and Brady, A.J. (1997): Summer drought in northern midlatitudes in a time-dependent CO2 climate experiment. J. Climate, 10:662-686.

Hack, J.T. (1957): Studies of longitudinal stream profiles in Virginia and Maryland. U.S.G.S. Professional Paper 294-B.

Haggett, P., Chorley, R.J. (1969): Network analysis in geography. London : Arnold,. XII, 348 p.

Haldorsen, H. and Damsieth, E. (1990): Stochastic Modelling. SPE 20321.

Hansen, J. and Lebedeff. S. (1988): Global surface air temperatures: Update through 1987. Geophysical Research Letters 15:323-326.

Hammersley, J.M. and Handscomb, D.C. (1964): Monte Carlo methods. London: Methuen. 178 p.

Hammond, A. (1996): Which World? Island Press, Washington, D.C., 306 p.

Hamon, W.R. (1961): Estimating potential evapotranspiration. Proceedings of the American Society of Civil Engineers, Journal of Hydraulic Division, 87(HY3):107-120.

Henderson-Sellers, A. (1990): Predicting generalized ecosystem groups with the NCAR GCM: First steps towards an interactive biosphere. J. Climate, 3:917-940.

Henderson-Sellers, A. (1992): Assessing the sensitivity of a land-surface scheme to parameters used in tropical-deforestation experiments, Quart. Quart. J. Roy. Meteor. Soc., 118:1100-1116.

Henderson-Sellers A. (1993): Continental vegetation as a dynamic component of global climate model: a preliminary assessment. Climatic Change 23:337-378.

Hennessy, K.J., Gregory, J.M., and Mitchell, J.F.B. (1997): Changes in daily precipitation under enhanced greenhouse conditions. Clim. Dyn., 13:667-680.

Hess, P. und Brezowsky, H. (1969): Katalog der Großwetterlagen Europas. Berichte des Deutschen Wetterdienst 113(15), 2 neu bearbeitete und ergänzte Aufl., Offenbach a. Main, Selbstverlag des Deutschen Wetterdienst.

Hewlett, J.D. and Hibbert, A.R. (1967): Factors affecting the response of small watersheds to precipitation in humid areas. In W.E. Sopper and H.W. Lull (Eds.), International Symposium on Forest Hydrology. The Pennsylvania State University. Oxford: Pergamom.

Horton, R.E. (1945): Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology. Bulletin of the Geological Society of America, 56:275-370.

Houghton, J.T., Jenkins, G.J., and Ephraums, J.J. (Ed.) (1990): Climate Change - The IPCC Scientific Assessment. Cambridge, UK: Cambridge University Press. 365 p.

Houghton, J.T., Callander, B.A., and Varney, S.K. (Ed.) (1992): Climate Change 1992 - The Supplementary Report to the IPCC Scientific Assessment. Cambridge, UK: Cambridge University Press. 200 p.

Houghton, J.T. et al. (eds.) (2001): Climate Change 2001: The Scientific Basis. Cambridge Univ. Press. Online: http://www.grida.no/climate/ipcc_tar/wg1/index.htm.

Huber, P.J. (1981): Robust Statistics. New York: Wiley.

IMSL (1997): Fortran Subroutines for Mathematical Applications. Visual Numerics, Inc.

Institut für Photogrammetrie und Fernerkundung, Universität Karlsruhe (1995): Erstellung einer Landnutzungskarte des Landes Baden-Württemberg als Eingabedatensatz für ökologishe Bewertungen auf der Grundlage von LANDSAT TM Satellitenbilddaten.

Jenson, S.K. and Domingue, J.O. (1988): Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis, Photogrammetric Engineering and Remote Sensing. 54(11)1593-1600.

Johns, T.C. (1996): A description of the Second Hadley Centre Coupled Model (HadCM2). Climate Research Technical Note 71, Hadley Centre, United Kingdom Meteorological Office, Bracknell Berkshire RG12 2SY, United Kingdom, 19 p.

Jolliffe, I.T. (1986): Principal component analysis. New York: Springer. University of Geneva.

Jones, P.D. (1994): Recent warming in global temperature series. Geophysical Research letters, 21:1149-1152.

Jones, J.A. (1997): Global hydrology: processes, resources and environmental management. Harlow: Longman.

Jordan et al. (2000): On Assessment of Potential Effects and Adaptations for Climate Change in Europe: The Europe ACACIA Project. Parry, M.L. (Editor). Jackson Environment Institute, University of East Anglia, Norwich, UK, , 320 p.

Karl, T. and Trenberth, K. (1999): The Human Impact on Climate, Scientific American. December,1999, in Earth from the inside out (2000).

Karlqvist, A. (1978): Spatial interaction theory and planning models. Amsterdam: North-Holland Publ. Co., 388 p. (Studies in regional science and urban economics ; 3).

Kirby, C. Newson, M.D., and Gilman, K. (1991): Plynlimon Research: the first two decades. Institute of Hydrology, Wallingford, Rep. 109 p.

Kleeberg, H.B. and Cemus, J. (1992): Regionalisierung hydrologischer Daten. In: Kleeberg, H.B. (Ed.), Regionalisierung in der Hydrologie, Weinheim, 1-15.

van Laarhoven, P.J.M. and Aarts, E.H.L. (1992): Simulated annealing: theory and applications Dordrecht : Kluwer, 1992. XI, 187 p.

Laenen, A. (1980): Storm runoff as related to urbanization in the Portland, Oregon- Vancouver, Washington Area: United States Geological Survey Water Resources Investigations, Open File Report 80-689, 71 p.

Landesanstalt für Umweltschutz Baden-Württemberg (1960-1993): Hydrologisch-statistische Jahrbücher von 1960 bis 1993 für verschiedene Abflusspegel. Karlsruhe.

Landesanstalt für Umweltschutz Baden-Württemberg (1999): Daten des Landschaftsrahmen-programms Baden-Württemberg. (LfU-IER-ILPÖ). Karlsruhe.

Lane, D.M. (2001): Online: http://davidmlane.com/hyperstat/index.html.

Launer, R.L. and Wilkinson, G.N. (eds.) (1979): Robustness in Statistics. New York: Academic Press.

Law, F. (1956): The effects of afforestation upon water yields of catchment areas. Journal of British Waterworks Association 38:484-494.

Leavesley, G.H., Lichty, R.W., Troutman, B.M., and Saindon, L.G. (1983): Precipitation-Runoff Modeling System: User's Manual, U.S. Geological Survey Water- Resources Investigations 83-4238. Denver - Colorado, 208 p.

Leavesley, G.H. and Stannard, L.G. (1995): The Precipitation-Runoff Modeling System - PRMS, in Computer Models of Watershed Hydrology, V.P. Singh (ed.), Water Resources Publications, Highlands Ranch, Colorado, 281 - 310.

Leopold, L.B. and Miller, J.P. (1956): Ephemeral streams: hydraulic factors and their relations to the drainage net. U.S.G.S. Professional Paper 282-A.

Lettenmaier, D.P. and Wood, E.F. (1993): Hydrologic Forecasting. In Handbook of Hydrology, Chapter 26. Maidment, D.R. (ed. in chief). New York: McGraw-Hill.

Lindsey, J. K. (1999): Applying generalized linear models. Springer, New York, 256 p.

Linsley, R.K., Kohler, M.A., and Paulhus, J.L.H. (1982): Hydrology for Engineers, 3d ed., New York: McGraw-Hill.

Mann, M.E., Bradley, R.S., and Hughes, M.K. (1999) (1): Northern Hemisphere Temperatures During the Past Millennium: Inferences, Uncertainties, and Limitations. Geophysical Research Letters 26(6):759.

Mann, M.E., Bradley, R.S., and Hughes, M.K. (1999) (2): Northern Hemisphere Temperature Reconstruction for the Past Millennium, IGBP PAGES/World Data Center-A for Paleoclimatology Data Contribution Series # 1999-014. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA.
Online: ftp://ftp.ngdc.noaa.gov/paleo/contributions_by_author/mann1999/

Matheron, G. (1970): The theory of regionalized variables and its applications. Fascicule 5, Les Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontaineblueau.

McArthur, T. (ed.) (1992): The Oxford Companion to the English Language. Oxford: Oxford University Press, 1996. - XVIII, 1053 p.

McCarthy, J.J. (2001): Climate Change 2001: Impacts, Adaptation and Vulnerability - Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). Online: http://www.grida.no/climate/ipcc_tar/wg2/index.htm.

McCuen, R.H. (1973): The role of sensitivity analysis in hydrologic modeling, Journal of hydrology, 18:37-53.

McCuen, R.H. (1998): Hydrologic Analysis and Design. New Jersey: Prentice Hall, 814 p.

McGuffie, K., Henderson-Sellers, A., Holbrook, N., Kothavala, Z., Balachova O., and Hoestra, J. (1999): Assessing simulations of daily temperature and precipitation variability with global climate models for present and enhanced greenhouse climates. Int. J. Climatol., 19:1-26.

McNeill, J. (rapporteur) et al. (1994): Toward a Typology and Regionalization of Land Cover and Land Use Change. In: Meyer W. and B. L. Turner II (eds.). Changes in Land Use and Land Cover: A Global Perspective. Cambridge: University Press.

Mein, R.G. and Brown, B.M. (1978): Sensitivity of optimised parameters in watershed models. Water Resources Research, 14(2):299-303.

Melton, M.A. (1958): An analysis of relations among elements of climate, surface properties, and geomorphology. Office of Naval Research, Geographic Branch, Project NR 389-042, Technical Report, II.

Melloh, A.R (1999): A Synopsis and Comparison of Selected Snowmelt Algorithms. U.S. Army Cold Regions Research and Engineering Laboratory: CRREL Report 99-8. Online: http://www.crrel.usace.army.mil/techpub/CRREL_Reports/reports/CR99_08.pdf .

Metropolis, N., Rosenbluth, A., Teller, M., Teller, A., and Teller, E. (1953): Equation of state calculations by fast computing machines. Journal Chem. Phys. 21:1087.

Meyer, W. and Turner, B.L.II (eds.) (1994): Changes in Land Use and Land Cover: A Global Perspective. Cambridge: University Press.

Monteith, J.L. (1965): Evaporation and the environment. Symposium of the Society of Exploratory Biology 19:205-234.

Montgomery, D.C. and Peck, E.A. (1982): Introduction to linear regression analysis. New York: Wiley, XIII, 504 p.

Moore, R.D. (1997): Storage-outflow modeling of streamflow recessions, with application to a shallow-soil forested catchment. Journal of Hydrology 198:260-270.

Moore, I., Grayson, R., and Ladson, A. (1991): Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological Processes, 5:3-30.

Muller, M.R. and Middleton, J. (1994): A Markov model of land-use change dynamics in the Niagara Region, Ontario, Canada, Land- Winkler, G., 1995. Landscape Ecology, 9(2):151–157.

Nandakumar, N., Mein, R.G. (1997): Uncertainty in rainfall-runoff model simulations and the implications for predicting the hydrologic effects of land-use change. Journal of Hydrology 192:211-232.

Neave, H.R. and Worthington, P.L. (1988): Distribution-free tests. First Edition. London: Unwin Hyman.

New, M., Hulme, M., and Jones, P.D. (2000): Representing twentieth century space-time climate variability. Part 2: development of 1901-96 monthly grids of terrestrial surface climate. J. Climate, 13(13):2217-2238.

Nolsøe, K., Nielseny, J., and Madsenz, H. (2000): Optimal Weights in Prediction Error and Weighted Least Squares Methods. Online: http://www.imm.dtu.dk/documents/ftp/tr00/tr08_00.pdf

O'Loughlin, E.M. (1981): Saturated regions in catchments and their relations to soil and topographic properties, Journal of Hydrology, (53):229-226.

O'loughlin, E.M., Short, D.L. and Dawes, W.R. (1989): Modelling the hydrological response of catchments to land use change. In: Proceedings of the 1989 Australian Institution of Engineers Hydrology and Water Resources Symposium. Nat. Conf. Publ., 89(19):335-340.

Penman, H.L. (1948): Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society, London, A193, 120-146.

Perez-Trejo, F. (1993): Landscape response units: process-based self-organising systems. In, Haines-Young, R., Green, D.R. and Cousins, S. (eds), Landscape Ecology and Geographic Information Systems, Taylor and Francis, London, UK., 87-98.

Piechota, T.C. and Dracup, J.A. (1996): Drought and Regional Hydrologic Variations in the United States: Associations with the El Niño-Southern Oscillation. Water Resources Research, 32(5):1359-1373.

Pirt, J. (1983): Low flow estimation in ungauged catchments. Occasional paper 6, Department of Geography, University of Technology, Loughborough.

Potter, K.W. (1991): Hydrological impacts of changing land management practices in a moderate-sized agricultural catchment: Water Resources Research, 27(5):845-855.

Quenouille, M. (1949): Approximate tests of correlation in time series. Journal of the Royal Statistical Society, 11B:18-84.

Raudkivi, A.J. (1979): Hydrology : an advanced introduction to hydrological processes and modelling. 1st ed. Oxford, Kronberg-Taunus: Pergamon Press,. IX, 479 p.

Reddmont, T. and Koch, R. (1991): Surface climate and stream flow variability in the Western United States and their relationship to large-scale circulation indices. Water Resources Research 17:2381-2399.

Refsgaard, J.C., Alley, W.M., and Vuglinsky, V.S. (1989): Methods for distinguishing between man's influence and climatic effects on the hydrological cycle (IHP-III Project 6.3. UNESCO, 1989).

Refsgaard, J.C. (1997): Parameterisation, calibration and validation of distributed hydrological models. Elsevier, Journal of Hydrology 198:69-97.

Reimold, R.J. (ed.) (1998): Watershed Management: Practice, Policies, and Coordination. New York: McGraw-Hill.

Ripley, B.D. (1987): Stochastic simulation. New York: Wiley. XI, 237 p.

Rodriguez-Iturbe, I. (1969). Estimations of Statistical parameters for Annual River Flows. Water Resources Research, 5(6):1418-1426.

Rousseeuw, P.J. and Leroy A.M. (1987): Robust regression and outlier detection. New York : Wiley. XIV, 329 p.

Rowland, F.S. (2000): Memorandum to the President, subject: Climate change and its consequences. In: U.S. Policy and the Global Environment. Kennedy, D. and Riggs, J. (Eds.). Aspen Institute: Colorado.

Salati, E., Dall'Olio, A., Matsui, E., Gat, J.R. (1979): Recycling of water in the Amazon basin: an isotopic study. Water Resource Research, 15(5):1250-1258.

Samaniego, L. (1997): Watershed Management: a case study in Ecuador (Master Thesis). University of Stuttgart. (Unpublished).

Sartor, J. (1998): Mögliche Einflüsse der Bebauung auf den Hochwasserabfluß. Die Wasserwirtshaft, 88(3):122-126.

Savenije, H.H.G. (1995): New definitions for moisture recycling and the relationships with land-use changes in the Sahel. Elsevier, Journal of Hydrology 167:57-78.

Schumann, A.H. and Schultz, G.A. (2000): Detection of Land Cover Change Tendencies and their Effect on Water Management. Remote Sensing in Hydrology and Water Management. Heidelberg. 419-442.

Schultz, G.A. (ed.) (2000): Remote sensing in hydrology and water management. Berlin: Springer, XX, 483 p.

Shaw, J. (1997): Beyond VaR and stress testing in VaR: Understanding and Applying Value at Risk. Risk Publications, London, 211-224.

Shorthouse, C. and Arnell, N.W. (1997): Spatial and temporal variability in European river flows and the North Atlantic Oscillation. FRIEND'97: International Association of Hydrological Science Publications, 246, 77–85.

Shukla, J., Nobre, C. and Sellers, P. (1990): Amazon deforestation and climate change. Science 247:1322-1325.

Simonoff, J. S. (1996): Smoothing Methods in Statistics. Springer, New York.

Singh. V.P. (ed.) (1995): Computer models of watershed hydrology. Water Resources Publications. Colorado.

Skole, D.L. (1994): Data on Global Land Cover Change: Acquisition, Assessment, and Analysis. In: Meyer W. and B. L. Turner II (eds.). Changes in Land Use and Land Cover: A Global Perspective. Cambridge: University Press.

Statistisches Landesamt Baden-Württemberg (1984-1993): Ergebnisse der Flächenerhebung nach Gemeinden und Gemarkungen, Bodenfläche nach Art der tatsächlichen Nutzung (1981, 1985, 1989, 1993 und 1997).

Statistisches Landesamt Baden-Württemberg (1996): Arbeitslosenquote in der Region Stuttgart, Stuttgart. Online: http://www.statistik.baden-wuerttemberg.de.

Stahl, K. and Demuth, S. (1999): Linking streamflow drought to the occurrence of atmospheric circulation patterns. Hydrological Sciences Journal, 44(3):467-482.

Steiss, A.W. (1974): Models for the analysis and planning of urban systems. Lexington, Mass.: Lexington Books. - XI, 352 p.

Stern, P.C., Young, O.R., and Druckman D. (eds.) (1992): Global environmental change: Understanding the human dimensions. Washington, DC: National Academy Press.

von Storch, H., Bruce H., and Mearns L. (1999): Review of Empirical Downscaling Techniques. GKSS Research Center, Institute for Hydrophysics, Germany. Online: http://www.nilu.no/regclim/rapport_4/presentation02/presentation02.htm.

Swift, L.W. (1976): Algorithm for solar radiation on mountain slopes. Water Resources Research,12(1):108-112.

Tarboton D.G., Bras, R.L., Rodriguez-Iturbe, I. (1991): On the Extraction of Channel Networks from Digital Elevation Data, Hydrological Processes. 5:81-100.

The American Heritage Dictionary of the English Language, 4th ed. (2000): Boston: Houghton Mifflin.

Treuner, P. (1995): Introduction to Scenario Methods in Regional Development Planning, in: Methods and Experience of Regional Development Scenarios. Shanghai.

Tukey, J. (1958): Bias and confidence in not quite large samples. Abstract, Annals of Mathematical Statistics, 29:614.

Turner, B.L. and Meyer, W.B. (1991): Land use and land cover in global environmental change: considerations for study. International Social Sciences Journal 130:669–667.

Turner, M.G., (1987): Spatial simulation of landscape changes in Georgia: A comparison of three transition models, Landscape Ecology, York, N.Y., 537(1):29–36.

Turner, B.L., Moss, R.H., and Skole, D.L. (Eds.) (1993): Relating land use and global land-cover change: A proposal for an IGBP-HDP core project. Report from the IGBP-HDP Working Group on Land-Use/Land-Cover Change. Joint publication of the International Geosphere-Biosphere Programme (Report No. 24) and the Human Dimensions of Global Environmental Change Programme (Report No. 5). Stockholm: Royal Swedish Academy of Sciences.

USDA-SCS (1985): National Engineering Handbook, Section 4 - Hydrology. Washington, D.C.: USDA-SCS.

USGS (2002): Online: http://water.usgs.gov/cgi-bin/man_wrdapp?prms

Venables, W.N. and Ripley, B.D. (1997): Modern applied statistics with S-Plus. 2. ed. New York: Springer.

Vertessy, R.A., Hatton, T.J., O'Shaughnessy, P.J., and Jayasuriya, M.D.A. (1993): Predicting Water Yield from a Mountain Ash Forest Catchment Using a Terrain Analysis-Based Catchment Model. Journal of Hydrology, 150:665-700.

Vinnikov, K.Ya., Groisman, P.Ya., and Lugina. K.M. (1990): Empirical data on contemporary global climate changes (temperature and precipitation). Journal of Climate 3:662-677.

Walker, P.M. (ed.) (1999) Dictionary of Science and Technology. Edinburgh: Chambers Harrap Publishers Ltd.

Ward, R.C. and Robinson, M. (2000): Principles of hydrology (4th ed.). London: McGraw-Hill, - XIV, 450 p.

WCED (World Commission on Environment and Development), (1987): Our Common Future. London: Oxford University Press.

Wigley, T.M.L. and Raper, S.C.B. (1992): Implications for climate and sea-level of the revised IPCC emissions scenarios. Nature 357(6376):293-300.

Wilby, R.L. (ed) (1997): Contemporary hydrology: towards holistic environmental science. Chichester: Wiley, 354 p.

Wilks, D. (1995): Statistical Methods in the Atmospheric Sciences: *An Introduction*. San Diego: Academic Press, 465 p.

Wischmeier, W.H. and Smith, D.D. (1978): Agriculture Handbook No. 537. Predicting rainfall erosion losses - A guide to conservation planning with the Universal Soil Loss Equation (USLE) -. United States Department of Agriculture, Springfield, USA.

Wolfe, P. (1963): Methods of Nonlinear Programming on Recent Advances in Mathematical Programming, Graves, R. L., and Wolfe, P. (Eds.). New York: McGaw-Hill, 67-86.

Wolock, D.M. (1993): Simulating the variable-source-area concept of watershed hydrology with TOPMODEL: USGS Water-Resources Investigation Report 93-4124, 33 p.

Wonnacott, T.H. and Wonnacott, R.J. (1990): Introductory statistics. 5th ed. New York: Wiley, XVI, 711 p.

Zwiers, F.W. (2002): Climate change. The 20-year forecast. Nature, 416:690-691.

Appendix 1

Correspondence table showing the composition of the spatial units (i.e. basins) based on subunits called sub-catchments derived from a DEM (30×30 m, LfU) (see Figure 2.5).

Spatial Unit	Composed of Sub-catchments
1	1 46
2	2
3	3 6 40
4	4 30
5	5 7 9
6	6 27 25 39 15
7	7
8	8 17 28
9	9 20 12
10	10 35 41
11	11
12	12 33 22 37
13	13
14	14
15	15 26 14 16
16	16
17	17
18	18
19	19 29
20	20
21	21
22	22 23
23	23 21
24	24 44 2 34 11 10 5
25	25
26	26 32
27	27
28	28
29	29
30	30
31	31
32	32
33	33
34	34 31
35	35 38
36	36
37	37
38	38 18
39	39
40	40 42 8
41	41
42	42 19 24 1 4
43	43
44	44 36 45
45	45
46	46

Appendix 2

Basic information of the spatial units and their gauging stations located within the Study Area.

Spatial Unit	Gauging Station Name	River Name	Area [km^2]	nobs [year]	Outliers [%]
1	Wannweil-Bahn	Echaz	161.31	32	16.3
2	Pfäffingen	Ammer	133.52	33	18.5
3	Plochingen	Neckar	4002.00	33	12.1
4	Riederich	Erms	161.25	33	18.2
5	Horb	Neckar	1119.74	33	15.5
6	Plochingen	Fils	701.61	33	11.5
7	Hopfau-2	Glatt	202.34	32	16.9
8	Wendlingen	Lauter	190.00	33	13.6
9	Oberndorf	Neckar	694.71	33	12.7
10	Bad Imnau	Eyach	331.08	33	10.3
11	Rangendingen Wehr	Starzel	122.50	33	9.7
12	Rottweil-Gaswerk	Neckar	454.78	33	13.6
13	Denkendorf-Sägewerk	Körsch	126.29	33	14.8
14	Geislingen	Eyb	123.20	33	16.1
15	Süßen	Fils	357.00	33	12.7
16	Süßen	Lauter	68.20	33	13.6
17	Kirchheim	Lindach	92.10	32	12.5
18	Frommern	Eyach	72.90	28	7.1
19	Oberensingen-2	Aich	178.00	32	10.0
20	Epfendorf	Schlich	106.00	32	10.6
21	Horgen	Fischbach	120.40	22	10.9
22	Bühlingen	Eschach	218.52	21	6.7
23	Horgen-Kläranlage	Eschach	208.00	33	20.0
24	Kirchentellinsfurt	Neckar	2321.83	33	32.4
25	Baiereck-Typ	Herrenbach	4.50	31	13.9
26	Geislingen-Brücke	Fils	137.60	32	12.8
27	Reichenbach	Lützelbach	14.57	22	15.0
28	Unterlenningen	Lauter	58.80	32	22.2
29	Neuenhaus-Brücke	Schaich	38.20	25	7.5
30	Bad Urach-Kurgebiet	Erms	108.30	32	19.4
31	Dußlingen-Pulvermühle	Wiesaz	38.10	30	9.3
32	Wiesensteig-Ort	Fils	30.30	33	17.3
33	Göllsdorf	Prim	124.90	31	4.8
34	Tübingen-Bläsiberg	Steinlach	139.00	31	10.0
35	Owingen-Ort	Eyach	206.20	32	8.4
36	Bebenhausen	Goldersbach	34.32	29	12.1
37	Deißlingen	Neckar	37.90	26	10.4
38	Balingen	Eyach	122.90	23	4.8
39	Eislingen	Krumm	25.80	19	3.2
40	Wendlingen Kläranlage	Neckar	3270.10	22	15.0
41	Gruol	Stunzach	75.80	19	14.2
42	Wendlingen Wehr	Neckar	3075.00	16	6.9
43	Schömberg-Zulauf	Schlichem	28.40	13	2.3
44	Tübingen-Lustnau	Goldersbach	68.18	9	4.4
45	Tübingen-Lustnau	Kirnbach	8.90	13	16.9
46	Wannweil	Firstbach	6.70	11	17.3

Appendix 3

Sample statistics of the explanatory variables employed in this Study.

Variable Name	Unit	nobs	Min	Max	Mean	Median	Std	C_v [%]
x_1	[km²]	46	4.50	4002.00	433.08	124.05	895.75	206.8
x_2	[°]	46	3.02	11.07	6.97	7.27	2.07	29.7
x_3	[°]	46	2.20	10.01	6.00	6.00	1.53	25.5
x_4	[°]	46	4.92	9.14	6.68	6.73	1.09	16.3
x_5	[°]	46	3.24	8.86	6.13	6.10	1.36	22.2
x_6	[°]	46	1.09	3.73	2.36	2.29	0.68	28.7
x_7	[°]	46	3.37	15.93	7.63	7.09	2.60	34.0
x_8	[1/km]	46	1.75	2.53	2.16	2.18	0.16	7.4
x_9	[-]	46	1.02	7.06	2.56	2.23	1.10	42.7
x_{10}	[-]	46	5.49	18.20	12.34	12.12	2.42	19.6
x_{11}	[-]	46	5.13	30.75	18.66	17.79	4.33	23.2
x_{12}	[m]	46	386.14	818.34	595.13	609.72	114.40	19.2
x_{13}	[m]	46	144.00	768.00	434.22	447.00	161.57	37.2
x_{14}	[-]	46	39.82	65.48	56.78	57.08	5.74	10.1
x_{15}	[mm]	46	66.88	136.19	98.58	97.28	14.64	14.8
x_{16}	[-]	46	0.00	99.26	20.62	8.03	28.40	137.7
x_{17}	[-]	184	8.50	98.69	42.33	38.23	18.54	43.8
x_{18}	[-]	184	0.00	31.05	4.89	3.44	4.89	100.0
x_{19}	[-]	184	1.26	87.91	52.78	56.19	17.12	32.4
x_{20}	[mm]	1518	487.68	1680.32	910.44	892.52	182.94	20.1
x_{21}	[mm]	1518	170.68	1052.33	399.67	388.77	119.42	29.9
x_{22}	[mm]	1518	237.96	910.88	510.78	510.13	103.72	20.3
x_{23}	[mm]	1518	1.46	4.60	2.52	2.48	0.50	19.7
x_{24}	[mm]	1518	1.01	5.57	2.25	2.20	0.62	27.8
x_{25}	[mm]	1518	1.45	4.85	2.79	2.76	0.59	21.3
x_{27}	[mm]	1518	41.25	182.84	79.84	77.82	19.00	23.8
x_{28}	[mm]	1518	38.40	255.52	90.84	86.93	27.42	30.2
x_{29}	[mm]	1518	29.22	138.19	61.63	60.35	17.36	28.2

Appendix 3

(Continuation). Sample statistics of the explanatory variables employed in this Study.

Variable Name	Unit	nobs	Min	Max	Mean	Median	Std	C_v [%]
x_{30}	[°C]	1518	-8.10	4.20	-0.92	-0.70	2.65	289.3
x_{31}	[°C]	1518	13.10	22.20	16.82	16.60	1.56	9.3
x_{32}	[°C]	1518	0.60	12.20	6.20	6.20	2.21	35.7
x_{33}	[°C]	1518	19.70	29.60	23.30	22.90	1.73	7.4
x_{35}	[K]	1518	280.71	296.51	291.84	291.95	2.02	0.6
x_{36}	[K]	1518	279.06	286.41	282.26	282.14	1.51	0.5
x_{37}	[K]	1518	289.63	296.51	292.09	291.97	1.24	0.4
x_{38}	[day]	860	0	79	22.00	17.00	19.40	88.3
x_{39}	[day]	33	37	127	80.45	80.00	19.81	24.6
x_{40}	[day]	976	0	30	4.68	3.00	4.71	100.6
x_{41}	[day]	1239	0	60	11.30	10.00	8.14	71.8

Note: *Variables x_{26} and x_{34} are not included in this table because they depend on the time point where the variable is to be evaluated.*

Appendix 4

Sample statistics of the explained variables modelled in this Study.

Variable Name	Unit	nobs	Min	Max	Mean	Median	Std	C_v [%]
Q_1	[mm]	1244	69.40	1 206.39	405.38	379.35	182.57	45.0
Q_2	[mm]	1254	31.89	793.67	248.90	234.74	117.35	47.1
Q_3	[mm]	1255	20.68	639.58	156.29	136.57	87.43	55.9
Q_4	[mm]	1312	0.67	59.48	8.80	7.58	5.44	61.8
Q_5	[mm]	1318	0.47	82.85	6.81	4.91	6.20	91.0
Q_6	[mm]	1307	1.60	119.94	30.68	27.83	16.13	52.6
Q_7	[mm]	1239	1.50	446.52	74.21	55.79	63.59	85.7
Q_8	[mm]	976	1.38	363.11	34.23	22.91	37.72	110.2
Q_9	[day]	1312	0	66	12.80	11.00	9.79	76.6
Q_{10}	[day]	1318	0	46	4.79	3.00	6.28	131.0
Q_{11}	[1/year]	1247	1	15	4.36	4.00	2.36	54.1
Q_{12}	[1/year]	977	1	16	3.02	2.00	2.31	76.6
Q_{13}	[day]	860	1	135	12.58	10.00	12.08	96.1
Q_{14}	[day]	860	1	145	31.76	25.00	27.2	85.6
Q_{15}	[mm/year]	834	0.14	304.91	19.97	15.82	18.95	94.9
Q_{16}	[mm]	834	0.00	33.63	1.57	0.70	2.37	151.3

Appendix 5

Existing valid runoff data for the gauging stations located within the Study Area from 1961 to 1993 (Gauging station No. ≡ Spatial unit No.).

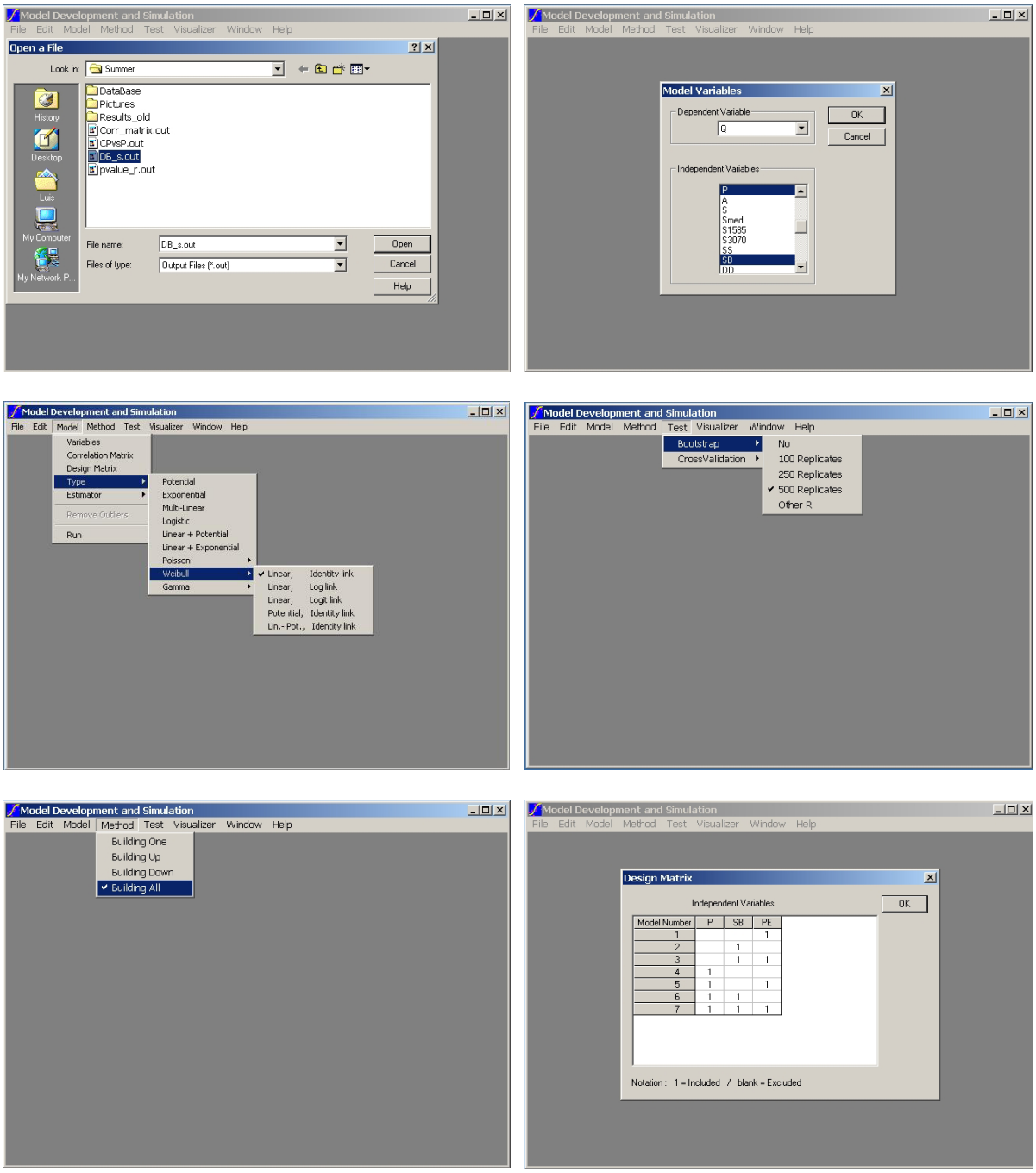
Station No.	Year (19xx)																															
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93
1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
3	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
4	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
6	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
7	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■
8	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
9	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
10	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
11	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
12	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
13	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
14	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
15	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
16	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
17	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□
18	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□
19	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■
20	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
21	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□	□	□
22	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□	□	□
23	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
24	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
25	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■
26	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
27	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□	□	□
28	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
29	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	□	□	■	■	■	■	■	□	□	□	□	□
30	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■
31	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
32	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
33	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■
34	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■	■
35	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■	■	■	■	■	■	■	■	■	■	■
36	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
37	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
38	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
39	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□	□	□	□
40	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
41	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
42	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	□	□
43	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
44	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	□	■	■	■	■	■	□	□	□	□	□	□	□
45	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■
46	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■

Legend:

■ Valid data exists for a given gauging station and a water year.

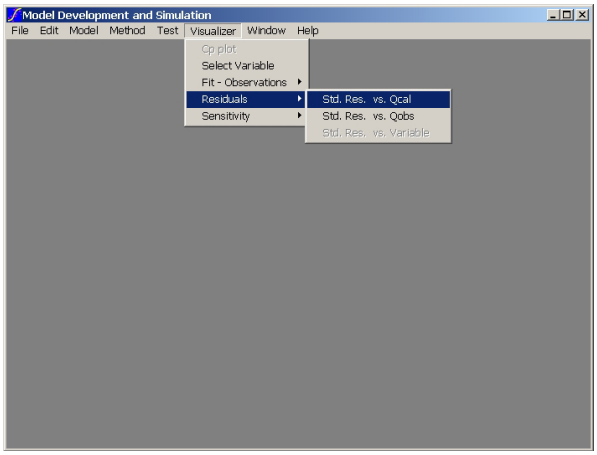
Appendix 6

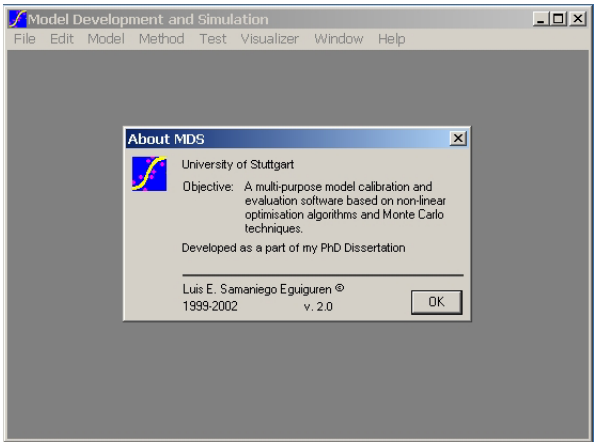
Sequence of screen captures showing the user interface of the MDS program. From top-left to bottom-right: 1) reading a database; 2) selecting variables; 3) selecting the model type; 4) selecting the number of replicates for the permutation test; 5) selecting the method for searching the best model; 6) displaying the design matrix of all possible combinations of variables.

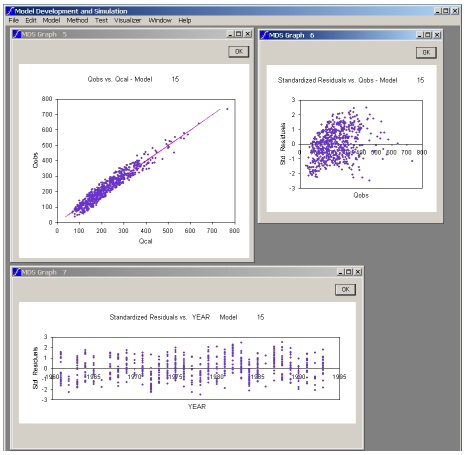


Appendix 6

(Continuation). From top-left to bottom-right: 7) selecting the visualizer; 8) authorship; 9) working space showing many examples of possible plots included in the program that help the user to judge the goodness of the fit during the process of calibration of a model.







Appendix 7

European Circulation Patterns according to Hess and Brezowsky (1969).

Major Type	Sub-type	No.	Description	Abbreviation
Zonal circulation	W	1	West, anticyclonic	Wa
		2	West, cyclonic	Wz
		3	Southern, West	WS
		4	Angleformed West	WW
Mixed circulation	SW	5	Southwest, anticyclonic	SWa
		6	Southwest, cyclonic	SWz
	NW	7	Northwest, anticyclonic	NWa
		8	Northwest, cyclonic	NWz
	HM	9	Central European high	HM
		10	Central European ridge	BM
	TM	11	Central European low	TM
Meridional circulation	N	12	North, anticyclonic	Na
		13	North, cyclonic	Nz
		14	North, Iceland high, anticyclonic	HNa
		15	North, Iceland high, cyclonic	HNz
		16	British Isles high	HB
		17	Central European trough	TRM
	NE	18	Northeast, anticyclonic	NEa
		19	Northeast, cyclonic	NEz
	E	20	Fennoscandian high, anticyclonic	HFa
		21	Fennoscandian high, cyclonic	HFz
		22	Norwegian Sea-Fennoscandian high, anticyclonic	HNFa
		23	Norwegian Sea-Fennoscandian high, cyclonic	HNFz
		24	Southeast, anticyclonic	SEa
		25	Southeast, cyclonic	SEz
	S	26	South, anticyclonic	Sa
		27	South, cyclonic	Sz
		28	British Isles low	TB
		29	Western Europe trough	TRW
Unclassified	U	30	Classification not possible	U