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).

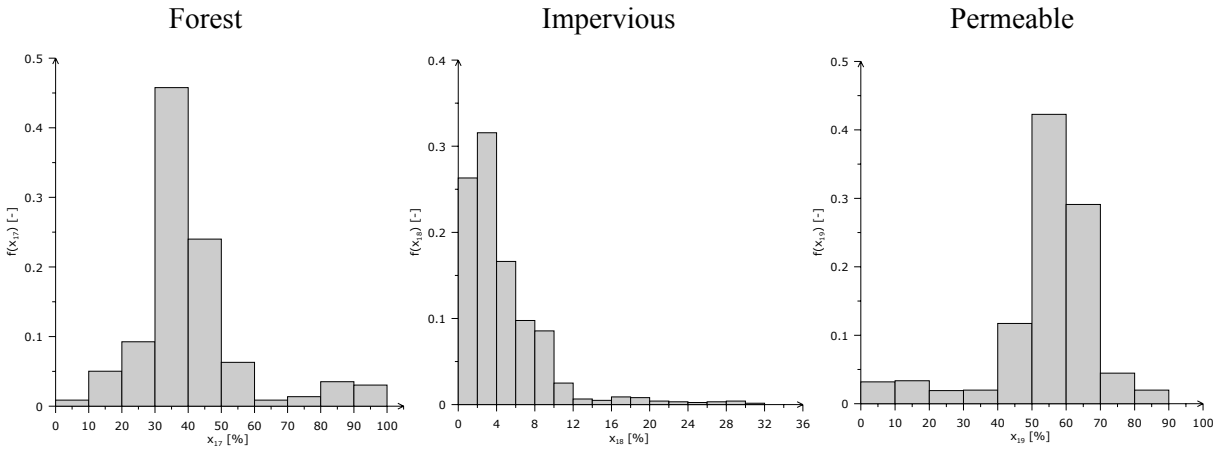

Figure 4.1 Histograms depicting the empiric PDF of the land cover types for all spatial units ($\mathcal{L}_i \subseteq \Omega_i$) from 1961 to 1993 (Number of observations for each histogram = 184).

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.

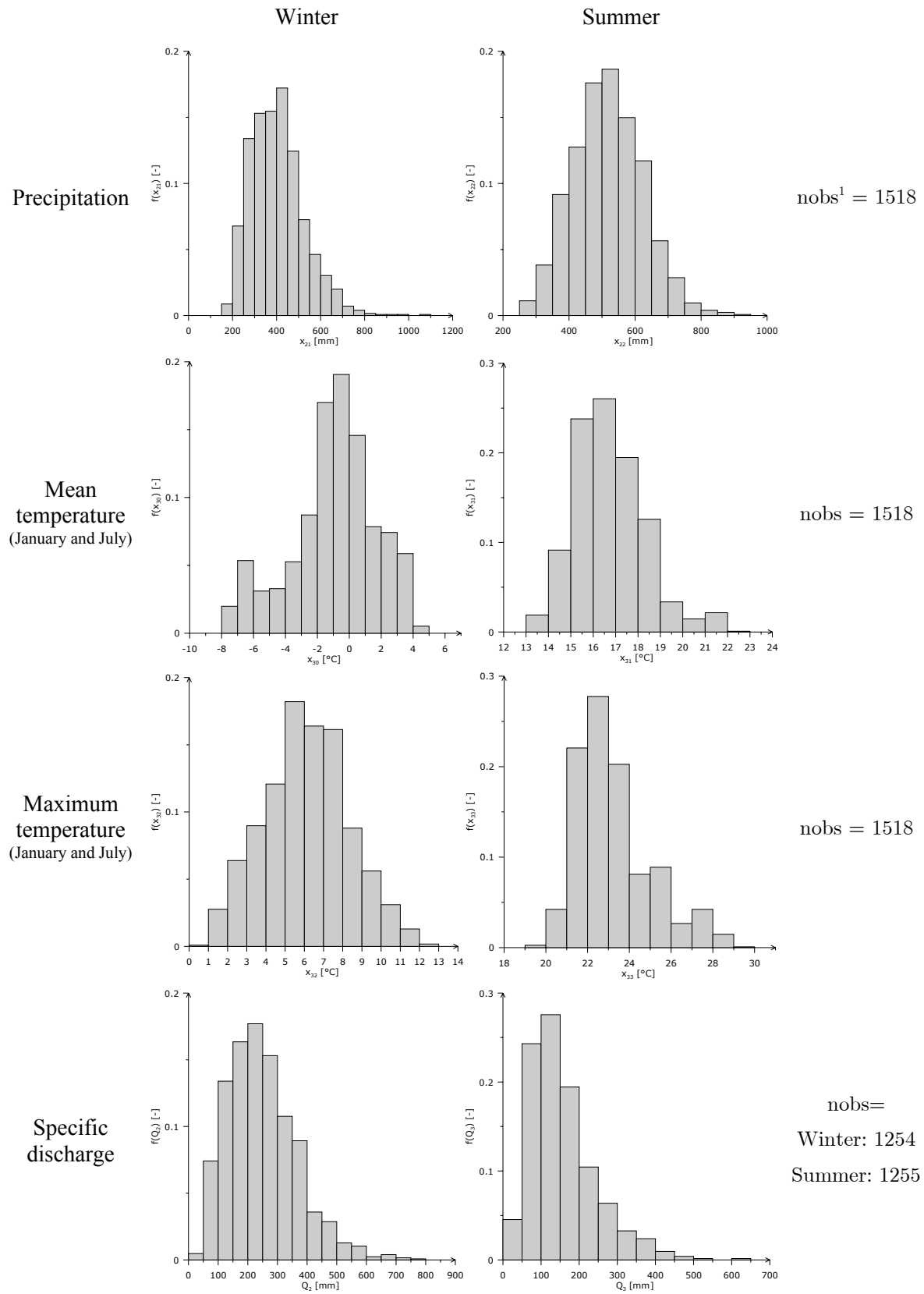

Figure 4.2 Histograms depicting the empiric PDF of climatic factors and specific discharge for all spatial units from 1961 to 1993.

[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

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

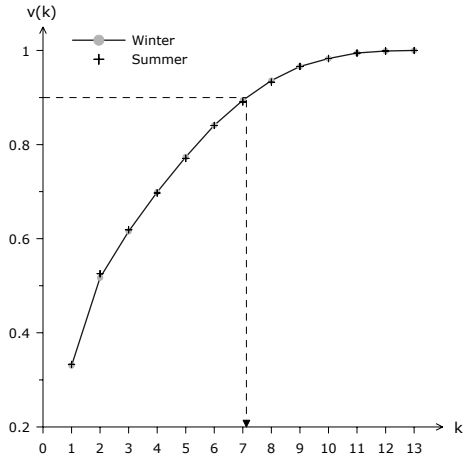

Figure 4.3 Curve showing the relative variance retained by the k first eigenvectors of the matrix $[\mathbf{R}]$ for winter. Additionally, the crosses show the results for the summer season. The correlation matrices have been calculated with time series from 1961 to 1993.

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.

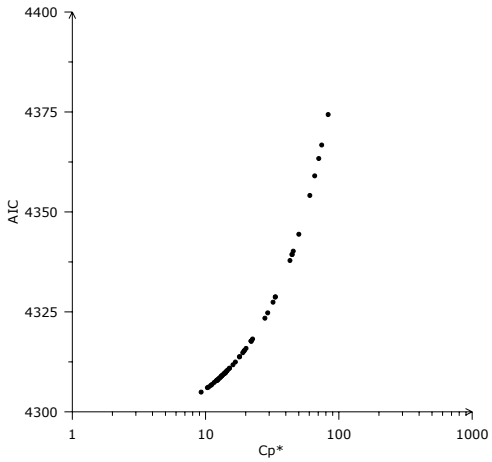

Figure 4.4 Curve depicting the non-linear relationship between the Mallows' C_{p^*} statistic and the AIC for the sample of best performing models described in Table 4.2. The best models in both cases exhibit small values.

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.

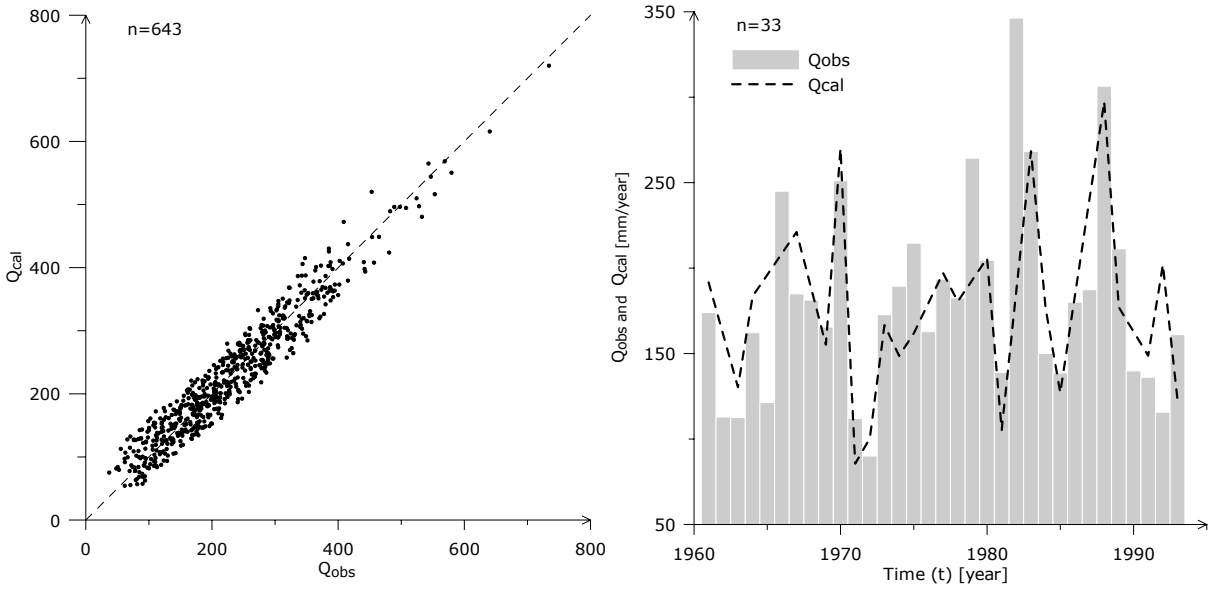

Figure 4.5 At the left panel, a scatterplot shows the relationship between observed and calculated values using model No. 3733 for winter. The samples size is 643. The right panel illustrates a time series of the observed specific discharge in winter and their corresponding calculated values for Basin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

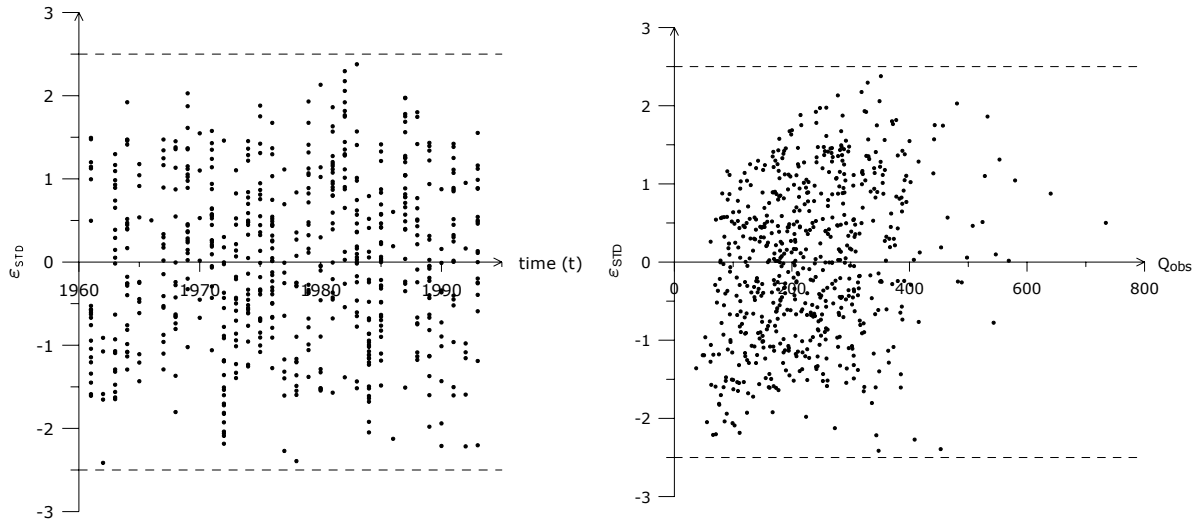

Figure 4.6 The left panel shows a plot of the standardized residuals for winter obtained with model No. 3733 versus time. At the right panel, a standardized residual plot for the same model is presented.

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).

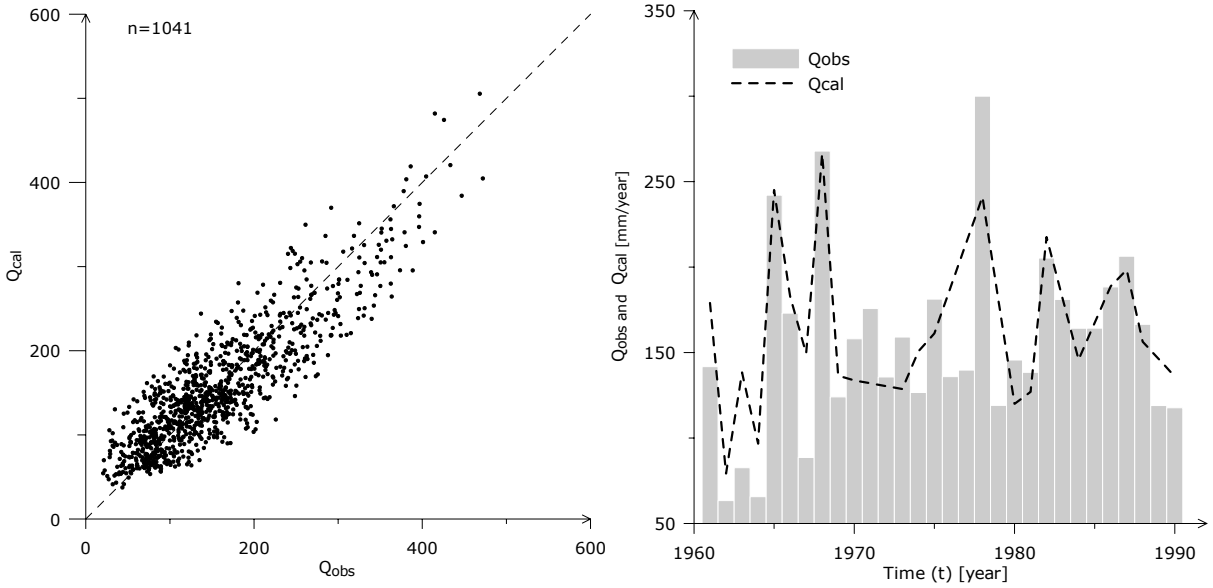

Figure 4.7 The left panel shows a scatterplot of the observed values versus calculated ones for summer obtained with model No. 3965. The right panel illustrates a time series of the observed specific discharge in summer and their corresponding calculated values for Basin No. 13.

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).

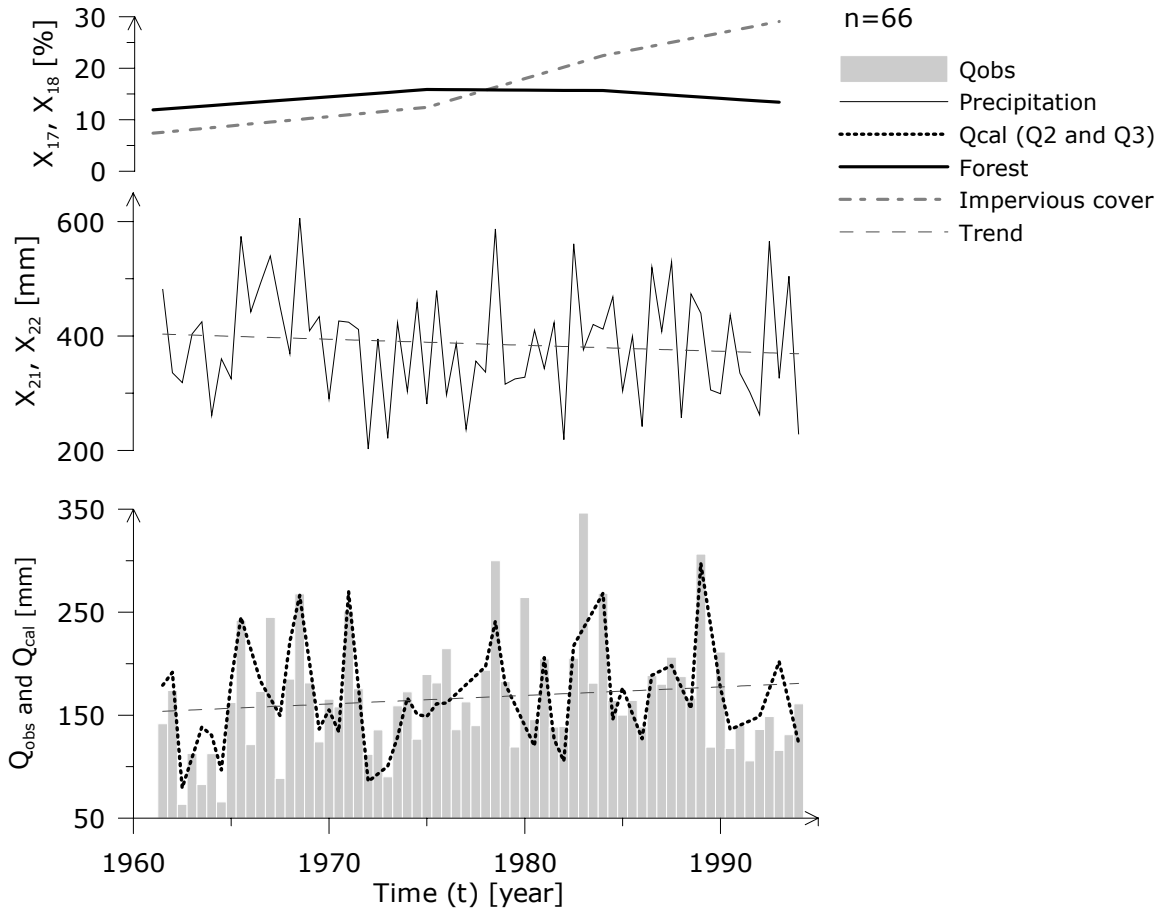

Figure 4.8 Comparison of time series of land cover, precipitation and specific discharge in winter and summer for Basin No. 13. Calculated values using models No. 3733 for winter and No. 3965 for summer are also displayed.

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.

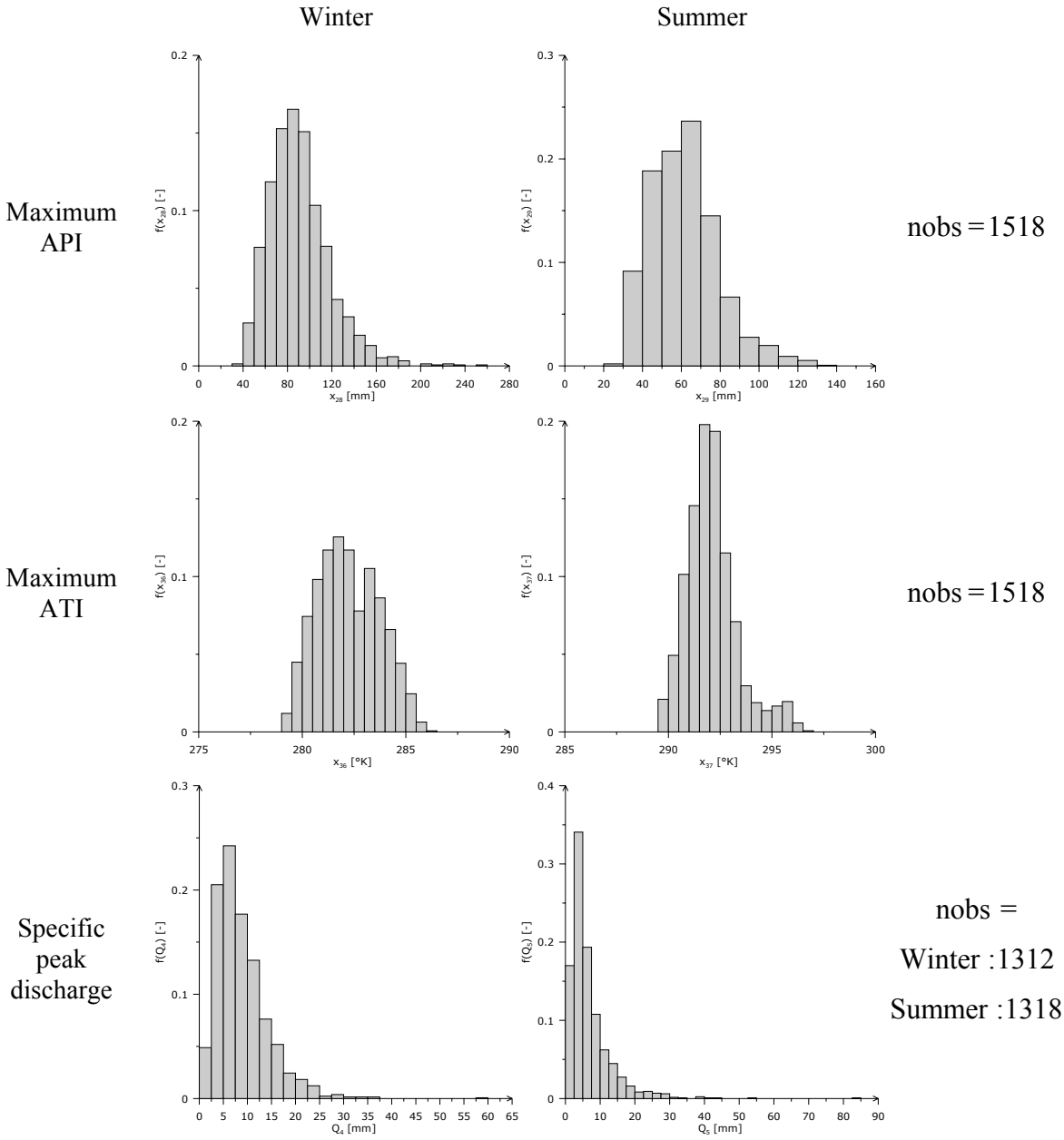

Figure 4.9 Histograms depicting the empiric PDFs for both maximum API and ATI indices for winter (left panel) and summer (right panel), as well as the specific peak discharge considering all spatial units during the period from 1.11.1960 to 31.10.1993.

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.

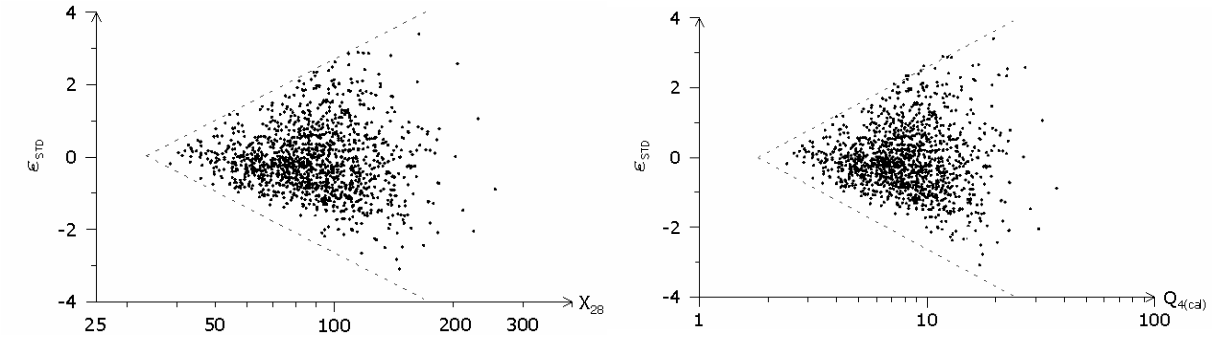

Figure 4.10 Scatterplot of residuals shows a clear heteroscedasticity of the errors with respect to variable x_{28} and the estimated values $\hat{Q}_4$ using model No. 1308.

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.

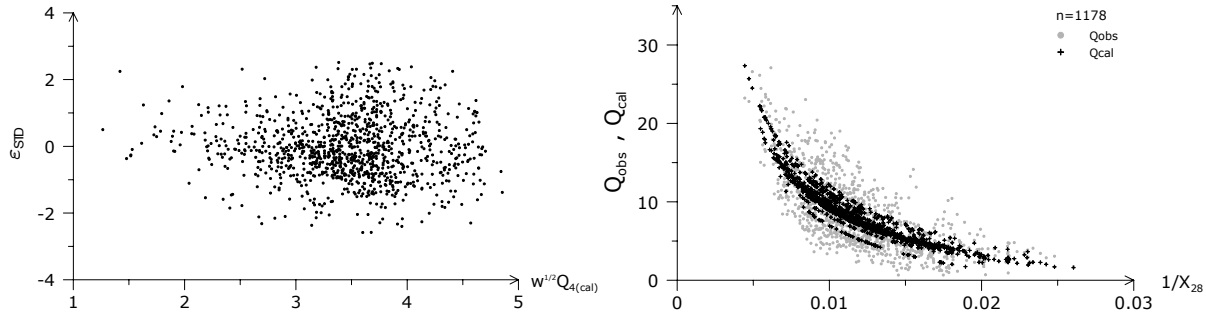

Figure 4.11 The left panel shows a scatterplot of residuals obtained for model No. 1308 using the estimator described by (4.20) and (4.21). The graph at the right panel shows the nonlinear relationship among the calculated/observed specific peak in winter and the inverse of the maximum precipitation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

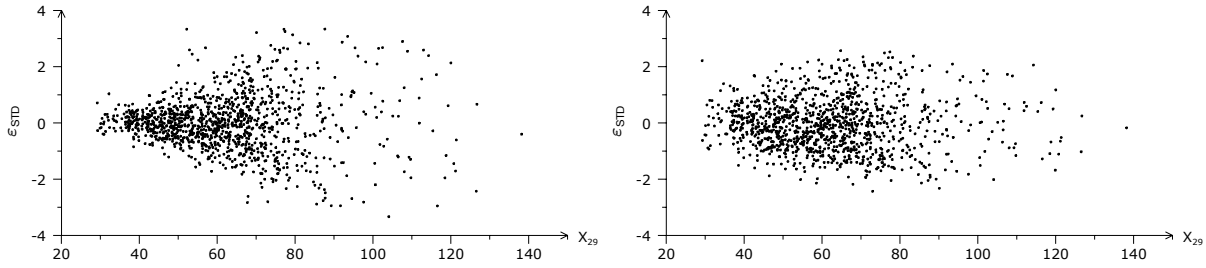

Figure 4.12 Scatterplots of residuals of model No. 3954 before (left panel) and after (right panel) the heteroscedasticity of the errors with respect to variable x_{29} has been removed.

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.

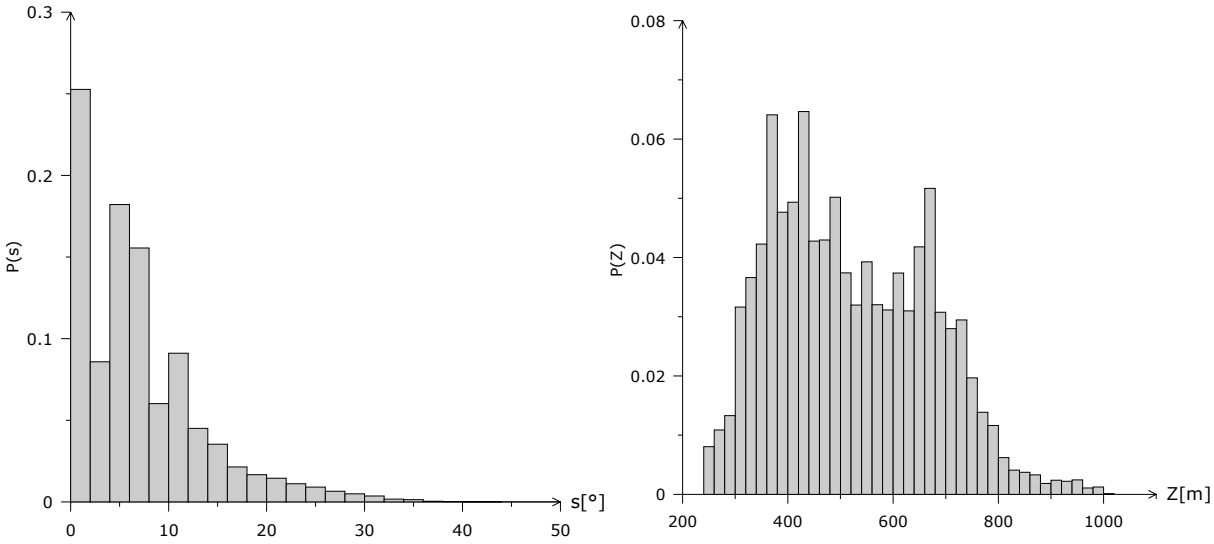

Figure 4.13 PDF showing the likelihood of a given place to endure a land cover change based on its slope and elevation. These curves take into account all locations that have undergone a land cover change from 1960 to 1993.

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.

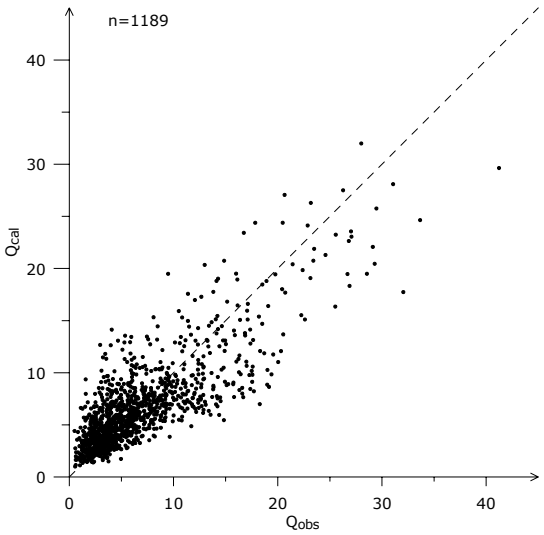

Figure 4.14 This scatterplot shows the relationship between calculated and observed specific peak flows using the potential model No. 3954.

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.

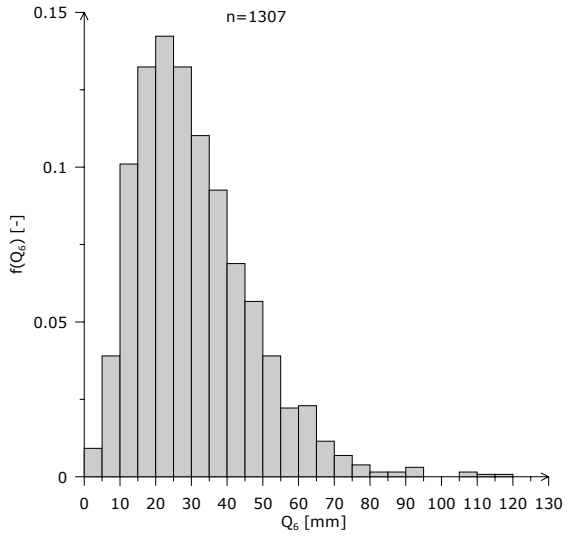

Figure 4.15 Histogram depicting the PDF of the cumulative specific discharge of the annual peak event (Q_6) considering the time series from 1.11.1960 to 31.10.1993 for all spatial units.

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).

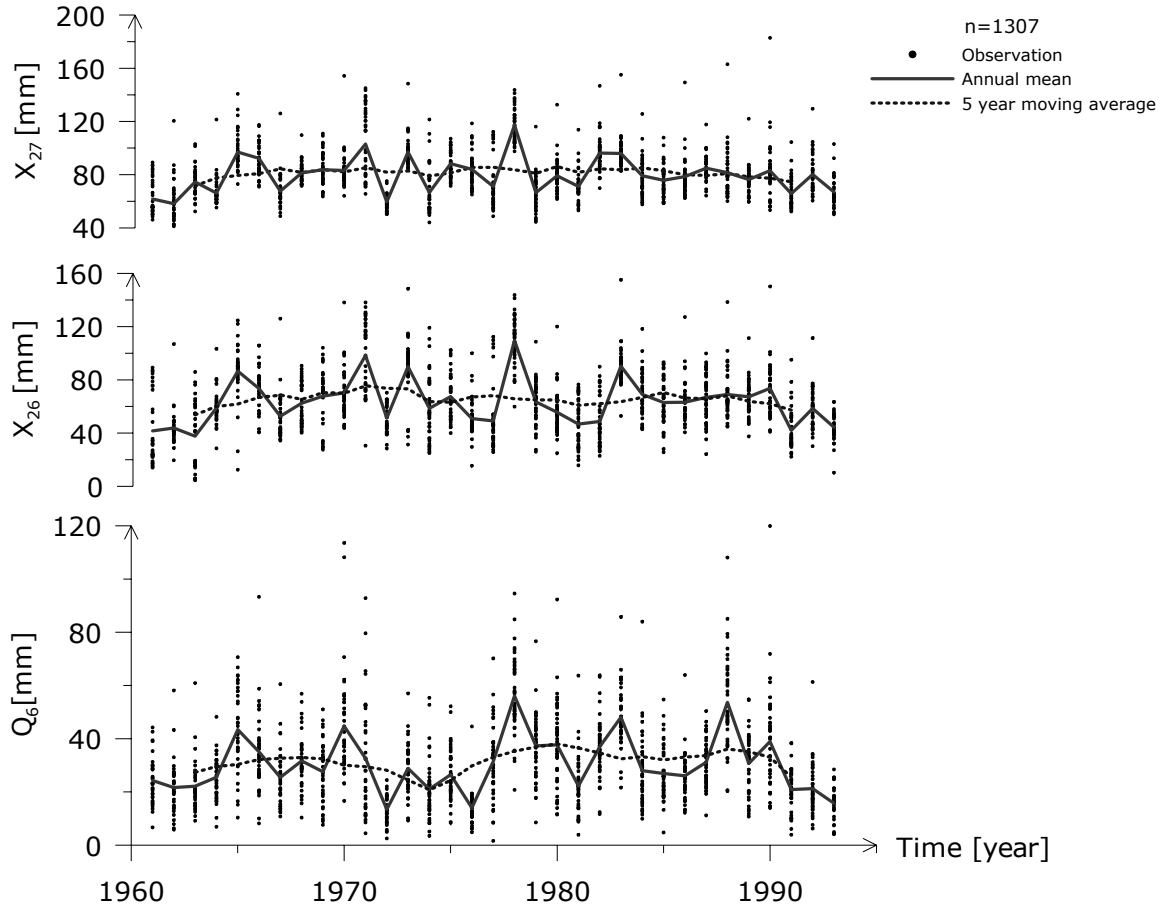

Figure 4.16 Comparison of time series showing the variability of the explained variable (Q_6) and two climatic factors (x_{26}) and (x_{27}). Each observation is represented by a point during the period from 1.11.1960 to 31.10.1993. The annual mean is depicted by a continuous line. The trend of these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

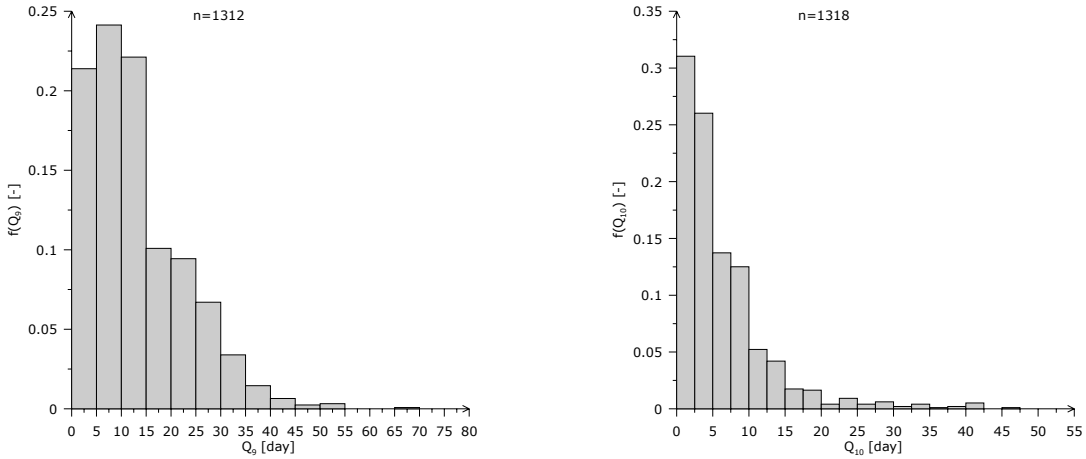

Figure 4.17 Histograms depicting the empiric PDFs for both total duration of high flows in winter (left panel) and summer (right panel) considering all spatial units during the period from 1.11.1960 to 31.10.1993.

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.

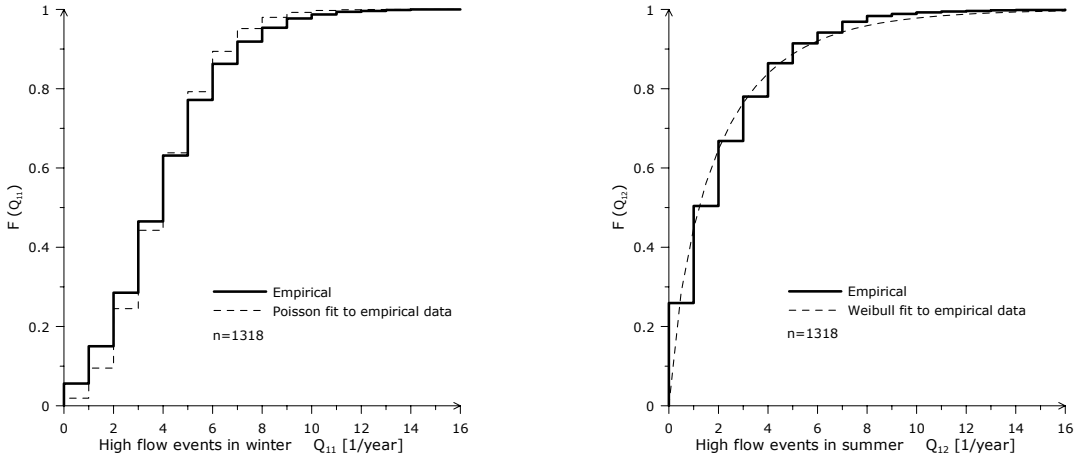

Figure 4.18 Empirical and fitted CDFs for both frequency of high flow events in winter (left panel) and summer (right panel) considering all spatial units during the period from 1.11.1960 to 31.10.1993.

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.

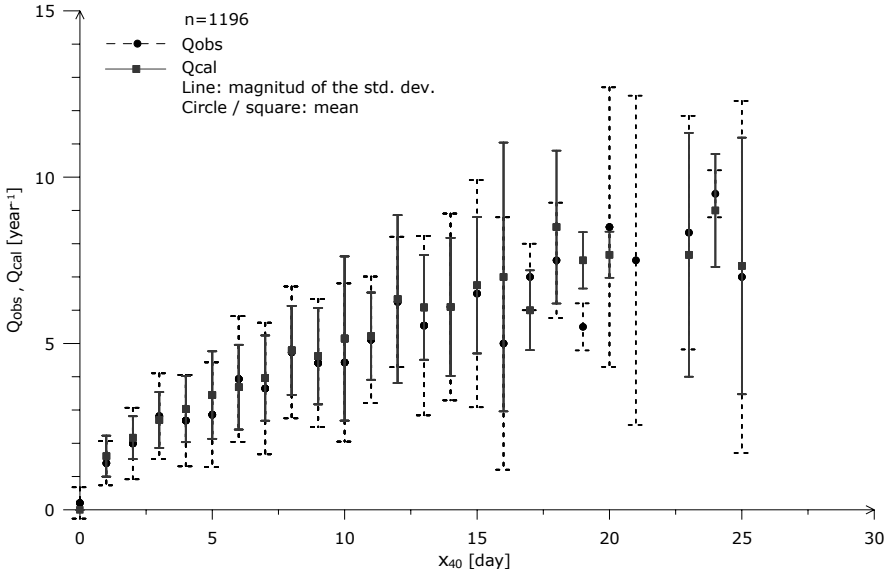

Figure 4.19 Plot showing the variation of the dispersion of the explained variable Q_{12} (observed and calculated by model No. 3052) as a function of the predictor x_{40}. Both continued and dashed lines represent the magnitude of the standard deviation whereas dots and rectangles represent the mean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