Universität Stuttgart
Geodätisches Institut

Least-squares prediction of runoff

Bachelorarbeit im Studiengang
Geodäsie und Geoinformatik
an der Universität Stuttgart

Robin Thor

Stuttgart, September 2013

Betreuer: Prof. Dr.-Ing. Nico Sneeuw
Universität Stuttgart

M. Sc. Mohammad Tourian
Universität Stuttgart

Erklärung der Urheberschaft

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum Unterschrift

When modelling hydrological cycles, the runoff of a drainage basin is an important variable, being an output from hydrological models and an input to many hydrological interactions as a quantity used for validation and calibration. In this context, the decrease of the availability of in situ runoff measurements that has been observed over the last years poses a challenge which this study aims to tackle with the methods of least-squares prediction. This research uses the spatial correlations between in situ runoff measurements generated in a training period to predict values for a validation period during which one of the catchments is assumed ungauged. Different methods include the usage of covariance matrices, which are formed

1. on the signal level, or
2. separately for each of the 12 months of the year, or
3. after the reduction of the monthly mean, or
4. after the reduction of the long-term mean

for prediction purposes. For validation, the Nash-Sutcliffe model efficiency coefficient, correlation, and RMSE are computed. The impacts of variations in the length of the training period and of the choice of catchments whose observed measurements are used in the prediction process are analysed. The errors then undergo a spectral analysis to test, which prediction methods are able to capture cyclostationary behaviour best.
Most of the methods provide viable results, although the prediction based on covariance matrices generated out of residuals is slightly better than the other methods in a vast majority of configurations. After a training period of 20 years of simultaneous data and with a selection of three catchments used in each prediction process, this method can reach a Nash-Sutcliffe coefficient of over 0.4 for about 90% and of over 0.75 for about 50% of the 25 analysed catchments, although viable results can already be achieved with much shorter training periods of one to three years, depending on the predicted catchment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The hydrological cycle is the process of transport of water through atmosphere, land areas and oceans (Hendriks, 2010). Its analysis is the main object of hydrology. The hydrological cycle consists of evaporation, precipitation, surface runoff, infiltration into the groundwater, groundwater flow, and discharge into the oceans (Maidment, 1993). Among these, discharge is a variable that can be monitored by gauging stations which measure the amount of water flowing through a river when the profile of the riverbed is known (Bárdossy, 2005). The discharge can be computed out of the water level at the gauging station and the measured velocity of the flow which is also called runoff (Bárdossy, 2005). Discharge is usually given in units of volume per time and runoff is mostly referred to in units of distance per time (Maidment, 1993). A useful unit to study the hydrological cycle is the drainage basin, or catchment. This is a geographical area that drains through a river, determined by the height configuration of the land surface which sets the direction of the surface water flows as these obey gravity (Hendriks, 2010). The boundary between a pair of neighbouring catchments is called watershed (Bárdossy, 2005).
Hydrologists want to analyse the hydrological cycle and model the processes that define it. A hydrological model includes variables like precipitation, evaporation, leakage from groundwater, and discharge. The most downstream gauging station often offers a sufficient value for the modelling of its catchment (Fekete and Vörösmarty, 2007). When precise runoff measurements for a catchment are available, these can be used as the output of the water balance for this catchment, allowing to draw conclusions about the other parameters that define the hydrological cycle (Sneeuw et al., 2013). They are also useful for the validation and calibration of other measurement methods (Milzow et al., 2011).
There have been attempts to replace traditional in situ runoff measurements by remote sensing applications, for example by Microwave Radiometer (Vörösmarty et al., 1996), Synthetic Aperture Radar (Smith et al., 1996), interferometric radar measurements (Alsdorf et al., 2000) and radar altimeter data (Birkett, 1998), and to combine the satellite applications radar altimetry, Synthetic Aperture Radar surface soil moisture estimates and GRACE gravity data for calibration purposes (Milzow et al., 2011). Bjerklie et al. (2005) use aerial orthophotos and SAR images together with channel slope data obtained from topographic maps to estimate discharge.
However, all of these techniques have their limitations. Especially the lack of large open bodies of water often poses a problem. In fact, Fekete and Vörösmarty (2007) have come to the conclusion that "most of the sensors considered today for river monitoring require a minimum 200 m river width". In the case of interferometric radar measurements large bodies of water are needed which are not completely open but extensively provided with vertical objects like emerging trees (Alsdorf et al., 2000). Most of the current radar altimeters are originally designed to monitor ocean levels. In order to improve their suitability for the measurement of rivers, different sensor specifications and antenna sizes would be necessary (Fekete and Vörös-

marty, 2007). Remote sensing methods mostly rely on measuring the water height, whereas informations about the relationship between the water stage and the actual amount of discharge can usually only be gained from field surveys or long time calibrations (Fekete and Vörösmarty, 2007). Moreover, all of the space-based techniques are restricted by the satellite repeat cycle (Sun et al., 2010) while the methods presented in this study can potentially achieve any temporal resolution, eliminating the need to rely on interpolations. Conventional approaches using the hydrological and hydro-meteorological balance equations encounter problems when the amplitude of runoff for the estimated catchment is low (Sneeuw et al., 2013).
While in situ runoff measurements are sometimes unavailable in larger catchments, the situation is even worse for small drainage basins that are situated near to the coast (Blöschl et al., 2013), adding to the impracticability of deriving discharge measurements for these from remote sensing missions. However, these small catchments are usually in the proximity of larger catchments that may come with a better availability of data from gauging stations.
This study focuses on modelling runoff by least-squares prediction. Modelling is necessary where measurements have not been carried out or the results of the evaluation of measured data are not yet known and their quality uncertain, as in the case of remote sensing data and especially smaller catchments draining through a river of less than 200 metres width. Least-squares prediction uses correlations to model runoff of one catchment based on the measurements that have been done in other catchments. This method is particularly effective when there are catchments that show a very similar behaviour to the catchment whose runoff is to be predicted. Spatial proximity often leads to a high similarity and therefore correlation. The spatial distribution of ungauged catchments argues for a convenient modelling by least-squares prediction. Increasing the availability of accurate runoff data to virtually any catchment, potentially even catchments which have never been gauged before by using theoretical approaches at the creation of covariance matrices, can be a basis for a further analysis of various hydrological problems, including the calibration of satellite based experiments which attempt to derive runoff data from their measurements.
The least-squares prediction of runoff can possibly be a partial substitute for costly satellite missions or the installation of additional gauging stations without the constrictions posed by development and construction times.
To approach the problem, a sample dataset of catchments has been chosen which is investigated in this study. For these catchments some characteristic parameters have been defined to analyse them further (chapter 2). Consequently, a short overview of the principles of least-squares prediction (chapter 3) serves as a foundation for some different approaches at the formation of covariance matrices which are needed for the prediction. These are evaluated (chapter 4) and validated with respect to different configurations and settings (chapter 5). A final discussion sums up and rates the results (chapter 6).

Chapter 2

The dataset

2.1 Origin

The initial set of data that has been dealt with in this thesis is derived from in situ river discharge measurements that have been made over the whole world. The dataset consists of the 255 largest catchments in the world by area, including those which are dischargeless. For each catchment with discharge one measurement station has been chosen out of the available stations from the Global Runoff Data Centre (GRDC). For those stations that have their original data given with a daily resolution, monthly runoff values have been computed, so that the data is available in a monthly interval for all catchments. While the original values are in units of cubic metres per second, the initial data set has these already converted to millimetres per month, referring to the average rainfall needed in the whole area of the catchment for that amount of runoff at the measurement station to come about. The area of each catchment which is needed for this conversion is also provided in the dataset. The conversion is a useful process because the runoff values of different catchments can be compared more easily this way.
Additionally, approximated coordinates in longitude and latitude of the centre of each catchment are provided (table 2.1) which will prove useful later on when the impact of distance between catchments on their correlation is analysed (see section 5.2).

2.2 Analysis

The time series in the initial dataset provided for this study start at the earliest in January of 1980. Therefore the time frame analysed in this thesis will begin in 1980. Figure 2.1 shows the decline of the number of catchments that have data available over the course of the last three decades. Out of the 255 largest catchments 75 are dischargeless, so in 1980, a quite large percentage of the catchments had measurement stations whose data have been made available. While data for some catchments have not been collected at all, the strong decline in the last few years also exists to a certain part because data have been collected, but not yet published. Then, there are a lot of catchments with incomplete data, be it for the majority of the time since 1980 or only for some months.

Table 2.1: *Longitude λ, latitude ϕ and area A of the catchments in the sample dataset.*

Catchment	$\lambda[^\circ]$	$\phi[^\circ]$	$A[\mathrm{km}^2]$
AMAZONAS	−55.50	−7.25	4672876
MISSISSIPPI RIVER	−90.91	41.00	2938538
YENISEI	86.50	58.00	2454961
LENA	127.65	61.50	2417932
ST LAWRENCE	−73.62	45.50	943769
ORANGE	17.73	−27.25	828475
DANUBE	28.73	46.25	771277
COLORADO RIVER (PACIFIC OCEAN)	−114.63	37.25	636508
NORTHERN DVINA	41.92	61.50	330709
FRASER RIVER	−121.45	52.50	228874
COLORADO RIVER (CARIBBEAN SEA)	−96.10	31.50	115821
BRAZOS RIVER	−95.76	32.25	106914
SUSQUEHANNA RIVER	−76.18	41.25	69694
VUOKSI	28.78	62.75	66393
APALACHICOLA RIVER	−85.02	32.50	57546
ALABAMA	−87.51	33.25	54249
MEZEN	45.62	64.00	54125
STIKINE RIVER	−132.13	57.75	52894
TOMBIGBEE	−88.13	33.00	51622
TRINITY RIVER (TEXAS)	−94.85	32.00	44365
GEORGE RIVER	−65.84	56.25	36171
ALTAMAHA RIVER	−81.83	33.00	33706
KYMIJOKI	26.82	62.00	33195
SKEENA RIVER	−128.43	55.50	31555
POTOMAC RIVER	−77.13	39.00	31151

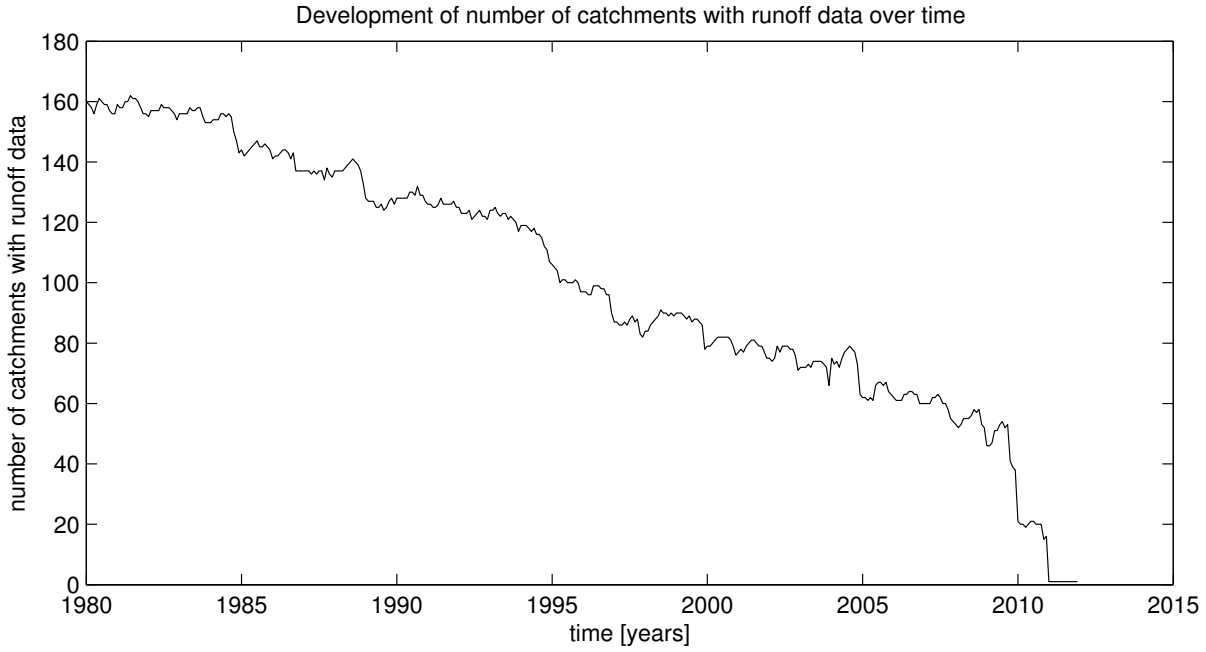

Figure 2.1: *Development of the number of catchments that have runoff data available for the respective month.*

2.3 Catchments with complete time series

The incompleteness of the dataset has resulted in the decision to take a data sample of all catchments that have a complete time series of runoff data from January 1980 to December 2008, accounting for a total of 29 years or 348 months covered. This sample is used in matrix form, named $\boldsymbol{R}$, and contains the time in the first dimension and the catchments in the second dimension. It is used during the rest of the thesis as it is necessary to compare predicted values with in situ measured values to evaluate the quality of a prediction method. These values measured by gauging stations are referred to as *true values* hereafter.
The sample dataset contains time series for 25 catchments. These are rather unrepresentative, mainly lying in highly developed countries, because elsewhere a continuous maintenance of gauge stations seems more unlikely. This means that the subarctic climate is covered in Russia and Canada, the subtropical climate in the Southern USA and the continental climate in Eastern and Northern Europe. But from this small sample of catchments a lot about prediction methods can be learnt already. Figure 2.2 shows a map of catchments whose runoff measurements are used in the sample. These catchments are also represented by table 2.1.

2.4 Signal parameters

The following four parameters describe a time series of runoff measurements. Their values for the sample dataset are shown in table 2.2.

Table 2.2: *Signal parameters: root mean square (RMS), range, cyclostationarity factor γ, trend.*

Catchment	RMS $\left[\frac{\text{mm}}{\text{month}}\right]$	range $\left[\frac{\text{mm}}{\text{month}}\right]$	γ	trend $\left[\frac{\text{mm}}{\text{month}}\right]$
AMAZONAS	100.3	127.9	0.88	9.15
MISSISSIPPI RIVER	18.0	46.1	0.67	0.12
YENISEI	29.0	107.7	0.78	2.37
LENA	29.8	109.8	0.81	2.78
ST LAWRENCE	24.2	17.3	0.88	−3.99
ORANGE	1.2	16.1	0.13	−0.03
DANUBE	23.2	41.5	0.74	1.36
COLORADO RIVER (PACIFIC OCEAN)	0.6	3.2	0.17	−0.73
NORTHERN DVINA	40.4	160.2	0.71	−0.32
FRASER RIVER	38.6	102.4	0.79	0.75
COLORADO RIVER (CARIBBEAN SEA)	3.1	21.4	0.19	0.35
BRAZOS RIVER	8.6	43.0	0.26	1.60
SUSQUEHANNA RIVER	53.5	256.8	0.51	10.31
VUOKSI	25.4	30.8	0.78	−3.83
APALACHICOLA RIVER	36.3	117.8	0.55	−10.34
ALABAMA	53.4	198.6	0.49	−13.49
MEZEN	50.4	218.2	0.71	1.90
STIKINE RIVER	104.1	267.7	0.81	−4.59
TOMBIGBEE	58.2	239.2	0.46	−11.91
TRINITY RIVER (TEXAS)	22.5	100.6	0.33	1.62
GEORGE RIVER	24.8	47.5	0.68	−0.58
ALTAMAHA RIVER	38.0	162.2	0.45	−6.85
KYMIJOKI	25.5	38.3	0.70	−5.89
SKEENA RIVER	104.7	354.0	0.75	5.67
POTOMAC RIVER	40.5	160.9	0.43	0.78

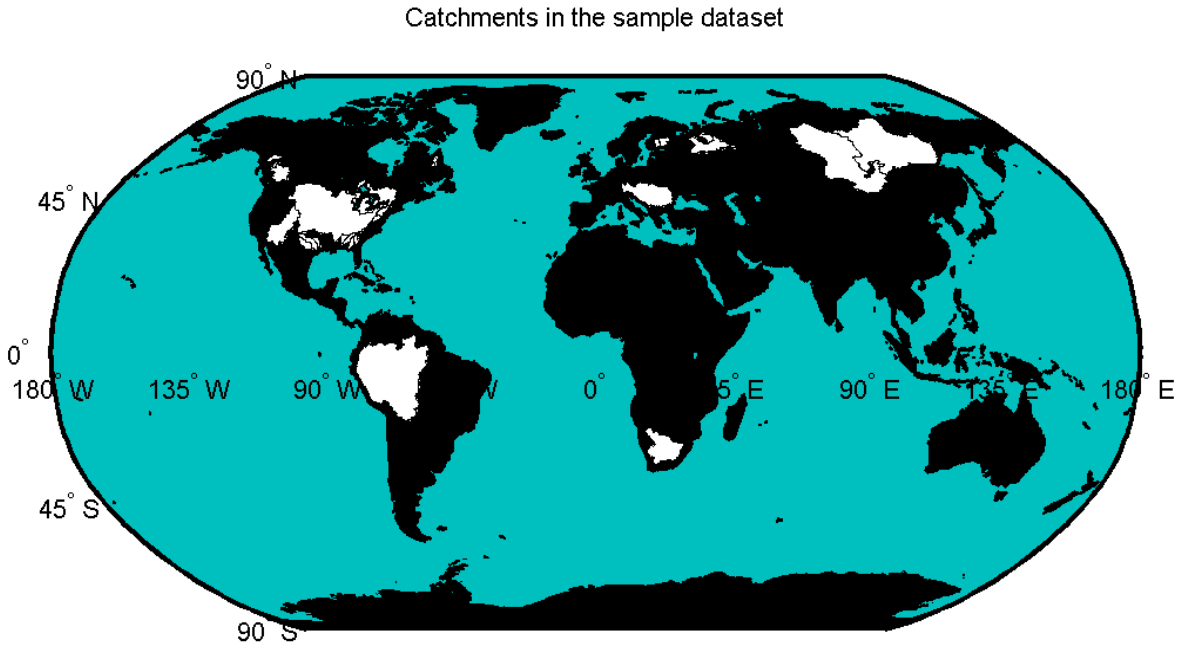

Figure 2.2: *Map showing the distribution of catchments used in the sample dataset (white) in the world.*

2.4.1 RMS

The root mean square (RMS) is defined as

$$\mathrm{RMS} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} r_t^2} \tag{2.1}$$

with

r_t : the values of a time series $\boldsymbol{r}$ at the time t, in this case the runoff values for one catchment

T : the length of the time series

When dealing with the full length signal, T is the number of months in the dataset: $T = 348$. The RMS gives information about the amount of runoff and its variation and is used for the computation of further parameters.

2.4.2 Range

A range can be defined as

$$\mathrm{range} = \max \boldsymbol{r} - \min \boldsymbol{r} \tag{2.2}$$

with $\boldsymbol{r}$ being the runoff time series for one catchment. The range is the margin between the highest and the lowest value.

2.4.3 Trend

When regarded in a broader context an important aspect of discharge variations is the long-time trend. Runoff data is provided as a time series of the length T, given in years. For each year, a mean runoff value is determined. The resulting values are used for a polynomial fit of degree one. Then, the trend can be defined as the difference in runoff of the fit in the last and the first year. When the linear coefficient of the fit is called a_1, the trend can be computed as

$$\text{trend} = a_1(T-1)\ . \tag{2.3}$$

It is given in units of $\frac{\text{mm}}{\text{month}}$. An Analysis of the relationship between trend and other characteristics may prove useful later on.

2.4.4 Cyclostationarity factor

Different catchments show different variations in their runoff values over the course of one year. Sometimes the oscillation follows very regular patterns whereas sometimes it does not. This is shown exemplified for two catchments in figures 2.3 and 2.4.

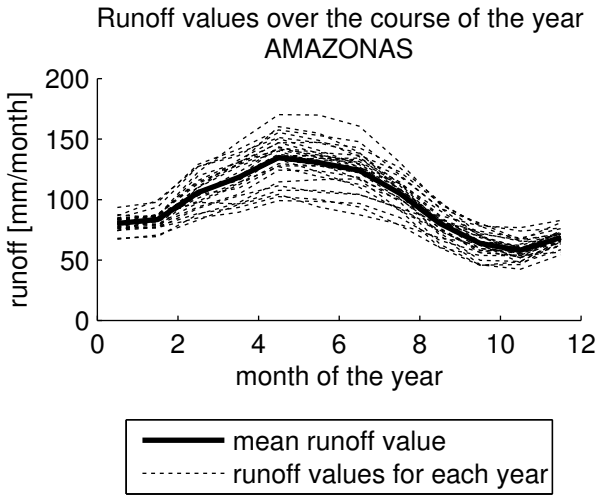

Figure 2.3: *Runoff oscillation of Amazon.*

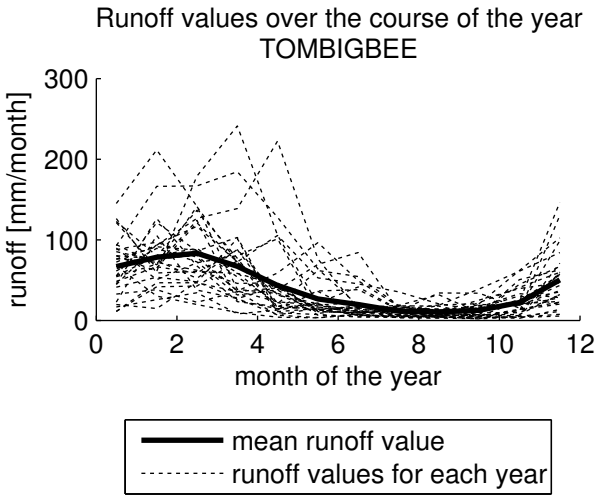

Figure 2.4: *Runoff oscillation of Tombigbee River.*

To quantify these variations, or the cyclostationarity, a new parameter is introduced:

$$\gamma = 1 - \frac{1}{\text{RMS}}\sqrt{\frac{1}{T-1}\sum_{t=1}^{T} r_{\text{res},t}^2} \tag{2.4}$$

$r_{\text{res},t}$ is the value of the residual of the runoff $\boldsymbol{r}_{\text{res}}$ at time t, defined as the difference between the runoff and its monthly mean:

$$\boldsymbol{r}_{\text{res}} = \boldsymbol{r} - \bar{\boldsymbol{r}}_{\text{cyc}} \tag{2.5}$$

where $\bar{\boldsymbol{r}}_{\text{cyc}}$ is a time series containing the mean runoff value for each month of the year. This means that the cyclostationarity factor γ is formed by the ratio of the sample standard deviation

of the residuals to the root mean square RMS. A cyclostationarity factor of $\gamma = 1$ means that the runoff behaves exactly the same way every year. The lower the factor, the lower is the cyclostationarity of the catchment.

Chapter 3

Least-squares prediction

3.1 Covariance and correlation matrices

3.1.1 Covariance

To predict a signal out of an observation it is essential to know the covariance information of both, signal and observation. Let the time series of runoff observations which are the column vectors of the matrix $\boldsymbol{R}$ containing the sample dataset be called $\boldsymbol{r}$. Then the covariance describing the relationship between two time series $\boldsymbol{r}_j$ and $\boldsymbol{r}_k$ is defined (Schlittgen and Streitberg, 1995) as

$$c_{jk} = \frac{1}{T-1} \sum_{t=1}^{T} (r_{tj} - \bar{r}_j)(r_{tk} - \bar{r}_k) \tag{3.1}$$

with

r_{tj} , r_{tk} : the element of the respective time series $\boldsymbol{r}_j$ or $\boldsymbol{r}_k$ at the time t

$\bar{r}_j$, $\bar{r}_k$: the mean of the respective time series

T : their length

The covariance matrix $\boldsymbol{C} = [c_{jk}]$ of the dataset $\boldsymbol{R}$ containing n time series can be formed out of the coefficients c_{jk} with $j = 1, ..., n$ and $k = 1, ..., n$. This matrix is a symmetric matrix containing the variances on the main diagonal.

3.1.2 Cross-covariance

When one set of runoff $\boldsymbol{R}_1$ including n_1 time series $\boldsymbol{r}_j$ of the length T and another set of runoff $\boldsymbol{R}_2$ including n_2 time series $\boldsymbol{r}_k$, also of the length T, are given, the cross-covariance between the two time series can be determined as in equation 3.1. The resulting cross-covariance matrix is called $\boldsymbol{C}_{12}$ and is formed out of the coefficients c_{jk} with $j = 1, ..., n_1$ and $k = 1, ..., n_2$. This matrix is generally not square and contains only covariances. This is a generalisation of equation 3.1.

3.1.3 Correlation

It may also be interesting to take a look at the correspondent correlation matrix. The sample correlation between two time series $\boldsymbol{r}_j$ and $\boldsymbol{r}_k$ is defined as

$$\rho_{jk} = \frac{1}{(T-1)\sigma_j\sigma_k} \sum_{t=1}^{T} (r_{tj} - \bar{r}_j)(r_{tk} - \bar{r}_k) = \frac{c_{jk}}{\sigma_j\sigma_k} \tag{3.2}$$

with σ_j and σ_k being the standard deviations of respective time series $\boldsymbol{r}_j$ and $\boldsymbol{r}_k$. Then the correlation matrix $[\rho_{jk}]$ of a data set $\boldsymbol{R}$ containing n time series is formed out of the coefficients ρ_{jk} with $j = 1, ..., n$ and $k = 1, ..., n$. It is a symmetric matrix with ones on the main diagonal. The other values lie between -1 and 1 and denote the correlation between time series $\boldsymbol{r}_j$ and $\boldsymbol{r}_k$.

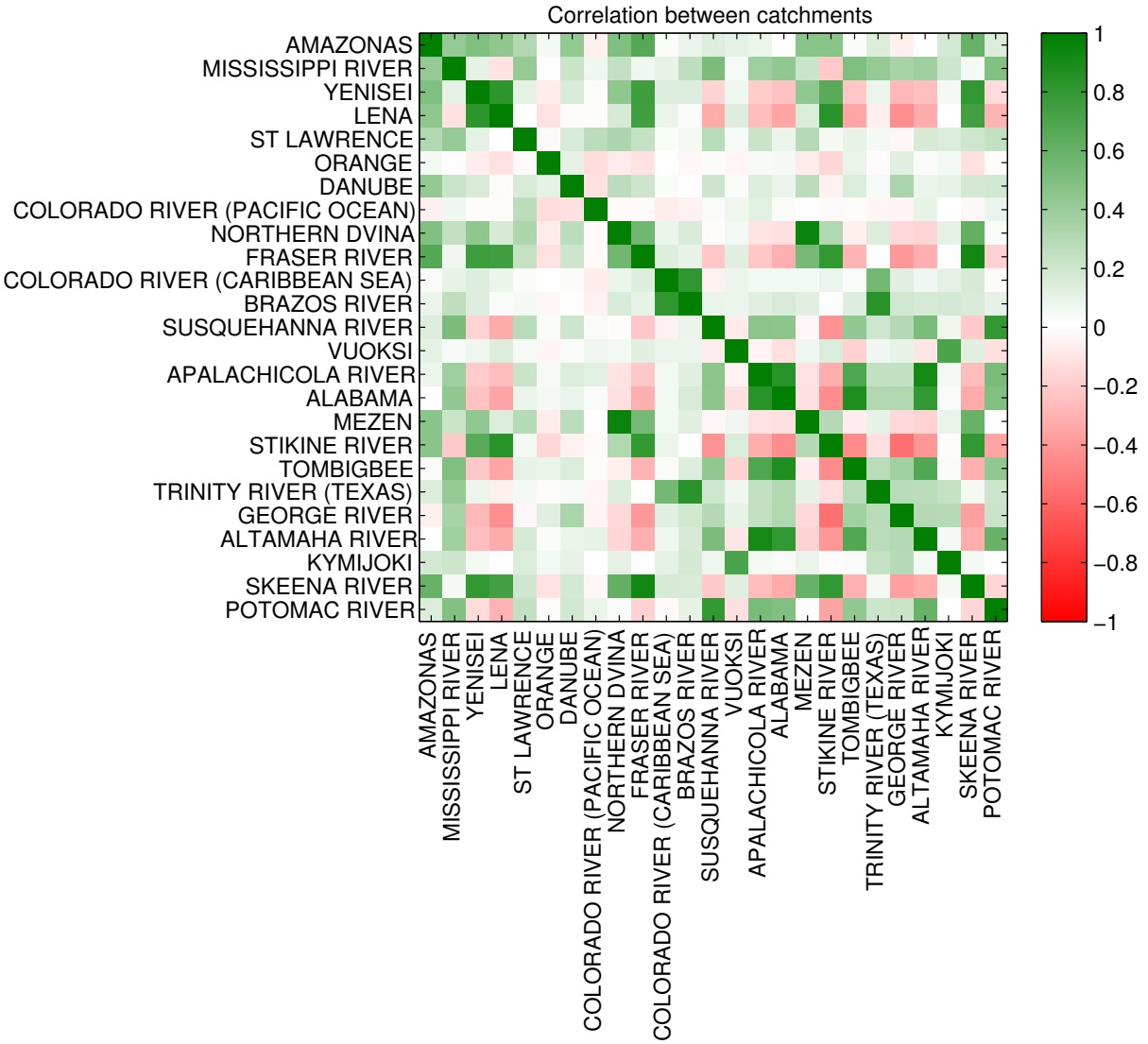

Figure 3.1: *Graphic representation of the correlation matrix of the sample dataset of 25 catchments.*

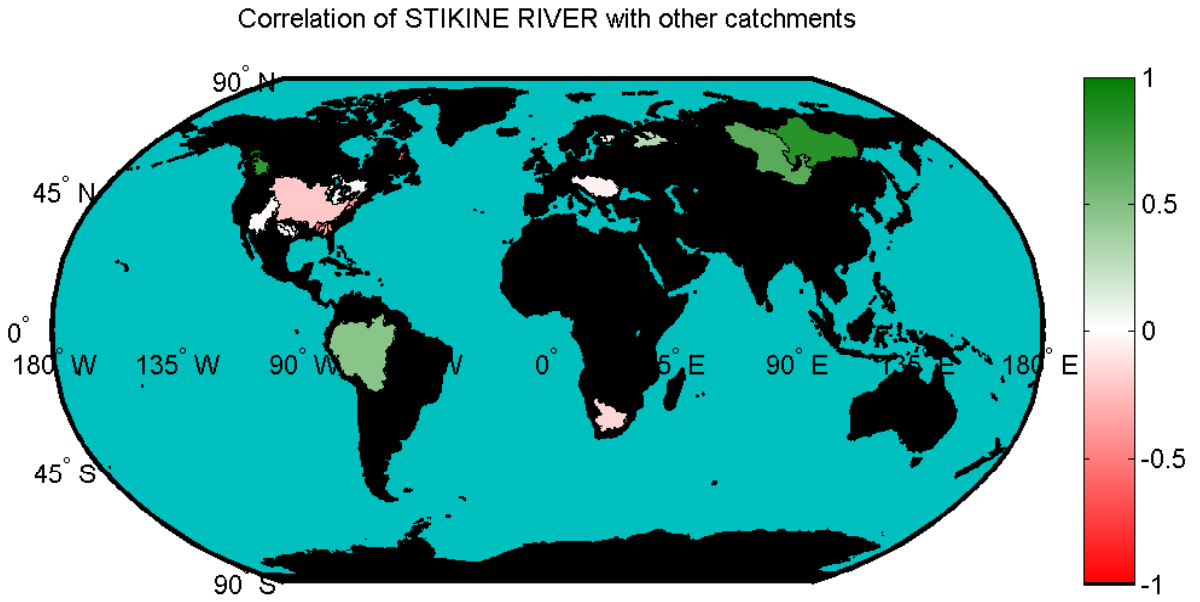

Figure 3.2: *Map showing correlation values for Stikine River, located in western Canada. High correlation occurs with other catchments in the boreal areas, whereas negative correlation values with catchments on the North American east coast indicate an opposite annual course of the runoff quantity.*

3.2 Prediction model

Fundamental information on least-squares prediction can for example be found in Moritz (1980), a source which is used for this short introduction. $\boldsymbol{s}$ and $\boldsymbol{l}$ are arrays of signals. The observations $\boldsymbol{l}$ are known and $\boldsymbol{s}$ is assumed to be unknown and therefore its prediction is computed as

$$\hat{\boldsymbol{s}} = \boldsymbol{H}\boldsymbol{l} \tag{3.3}$$

where $\boldsymbol{H}$ is some model matrix which is to be found. This prediction $\hat{\boldsymbol{s}}$ can later be compared with the true values for $\boldsymbol{s}$ to determine a model error

$$\boldsymbol{e} = \boldsymbol{s} - \hat{\boldsymbol{s}}\,, \tag{3.4}$$

also called inconsistency. Insertion of equation 3.3 into equation 3.4 yields

$$\boldsymbol{e}\boldsymbol{e}^{\mathsf{T}} = (\boldsymbol{s} - \boldsymbol{H}\boldsymbol{l})(\boldsymbol{s} - \boldsymbol{H}\boldsymbol{l})^{\mathsf{T}} = (\boldsymbol{s} - \boldsymbol{H}\boldsymbol{l})(\boldsymbol{l}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}} - \boldsymbol{s}^{\mathsf{T}}) = \boldsymbol{H}\boldsymbol{l}\boldsymbol{l}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}} - \boldsymbol{s}\boldsymbol{l}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}} - \boldsymbol{H}\boldsymbol{l}\boldsymbol{s}^{\mathsf{T}} + \boldsymbol{s}\boldsymbol{s}^{\mathsf{T}} \tag{3.5}$$

Then, the error covariance matrix $\boldsymbol{C}_{ee}$ can be computed as

$$\boldsymbol{C}_{ee} = \boldsymbol{H}\boldsymbol{C}_{ll}\boldsymbol{H}^{\mathsf{T}} - \boldsymbol{C}_{sl}\boldsymbol{H}^{\mathsf{T}} - \boldsymbol{H}\boldsymbol{C}_{sl}{}^{\mathsf{T}} + \boldsymbol{C}_{ss} \tag{3.6}$$

where $\boldsymbol{C}_{ll}$, $\boldsymbol{C}_{sl}$, and $\boldsymbol{C}_{ss}$ are the covariance matrices of the respective signal arrays. $\boldsymbol{C}_{ee}$ has to be minimised in order to create the best linear unbiased estimation. It is easily proved that $\boldsymbol{C}_{ee}$ is minimised for

$$\boldsymbol{H} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1} \tag{3.7}$$

which results into the formula of least-squares prediction

$$\hat{\boldsymbol{s}} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}\boldsymbol{l} \tag{3.8}$$

when inserted into equation 3.3. To meet the requirements posed by the composition of the dataset used in this study, a slightly modified version of the formula is used, transposing the signal arrays:

$$\hat{\boldsymbol{s}}^{\mathsf{T}} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}\boldsymbol{l}^{\mathsf{T}} \tag{3.9}$$

In this formula, Both $\boldsymbol{s}$ and $\boldsymbol{l}$ will be parts of the matrix of runoff time series $\boldsymbol{R}$, their sizes and positions in this matrix depend on which catchments are desired to be predicted at which period of time. $\boldsymbol{C}_{sl}$ is the cross-covariance matrix of signals $\boldsymbol{s}$ and $\boldsymbol{l}$ and $\boldsymbol{C}_{ll}$ is the covariance matrix of observations. Determining the cross-covariance matrix is a problem because the signal $\boldsymbol{s}$ is unknown and therefore it cannot be computed directly. This issue is dealt with in chapters 4 and 5.

Chapter 4

Prediction methods

4.1 Evaluation of predictions

In this section, some parameters are introduced which evaluate the quality of a prediction. Their values vary depending on the prediction method, several of which are presented in sections 4.2 and 4.3.

4.1.1 Mean of the model error

With formula 3.4 the model error $\boldsymbol{e}$ of a prediction can be computed if the predicted values have initially been known. The mean $\bar{e}$ of this model error has an expectancy of $\bar{e} = 0$. If the actual value deviates strongly from 0, this indicates a bias in the prediction process. The mean of the model error can be used to control the reliability of the prediction.

4.1.2 RMSE

Similar to the RMS, the root mean square of the model error (RMSE) is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} \tag{4.1}$$

with e_t: the values of the model error $\boldsymbol{e}$. For $\bar{e} = 0$ this is the standard deviation of the model error. The RMSE gives information about the size and variation of the model error, and therefore about the quality of the prediction.
It is provided in units of $\frac{\mathrm{mm}}{\mathrm{month}}$ and depends on the amplitude of the original signal. Therefore it cannot be directly compared to the RMSE of other catchments when the quality of the prediction is analysed.

4.1.3 Nash-Sutcliffe coefficient

The Nash-Sutcliffe model efficiency coefficient (Nash and Sutcliffe, 1970) is defined as

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T} (s_t - \hat{s}_t)^2}{\sum_{t=1}^{T} (s_t - \bar{s})^2} \tag{4.2}$$

with

s_t : the value of the signal $\boldsymbol{s}$ at the time t

$\hat{s}_t$: the value of the predicted signal $\hat{\boldsymbol{s}}$ (see equation 3.9) at the time t

$\bar{s}$: the mean of $\boldsymbol{s}$

The coefficient becomes 1 when the prediction perfectly reconstructs the original signal and 0 when the prediction is just as good as the mean of the original signal. Negative values of NSE denote a bad prediction model quality.
It is also useful to define a Nash-Sutcliffe coefficient with respect to the monthly mean instead of the long term mean. This coefficient can be computed as

$$\widetilde{\mathrm{NSE}} = 1 - \frac{\sum_{t=1}^{T} (s_t - \hat{s}_t)^2}{\sum_{t=1}^{T} (s_t - \bar{s}_{\mathrm{cyc}})^2} \tag{4.3}$$

with $\bar{s}_{\mathrm{cyc}}$: the mean value of the signal for the respective month of the year. This formula produces a value of 0 when the prediction is just as good as forming the monthly mean would be. In runoff prediction this comparison is very obvious because river discharge usually oscillates with the period of one year. A prediction with $\widetilde{\mathrm{NSE}} < 0$ is surpassed in its quality by a simple long-term trend of the runoff over the course of the year.
The Nash-Sutcliffe model efficiency coefficient will be used throughout the study to quantify the quality of the prediction results.

4.2 Spatio-temporal prediction approach

The spatio-temporal approach is the traditional way of predicting missing values in a time series. It will be shortly described in this thesis to see why a spatial prediction approach is more useful under the given conditions.
The assumption is made that runoff data for all catchments are known until a certain point in time T_l, and unknown after that. The vector of observations $\boldsymbol{l}$ is then defined as the last row of the known elements of the runoff data matrix $\boldsymbol{R}$ and the signal $\boldsymbol{s}$ to be predicted is the the first row of the unknown elements. This means that data at the time $T_l + 1$ is predicted out of data at the time T_l.
The covariances are not gained from these two rows of runoff data but from the whole set of known values. This means that the process described in section 3.1.2 can be applied by putting

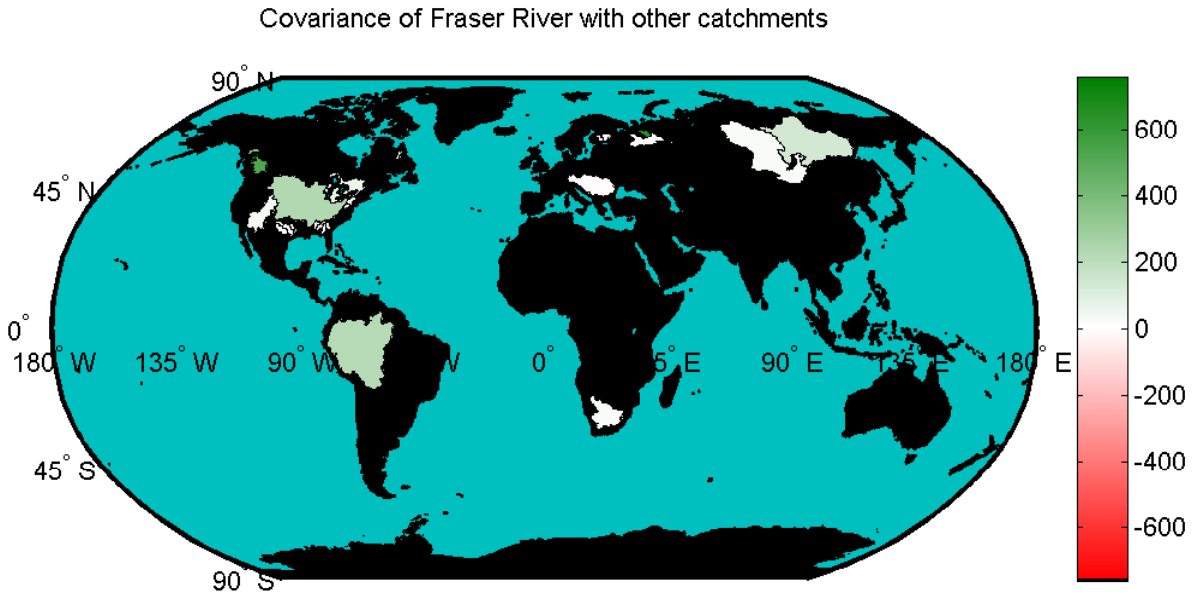

Figure 4.1: *Map showing covariances between Fraser River and other catchments from the $\boldsymbol{C}_{ll}$ matrix in the first iteration of a spatio-temporal prediction. The colour indicates values of covariance between Fraser River and the respective catchment in units of* $\frac{\text{mm}^2}{\text{month}^2}$.

$\boldsymbol{R}_1$ as the first $T_l - 1$ rows of $\boldsymbol{R}$ and $\boldsymbol{R}_2$ as the second to T_l-th row of $\boldsymbol{R}$ for the formation of $\boldsymbol{C}_{sl}$. $\boldsymbol{C}_{ll}$ is formed out of $\boldsymbol{R}_1$ only:

$$c_{jk} = \frac{1}{T_l - 2} \sum_{t=1}^{T_l-1} \left(\left(r_{tj} - \frac{1}{T_l - 1} \sum_{\tau=1}^{T_l-1} r_{\tau j} \right) \left(r_{tk} - \frac{1}{T_l - 1} \sum_{\tau=1}^{T_l-1} r_{\tau k} \right) \right) \tag{4.4}$$

$$c_{mn} = \frac{1}{T_l - 2} \sum_{t=1}^{T_l-1} \left(\left(r_{tm} - \frac{1}{T_l - 1} \sum_{\tau=2}^{T_l} r_{\tau m} \right) \left(r_{tn} - \frac{1}{T_l - 1} \sum_{\tau=1}^{T_l-1} r_{\tau n} \right) \right) \tag{4.5}$$

The covariance matrices are composed out of their components as $\boldsymbol{C}_{ll} = [c_{jk}]$ and $\boldsymbol{C}_{sl} = [c_{mn}]$ and their values are plotted in figures 4.1 and 4.2 for a prediction of the runoff of Fraser River. It can be seen that the covariance with a catchment located more to the south becomes a lot lower, but the value with a more northern catchment becomes higher when the covariance with a time shift of one month is considered.
By this process, only one month ahead is predicted. Therefore, the prediction result $\hat{\boldsymbol{s}}$ is then treated as a known element and the prediction is repeated until the whole dataset has been predicted. This way, $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ become larger with every iteration.
When the whole data matrix $\boldsymbol{R}$ has been predicted, the predicted values starting at row $T_l + 1$ can be compared to the true values for each step in time. Figure 4.3 shows the results for the exemplary catchment with $T_l = 180$ months, so that the calculation of the covariance matrices is based on a time period of 15 years in the first iteration. All other catchments from the sample dataset have been used in the prediction. This is generally a rather advantageous configuration, but still it can be seen clearly that the spatio-temporal prediction can only model the runoff performance for some months and quickly loses its ability to fit the periodic behaviour afterwards. This is because this prediction approach does not use any observations during the process. In a set up where no such validation at later times is possible a spatial prediction ap-

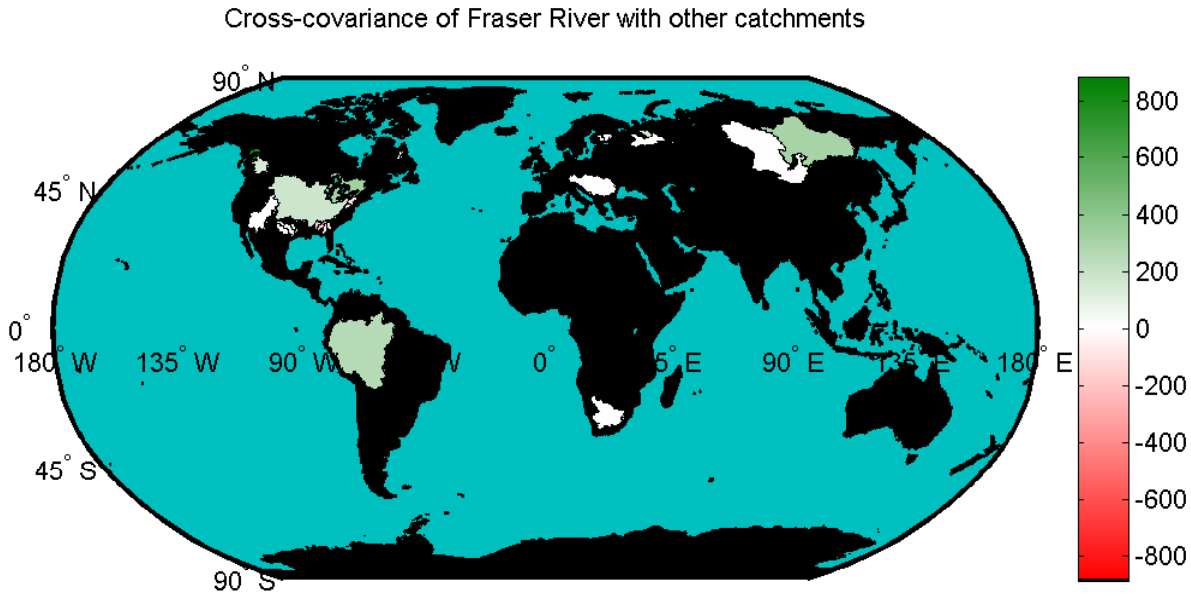

Figure 4.2: *Map showing cross-covariances between Fraser River and other catchments with a time shift of one month from the $\mathbf{C}_{sl}$ matrix in the first iteration of a spatio-temporal prediction. The colour indicates values of covariance between Fraser River and the respective catchment in units of* $\frac{\text{mm}^2}{\text{month}^2}$.

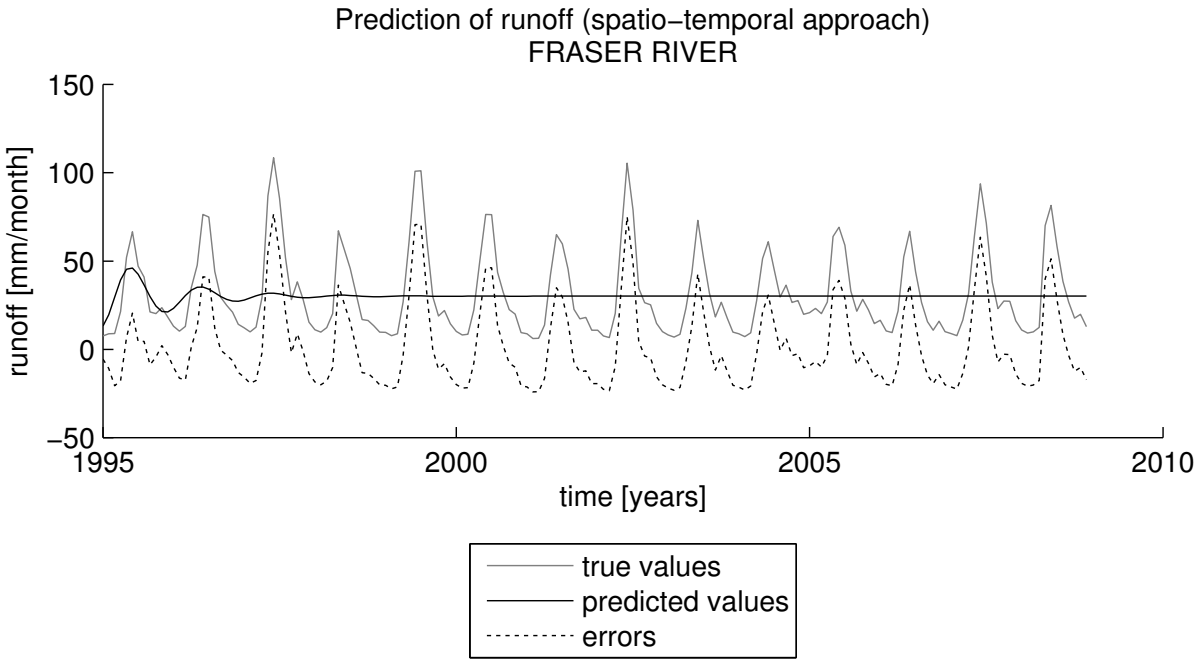

Figure 4.3: *Prediction of runoff for Fraser River with a spatio-temporal approach.* NSE $= 0.05$. *The value of the Nash-Sutcliffe coefficient* NSE *indicates a prediction which is slightly better than the mean of the true values.*

proach is more useful if data from other catchments are available.
Under realistic circumstances it is likely that there are catchments which only lack runoff val-

ues for a couple of months in their time series. For such a setting spatio-temporal prediction may prove valuable.

4.3 Spatial prediction approach

The spatial prediction approach uses different input variables for the formula of least-squares prediction (3.9) than the spatio-temporal approach. Now, the amount of runoff at one point in time is no longer predicted out of previous values, but instead the amount of runoff for one catchment is predicted out of the behaviour of other catchments. This means that one set of catchments forms the observation array $\boldsymbol{l}$ and another set of catchments forms the signal array $\boldsymbol{s}$. An iteration is not needed when using the spatial prediction approach because all the missing catchments are predicted by one calculation. It would be pointless to predict one catchment first and then use the result for the prediction of another catchment.
For this section the catchments used are fixed: Only Amazonas is predicted ($\boldsymbol{s}$) and the rest of the catchments of the sample are used as observations ($\boldsymbol{l}$). To approach the prediction technique, it is in a first step assumed that the whole data is known for the creation of the covariance matrices only. This is contrasted for each method with a more realistic approach which computes the covariance matrices out of a training period. The covariance information stored in these matrices is of course less accurate than the contents of a covariance matrix generated from the whole dataset. However, generating covariance matrices out of the whole dataset is good as a first step to test the general practicability of the spatial prediction approach and various methods. For the actual prediction that follows, the signal that is to be predicted is assumed unknown for both approaches.
When using least-squares prediction to actually find unknown values, the signal array $\boldsymbol{s}$ that is to be predicted is unknown. Therefore, covariance information cannot be generated from this part of the data. Instead, the covariance matrices are computed out of other, in most cases previous, points in time where data exists. This period of time where runoff data is available is called the training period and the unknown part is referred to as the validation period. The runoff data matrix $\boldsymbol{R}$ is split in four blocks of variable size:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{s}_{\mathrm{tr}} & \boldsymbol{l}_{\mathrm{tr}} \\ \boldsymbol{s} & \boldsymbol{l} \end{bmatrix} \tag{4.6}$$

Note that the matrix $\boldsymbol{R}$ does not necessarily have to have this block structure. For example, the unknown values can be a gap of data that is preceded, but also followed, by a set of known values. In this section, the length of the training period is fixed at $T_{\mathrm{tr}} = 120\,\mathrm{months} = 10\,\mathrm{years}$, meaning that the length of the validation period is $T_{\mathrm{val}} = 228\,\mathrm{months} = 19\,\mathrm{years}$, or the rest of the dataset. Therefore, $\boldsymbol{l}_{\mathrm{tr}}$ is a 120×24 matrix. $\boldsymbol{s}_{\mathrm{tr}}$ and $\boldsymbol{l}_{\mathrm{tr}}$ contain the data from the training period and are used for the computation of the covariances:

$$c_{jk} = \frac{1}{T_{\mathrm{tr}} - 1} \sum_{t=1}^{T_{\mathrm{tr}}} \left(\left(r_{tj} - \frac{1}{T_{\mathrm{tr}}} \sum_{\tau=1}^{T_{\mathrm{tr}}} r_{\tau j} \right) \left(r_{tk} - \frac{1}{T_{\mathrm{tr}}} \sum_{\tau=1}^{T_{\mathrm{tr}}} r_{\tau k} \right) \right) \tag{4.7}$$

For the formation of the covariance matrix $\boldsymbol{C}_{ll}$, j and k both run from 2 to 25 (all catchments except Amazon River) whereas for the formation of $\boldsymbol{C}_{sl}$, only k runs from 2 to 25 and j is always 1. The covariances are computed using only the signal arrays from the training period, therefore

the summation ends at T_{tr}. The prediction is carried out using those covariance matrices and the observation array $\boldsymbol{l}$ from the validation period. Prediction with covariance matrices from a training period operates under realistic assumptions concerning the availability of data and can be used in practical applications.

4.3.1 Method 1: Basic prediction without prearrangement

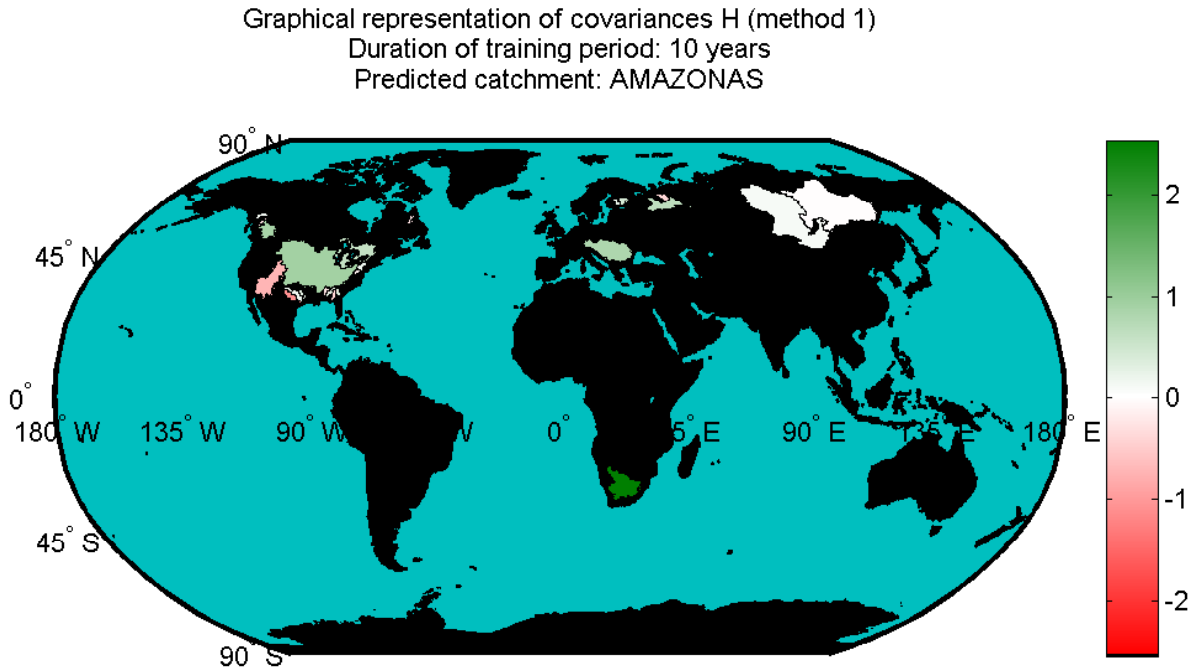

Figure 4.4: *Map showing the model matrix $\boldsymbol{H} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}$ between Amazon River and the catchments serving as observations for method 1.*

This method uses the least-squares prediction model in its simplest form. The covariances in this prediction method are computed out of all values, not according to formula 3.1, but instead using a simplified approach:

$$c_{jk} = \sum_{t=1}^{T} r_{tj} r_{tk} \tag{4.8}$$

with T being replaced by T_{tr} when the covariances are computed from a training period. The runoff data in this method are not yet normalised to zero mean, and instead the original signal is used. Also, the factor $\frac{1}{T-1}$ cancels down in the prediction formula in which a prediction for $\boldsymbol{s}$ is given by $\hat{\boldsymbol{s}} = \left(\boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}\boldsymbol{l}^{\mathsf{T}}\right)^{\mathsf{T}}$ and is therefore omitted in the calculation of the covariances already.
Figure 4.4 plots the values of the model matrix $\boldsymbol{H} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}$ generated from a training period of 10 years into the areas of the respective catchments. It shows that Amazon River is strongly influenced by Orange River, Mississippi, and catchments in western Canada whereas there is a negative correlation with catchments in the southern USA.
Figure 4.5 provides exemplary prediction results for the Amazon River catchment, plotting the

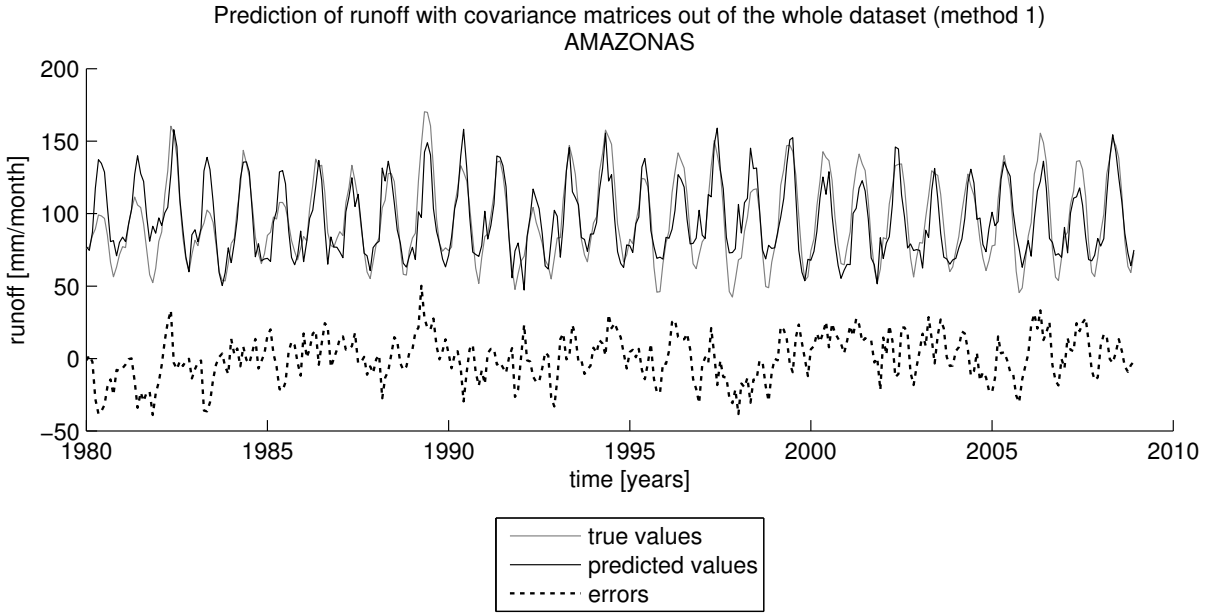

Figure 4.5: *Prediction of runoff for Amazon with covariance matrices out of the whole dataset (method 1),* $\bar{e} = 0.33 \frac{\text{mm}}{\text{month}}$, RMSE $= 15.66 \frac{\text{mm}}{\text{month}}$, NSE $= 0.70$.

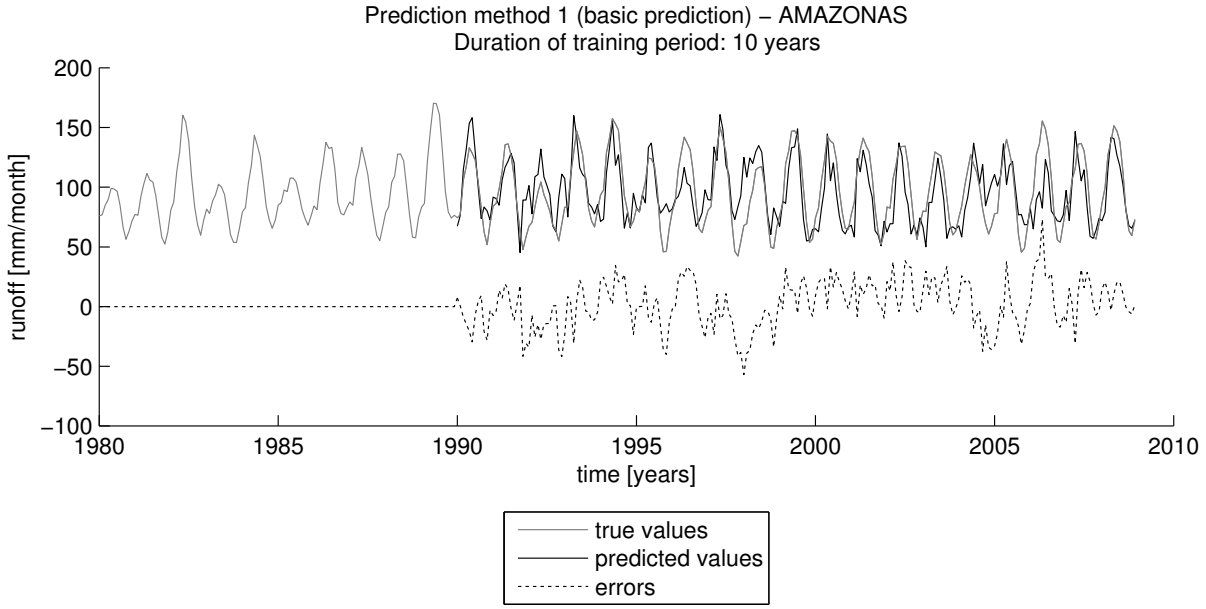

Figure 4.6: *Prediction of runoff for Amazon with covariance matrices from a training period (method 1),* $\bar{e} = 1.83 \frac{\text{mm}}{\text{month}}$, RMSE $= 20.46 \frac{\text{mm}}{\text{month}}$, NSE $= 0.52$.

arrays $\boldsymbol{s}$, $\hat{\boldsymbol{s}}$, and $\boldsymbol{e}$ (see equation 3.4). The prediction fits the general annual behaviour, but the Nash-Sutcliffe coefficient indicates that the fit is still far from perfect, especially considering the fact that there is covariance information available for the whole time. Also, the value for $\bar{e}$ deviates significantly from 0. This means that there is a bias which is what has been expected, seeing that the covariance matrices have not been normalised to zero mean.
These results are compared with a prediction which uses covariance information generated in

a training period (figure 4.6). The first ten years from 1980 to 1990 form the training period for which true runoff values s are provided. As a prediction is only carried out during the validation period, the errors e stay at zero during the training period. For the validation period, predicted runoff values $\hat{s}$ and the respective errors are also shown. It is obvious by the plot and the evaluation parameters that the prediction has a lower quality. Because of the shorter training period less information is stored in the covariance matrices and available for use in the prediction process. However, this more realistic configuration shows that the method is generally viable.

4.3.2 Method 2: Month-by-month prediction

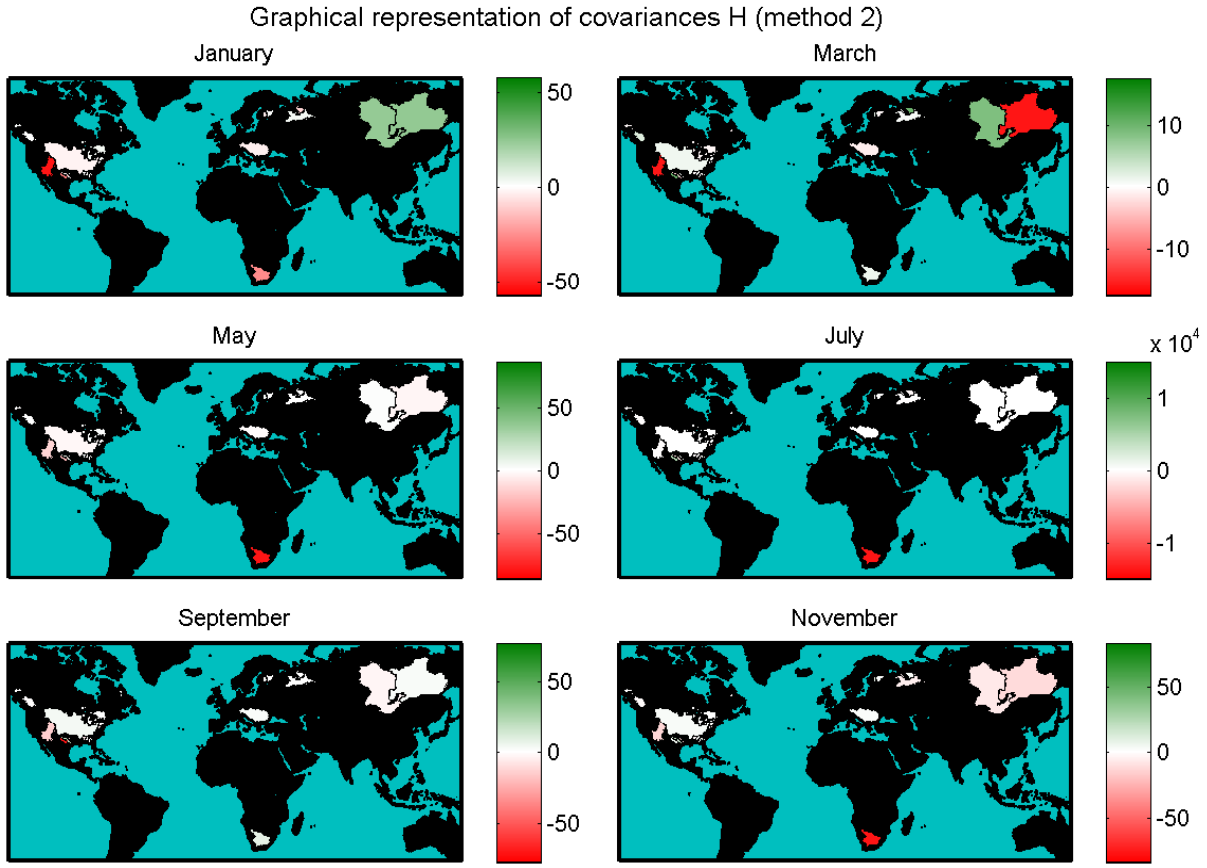

Figure 4.7: Map showing the model matrix $\boldsymbol{H} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}$ between Amazon River and the catchments serving as observations for method 2.

As runoff variation has a strongly periodic behaviour over the course of one year, the second method treats each month of the year for itself so that there are 12 separate covariance matrices. The prediction for the runoff value of one month only uses observations from that month of any year and covariance information which relates to that specific month. There is one set of covariance matrices for each month of the year, each of them following equation 4.8. This means that the prediction for one runoff value is not influenced at all by other values of that year.

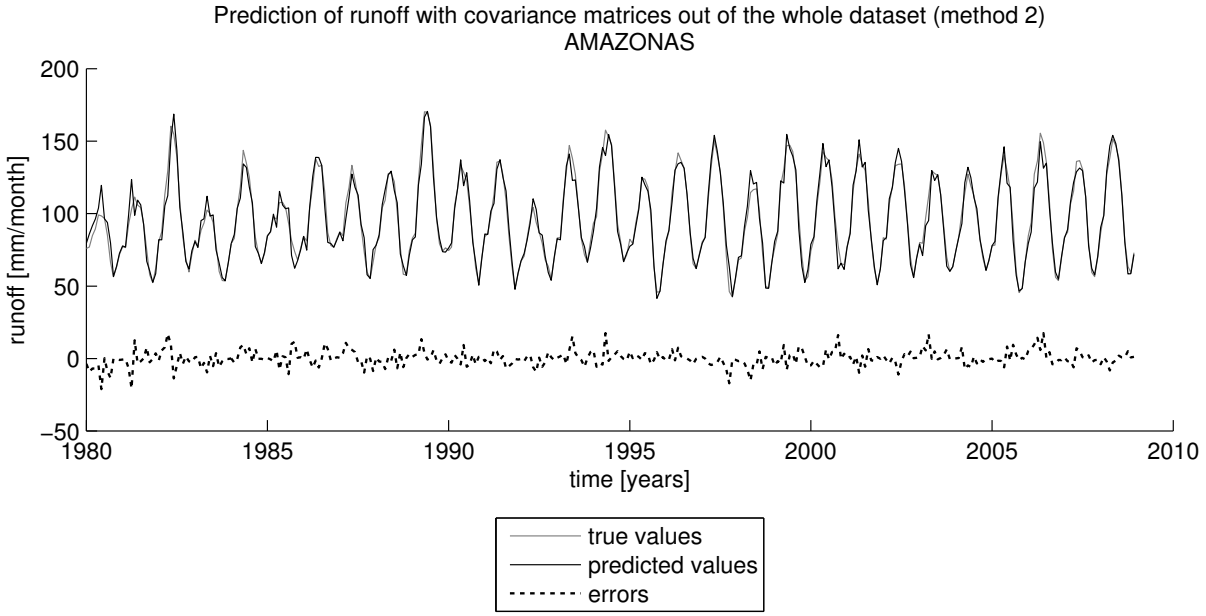

Figure 4.8: *Prediction of runoff for Amazon with covariance matrices out of the whole dataset (method 2),* $\bar{e} = 0.08 \frac{\text{mm}}{\text{month}}$, RMSE $= 5.36 \frac{\text{mm}}{\text{month}}$, NSE $= 0.96$.

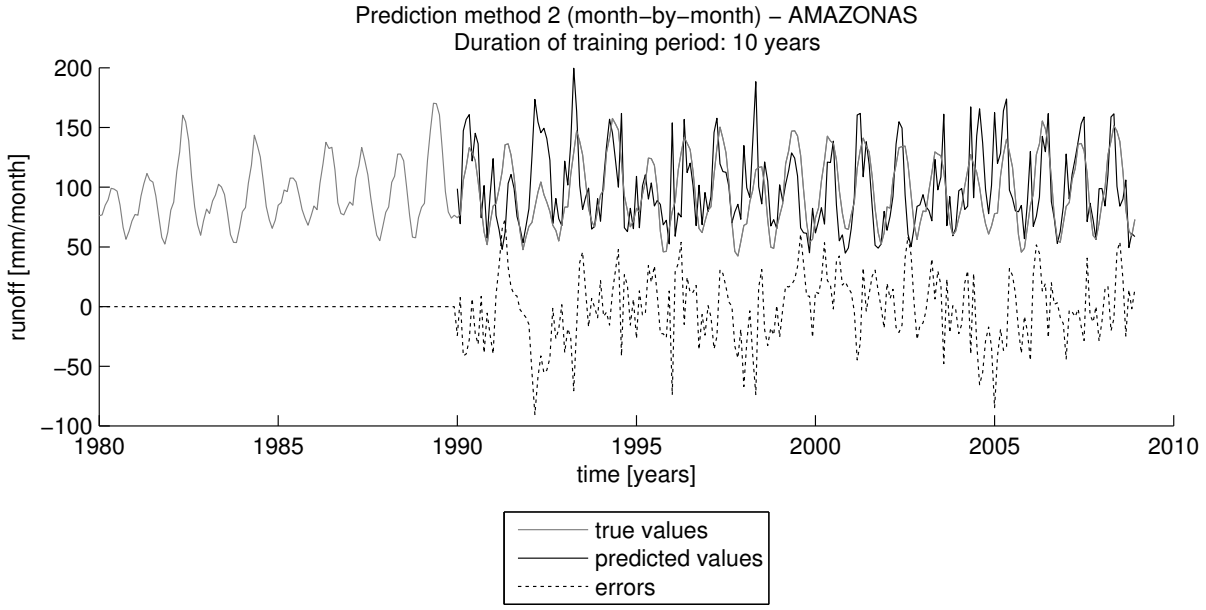

Figure 4.9: *Prediction of runoff for Amazon with covariance matrices from a training period (method 2),* $\bar{e} = -2.61 \frac{\text{mm}}{\text{month}}$, RMSE $= 29.92 \frac{\text{mm}}{\text{month}}$, NSE $= -0.02$.

Figure 4.7 shows 6 of the 12 covariance models $\boldsymbol{H}$. The values are very different for some of the months. For most of the year, there is a strong negative influence of Orange River, but in winter, the lowest value is reached for Colorado River (Pacific Ocean). The Yenisei catchment shows a positive behaviour in January, but a negative in March. The absolute covariance values also differ by a factor of 1000 between March and July. These differences may be the reason for the very good prediction quality indicated by the Nash-Sutcliffe coefficient of NSE $= 0.96$

(figure 4.8), seeing that these variations are not considered in the first prediction method.
There are 12 prediction processes, one for each month of the year. After all the predictions $\hat{s}$ are computed, these are joint and compared with the true values. Figure 4.8 provides exemplary prediction results. This method still has a small bias like the first method does, but the quality of the prediction is a lot better. This can be seen in the plot and in the better results for RMSE and NSE.
However, the inversion of the covariance matrix C_{ll} is a problem. When the amount of observed catchments is bigger than the duration of the training period in years, C_{ll} becomes singular and therefore the computation of an inverse becomes problematic. This is not the case in the example where C_{ll} is computed out of the whole dataset because there are 29 years of training period and 24 observed catchments. In more realistic approaches, these relations can easily change though, making this method exclusive to certain beneficial configurations. With a training period of 10 years (figure 4.9) and 24 observed catchments, C_{ll} is not invertible any more. Instead, the pseudoinverse is computed for the prediction:

$$\hat{s}^{\mathsf{T}} = C_{sl} C_{ll}^{+} I^{\mathsf{T}} \tag{4.9}$$

This computation delivers a prediction which is worse than the mean of the runoff curve (NSE $<$ 0). The errors are large and vary strongly. For these practical limitations, method 2 is not analysed any further in this study. It may be interesting to note that this problem can in fact occur for the other spatial prediction methods as well, but then the number of *months* has to be surpassed by the amount of catchments so that the resulting limitations are bearable.

4.3.3 Method 3: Prediction out of monthly residuals

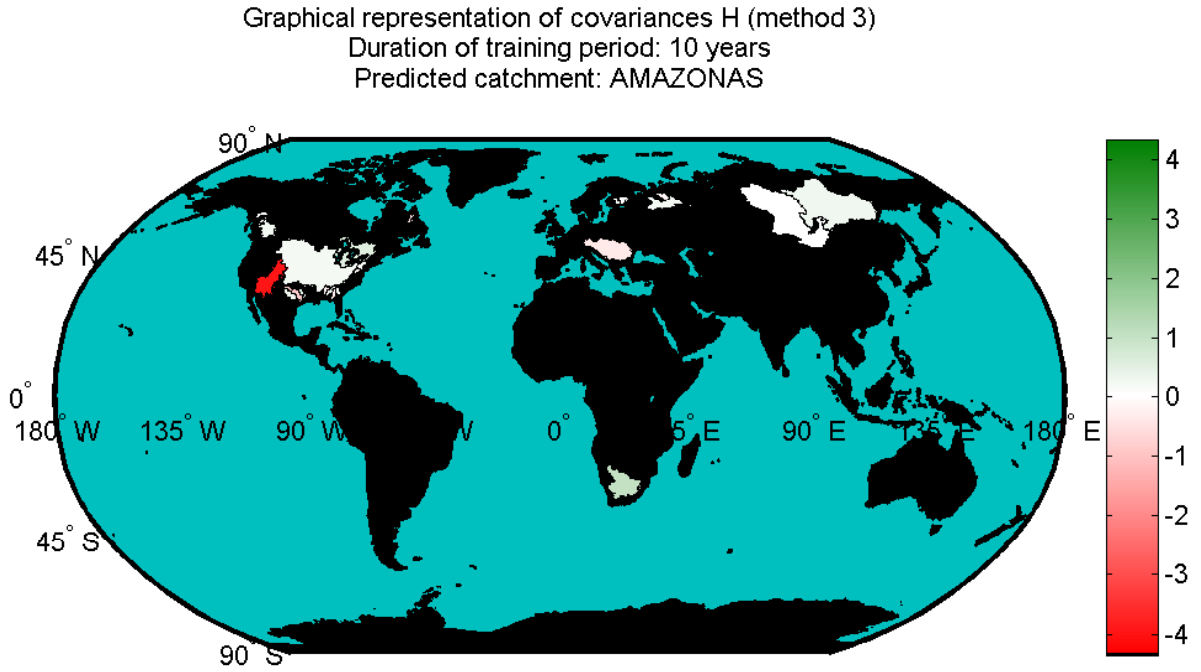

Figure 4.10: *Map showing the model matrix $H = C_{sl} C_{ll}^{-1}$ between Amazon River and the catchments serving as observations for method 3.*

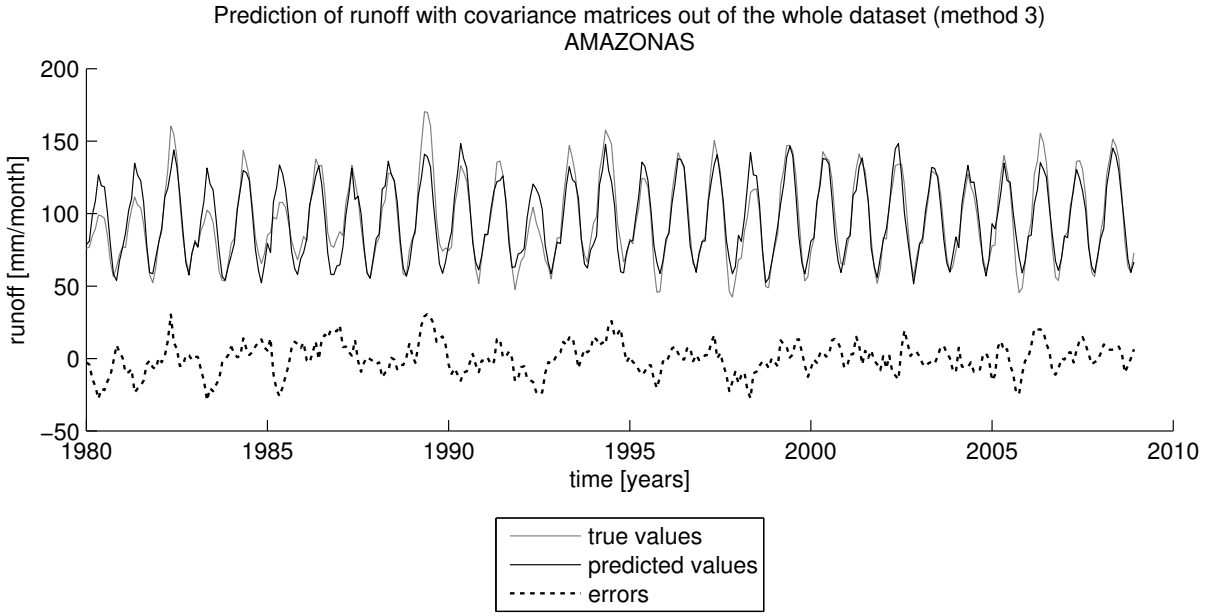

Figure 4.11: *Prediction of runoff for Amazon with covariance matrices out of the whole dataset (method 3),* $|\bar{e}| < 10^{-14}\,\frac{\text{mm}}{\text{month}}$, $\text{RMSE} = 10.93\frac{\text{mm}}{\text{month}}$, $\text{NSE} = 0.85$, $\widetilde{\text{NSE}} = 0.21$.

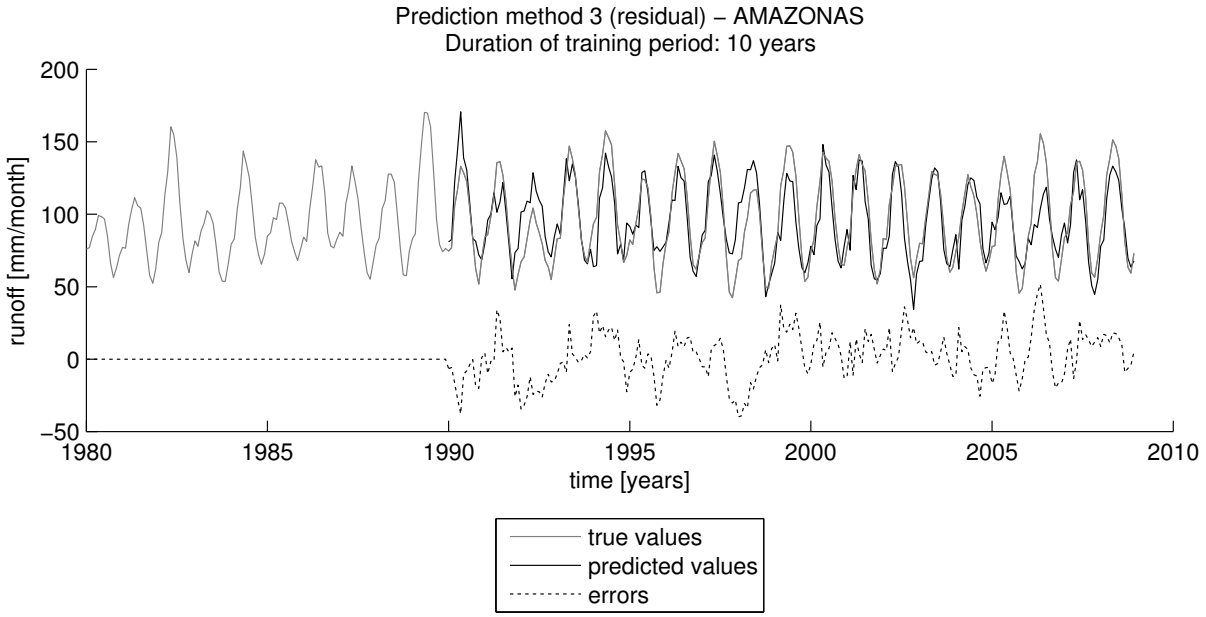

Figure 4.12: *Prediction of runoff for Amazon with covariance matrices from a training period (method 3),* $\bar{e} = 1.67\frac{\text{mm}}{\text{month}}$, $\text{RMSE} = 16.60\frac{\text{mm}}{\text{month}}$, $\text{NSE} = 0.69$, $\widetilde{\text{NSE}} = -1.45$.

The third method is based on the assumption that the high total runoff values strain the prediction results. Therefore all data is reduced by the monthly mean, only leaving residuals $\boldsymbol{R}_{\text{res}}$ to be used in the prediction process:

$$\boldsymbol{R}_{\text{res}} = \boldsymbol{R} - \bar{\boldsymbol{R}}_{\text{cyc}} \tag{4.10}$$

$\bar{\boldsymbol{R}}_{\text{cyc}}$ is the mean of the runoff for each month of the year, also called the monthly mean, and has to be computed out of either the training period, if applicable, or the whole dataset if this is assumed known for the creation of the covariances. After the subtraction of the monthly mean there are only the residuals left which are consequently used to form the covariance matrices and the signal arrays $\boldsymbol{l}$ and $\boldsymbol{s}$. The covariance matrices are formed out of their coefficients

$$c_{jk} = \sum_{t=1}^{T} r_{tj,\text{res}} r_{tk,\text{res}} \ , \tag{4.11}$$

again replacing T by T_{tr} if the covariance matrices are to be generated from a training period. The monthly mean is added again after the prediction. Then, visualisation and computation of error parameters can take place.
As the residuals of the runoff $\boldsymbol{R}_{\text{res}}$ can become negative, the correlation matrices may also contain negative values using this method. In fact, figure 4.10 shows that a large negative covariance value is reached with Colorado River (Pacific Ocean) and the other catchments are all centred around zero.
Figure 4.11 provides exemplary prediction results for one catchment. The resulting curve is a lot smoother here than in the first method. At the same time, the quality lies between those of method 1 and 2 and there is no more bias because residuals are normalised to zero mean. This method also features a value for the residual Nash-Sutcliffe coefficient as described in formula 4.3 which rates the prediction comparing it to the fit achieved by using the monthly mean. This comparison is useful in this case as the monthly mean is the base for the prediction method. The coefficient's value denotes a model quality which is only slightly better than the monthly mean. However, this is not very surprising because the predicted catchment is in a geographically isolated position in the sample dataset (see figure 2.2). High cyclostationarity ($\gamma = 0.88$ for Amazon) also means that the monthly mean already is a very good fit which is harder to surpass than the monthly mean of a catchment with low cyclostationarity.
The more realistic example generating the covariance matrices from a training period (figure 4.12) again provides prediction results which are slightly worse than when the covariance matrices are out of the whole dataset, similar to the behaviour observed when using method 1.

4.3.4 Method 4: Prediction out of zero-mean time series

As it has been shown that a subtraction of the monthly mean before starting the prediction is quite useful, the fourth method tries to decrease the runoff matrix $\boldsymbol{R}$ by the long term mean of each time series. All the parameters needed for the prediction are generated from this matrix. The covariances are generated from their coefficients

$$c_{jk} = \sum_{t=1}^{T} (r_{tj} - \bar{r}_j)(r_{tk} - \bar{r}_k) \tag{4.12}$$

where T and the way the means $\bar{r}_j$ and $\bar{r}_k$ are computed again have to be adjusted to whether a training period or the whole dataset is used for the formation of the covariance matrices. Once again, after the prediction, the mean is added.
A comparison of the cross-covariance matrix $\boldsymbol{C}_{sl}$ after a training period of 10 years (figure 4.13) with the matrix generated out of residual values as in method 3 (figure 4.10) depicts great

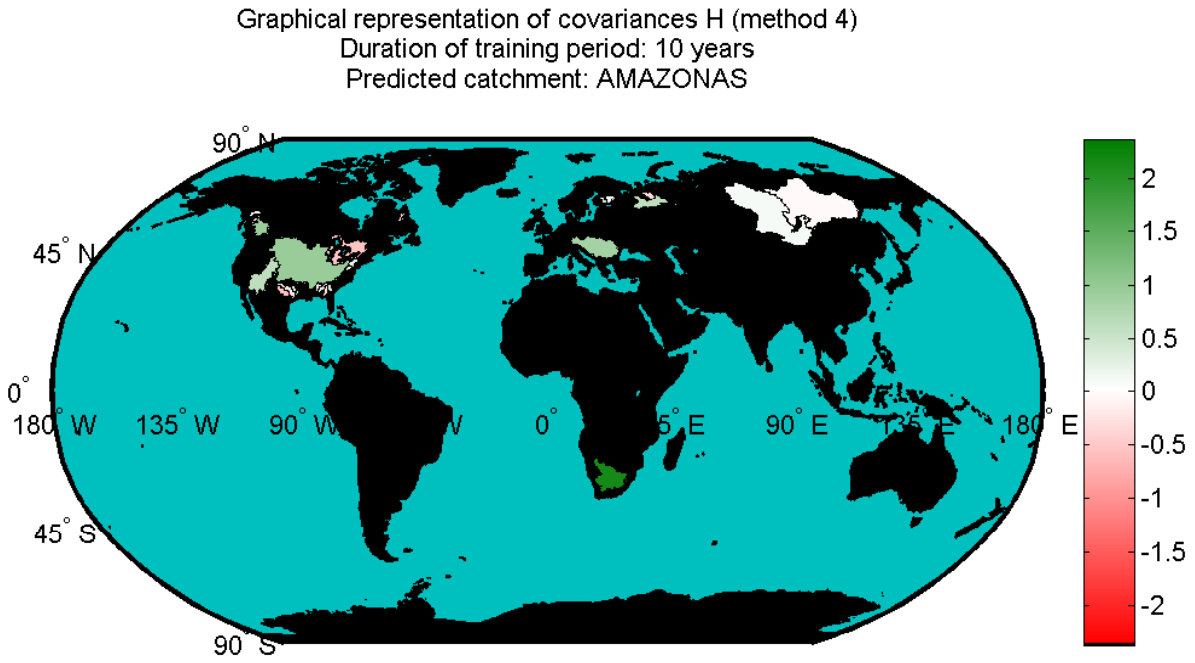

Figure 4.13: *Map showing the model matrix* $\boldsymbol{H} = \boldsymbol{C}_{sl}\boldsymbol{C}_{ll}^{-1}$ *between Amazon River and the catchments serving as observations for method 4.*

differences between the two matrices. For method 4 there is a big positive value on Orange River and the values in Northern America oscillate into the positive and negative realms. This example shows that the differences in the prediction quality between the various methods are already conceivable by observing their covariance matrices.
Figure 4.14 shows that the results of this method are very similar to those of the first method. This is not surprising because the subtraction of the mean is the only difference between the two methods. It causes the mean error to become 0 as in method 3 which in consequence also improves the model quality of method 4 slightly compared to method 1. However, it is a lot worse than the model quality of method 3 for the exemplary catchment.
When covariance matrices are generated from a training period (figure 4.15), the results are also very similar to those of the first method.

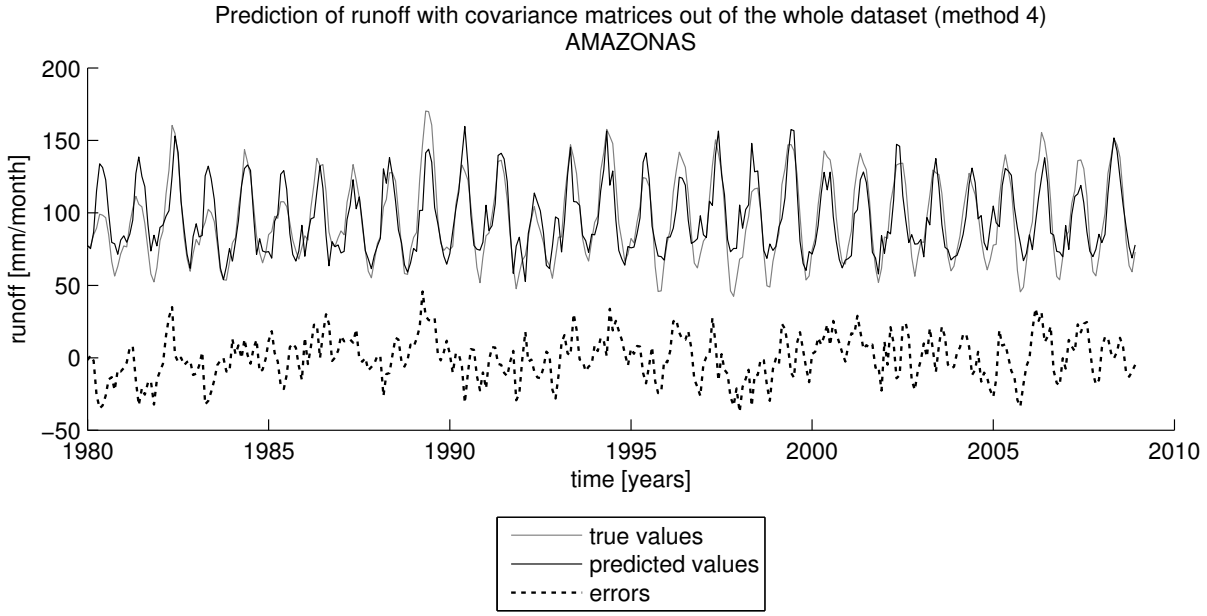

Figure 4.14: *Prediction of runoff for Amazon with covariance matrices out of the whole dataset (method 4),* $|\bar{e}| < 10^{-13}\frac{\text{mm}}{\text{month}}$, RMSE $= 15.28\frac{\text{mm}}{\text{month}}$, NSE $= 0.72$.

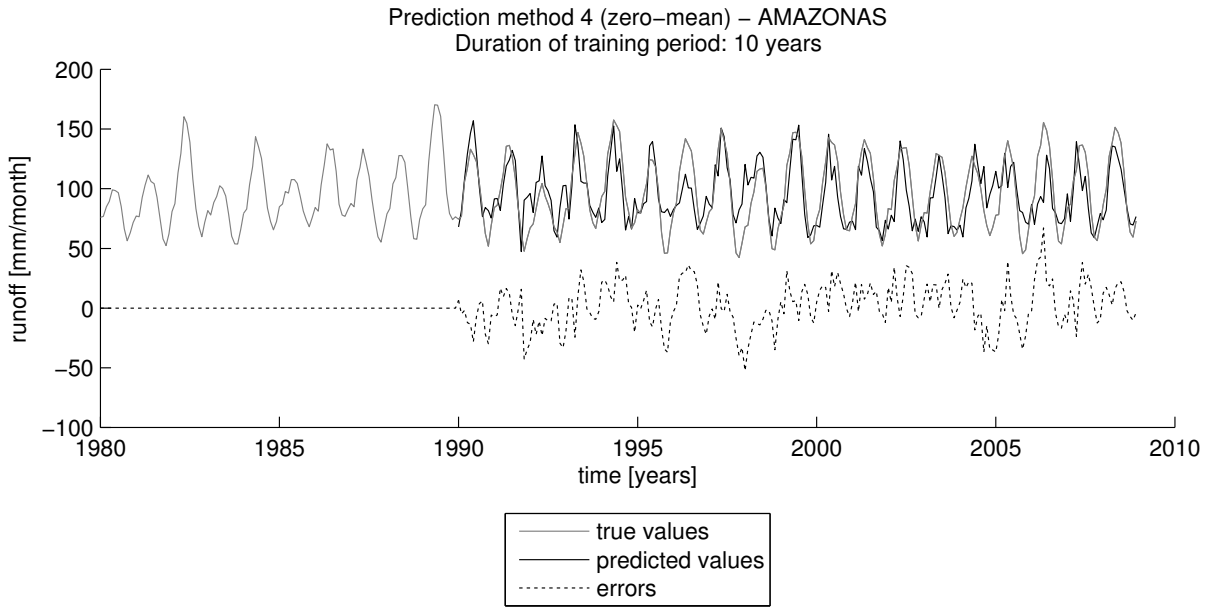

Figure 4.15: *Prediction of runoff for Amazon with covariance matrices from a training period (method 4),* $\bar{e} = 1.81\frac{\text{mm}}{\text{month}}$, RMSE $= 19.39\frac{\text{mm}}{\text{month}}$, NSE $= 0.57$.

Chapter 5

Validation

The prediction with covariance information from the whole dataset has shown that prediction of runoff is generally possible and viable under optimal conditions for all presented spatial methods. Under more realistic circumstances different results are achieved which are slightly worse. It is now interesting to analyse the efficiency of different configurations, mainly including the variation of the length of the training period and the set up of catchments which are used as observations. As investigations only take place in the sample dataset, the results can be validated and compared to determine which conditions prove favourable.
For this validation, only methods 3 and 4 will be used. The spatio-temporal prediction approach is not suitable for the lacking availability of control data, the first spatial method is only a biased version of method 4 without additional benefits, and the second method drops out because of the aforementioned inversion problems.

5.1 Duration of the training period

When generating covariance information from a training period for the prediction of runoff, the length of this training period is an important factor for the quality of the prediction. The assumption is made that a longer training period benefits the quality.
Figure 5.1 shows a prediction for the runoff of Amazon in the years 2000 until 2008 under optimal conditions: All other catchments in the sample dataset are used as observations and all of the previous time since 1980 is used as a training period, meaning that $\boldsymbol{l}_{\text{tr}}$ from equation 4.6 is a 240×24 matrix. As $\boldsymbol{l}_{\text{tr}}$ therefore contains a lot of data, the covariance matrices can also use a lot of information. In figure 5.2 the amount of catchments stays the same, but only the years 1980 until 1984 are used for the formation of the covariance matrices. These contain less information and therefore the prediction becomes worse as indicated by a lower value of the Nash-Sutcliffe coefficient.
Figure 5.3 directly shows the relationship between the length of the training period and the resulting value for NSE. It can be seen that a prediction only becomes viable after a minimum length of 3 years for the training period and after that a longer training period increases the model efficiency. This behaviour does not only occur with Amazon, but with most catchments from the sample dataset.
It is also useful to not only regard a change in the duration of the training period, but rather a simultaneous increase in the length of the training period and decrease in the length of the validation period. Until now, the length of the validation period has been set to a constant value of 9 years which means that depending on the duration of the training period certain values from the time series are neither used for training nor for validation. This is an unlikely scenario

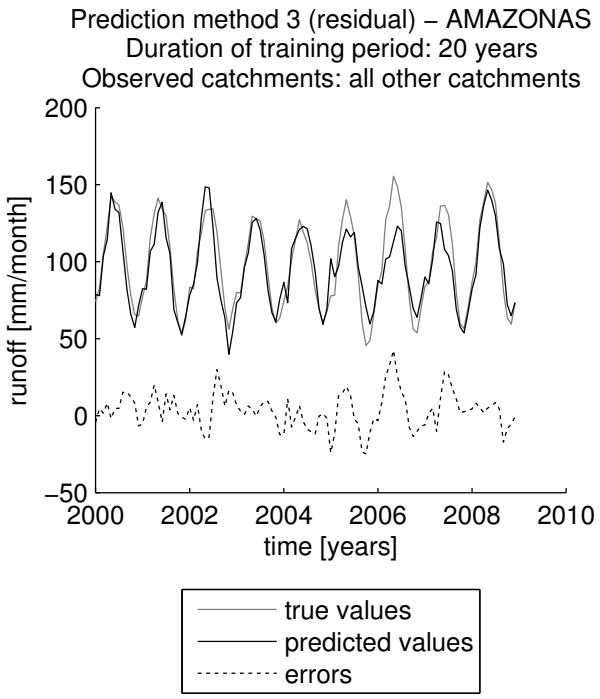

Figure 5.1: Prediction for Amazon with 20 years of training period, NSE = 0.82.

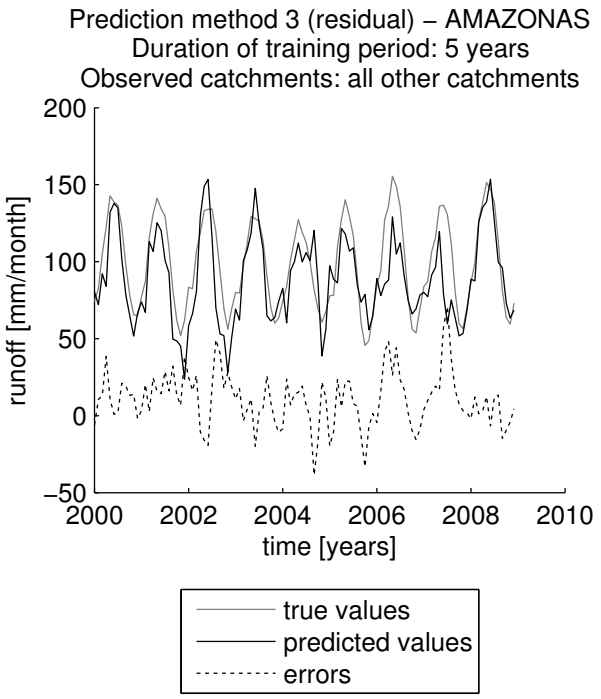

Figure 5.2: Prediction for Amazon with 5 years of training period, NSE = 0.50.

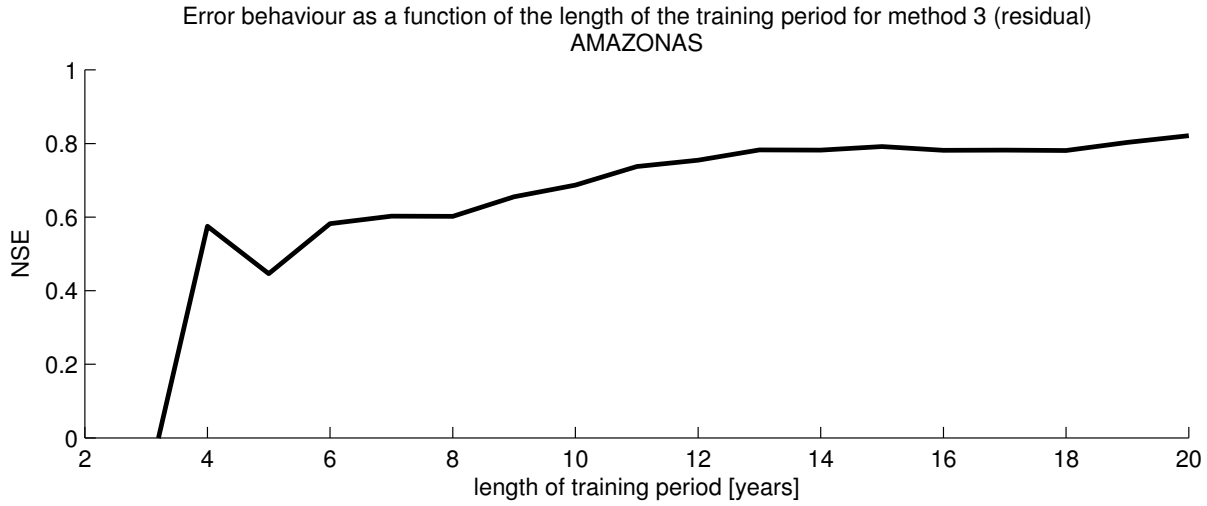

Figure 5.3: Relationship between the length of the training period and the Nash-Sutcliffe coefficient for Amazon. The length of the validation period is constantly at 9 years (2000-2008). The training period always begins in 1980.

because the length of the training period should be maximised in order to optimise the results. However, the Nash-Sutcliffe coefficient becomes more prone to small irregularities as the length of the validation period decreases. For very short validation periods, the small sample size of the predicted values causes a low reliability of the resulting model efficiency. This behaviour can be observed for validation periods which are shorter than 4 years (see figure 5.4) and has to be considered in all future applications of analyses comparing the length of the training period

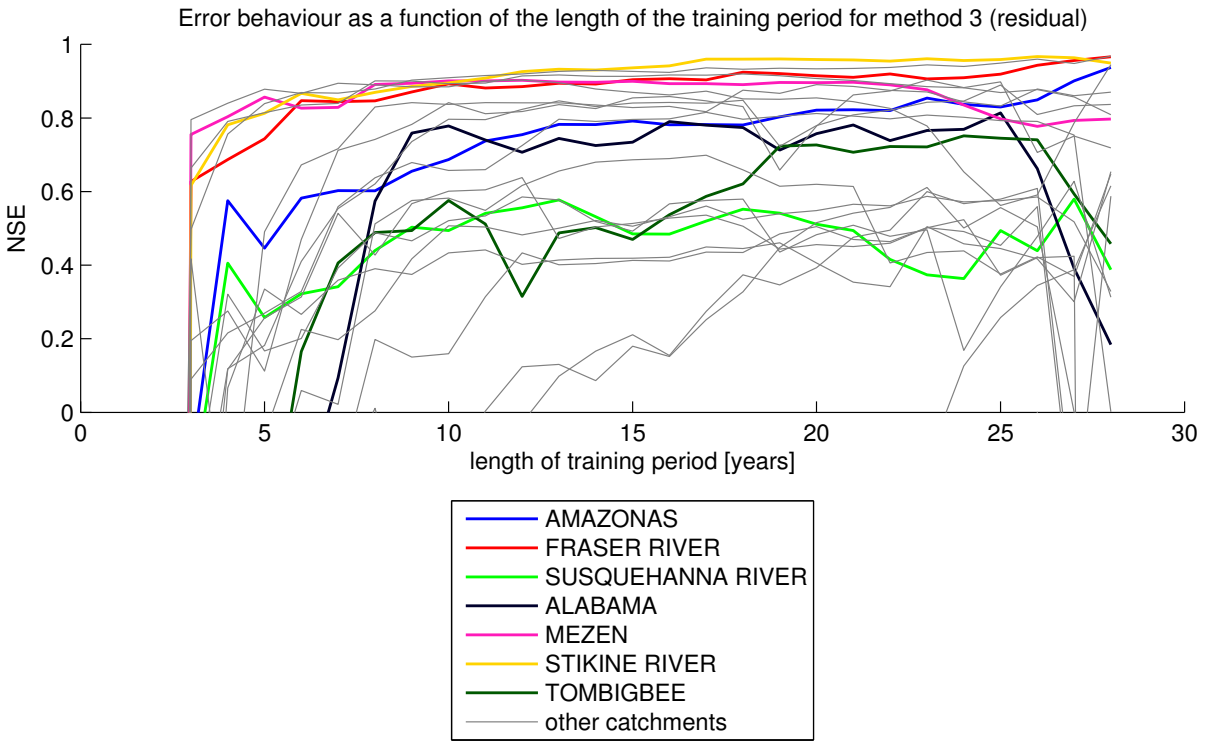

Figure 5.4: *Relationship between the length of the training period and the Nash-Sutcliffe coefficient for all catchments from the sample dataset, represented by the differently coloured lines, predicted by method 3. Colorado River (Pacific Ocean) and Orange River are not featured in this plot because* NSE *is negative for all possible lengths of the training period. These two catchments are very hard to predict for their lack of cyclostationarity (see table 2.2) and therefore do not provide good results. The training period always begins in 1980 and the validation period starts directly afterwards and lasts until 2008.*

to an error coefficient.
This figure also shows that for some catchments viable predictions are already possible after a training period of only 3 years. Generally, the quality of a prediction does not increase meaningfully any more when the training period becomes longer than 10 years. The amount of information stored in correlation matrices generated from a training period of 3 to 10 years of length, depending on the catchment, is sufficient to carry out a prediction. More data usually does not improve the quality above a certain level which lies at a value above NSE $= 0.4$ for about 80% of the catchments and above NSE $= 0.8$ for about half of the catchments in the dataset.
All of these computations are a result of the application of the third method: Prediction out of monthly residuals, as presented in chapter 4.3.3. Figure 5.5 and also table 5.1 show that for method 4, the results are of a similar, although slightly lower quality. It can also be seen that a high cyclostationarity factor often causes negative values for $\widetilde{\text{NSE}}$ because the monthly mean already is a good fit for catchments with a high cyclostationarity. For about two thirds of the catchments in the dataset, after a training period of 15 years method 3 predicts the runoff more accurately than the monthly mean ($\text{NSE}_3 > 0$). While method 4 may sometimes yield better results than method 3, the difference is usually very small in these cases, as indicated by figure 5.6. The $\widetilde{\text{NSE}}$ is never considerably better for method 4. However, for those catchments favouring a prediction using method 3, there is often a vast difference between the two predic-

Table 5.1: *Cyclostationarity factor γ and error parameters: Nash-Sutcliffe coefficient for methods 3 and 4 (NSE_3, NSE_4) and the respective Nash-Sutcliffe coefficients with respect to the monthly mean $\widetilde{NSE}$, as defined by equation 4.3. Length of training period: 15 years; length of validation period: 14 years.*

Catchment	γ	NSE_3	NSE_4	$\widetilde{NSE}_3$	$\widetilde{NSE}_4$
AMAZONAS	0.88	0.79	0.63	−1.21	−2.96
MISSISSIPPI RIVER	0.67	0.51	0.48	0.04	−0.01
YENISEI	0.78	0.83	0.77	−0.22	−0.65
LENA	0.81	0.91	0.87	−0.58	−1.41
ST LAWRENCE	0.88	−0.25	−0.20	−0.65	−0.58
ORANGE	0.13	−1.09	−0.85	−1.30	−1.04
DANUBE	0.74	0.41	0.18	0.04	−0.33
COLORADO RIVER (PACIFIC OCEAN)	0.17	−39.65	−35.97	−40.41	−36.66
NORTHERN DVINA	0.71	0.87	0.86	0.26	0.21
FRASER RIVER	0.79	0.90	0.89	0.30	0.19
COLORADO RIVER (CARIBBEAN SEA)	0.19	0.51	0.49	0.47	0.44
BRAZOS RIVER	0.26	0.80	0.83	0.78	0.82
SUSQUEHANNA RIVER	0.51	0.48	0.53	0.17	0.24
VUOKSI	0.78	0.21	0.22	0.18	0.18
APALACHICOLA RIVER	0.55	0.85	0.84	0.80	0.78
ALABAMA	0.49	0.73	0.76	0.59	0.63
MEZEN	0.71	0.90	0.86	0.20	−0.10
STIKINE RIVER	0.81	0.94	0.80	−0.12	−2.51
TOMBIGBEE	0.46	0.47	0.45	−0.05	−0.10
TRINITY RIVER (TEXAS)	0.33	0.69	0.71	0.62	0.64
GEORGE RIVER	0.68	0.42	0.43	−0.11	−0.09
ALTAMAHA RIVER	0.45	0.84	0.84	0.77	0.77
KYMIJOKI	0.70	0.18	0.24	0.08	0.15
SKEENA RIVER	0.75	0.93	0.87	0.48	0.05
POTOMAC RIVER	0.43	0.49	0.55	0.33	0.41

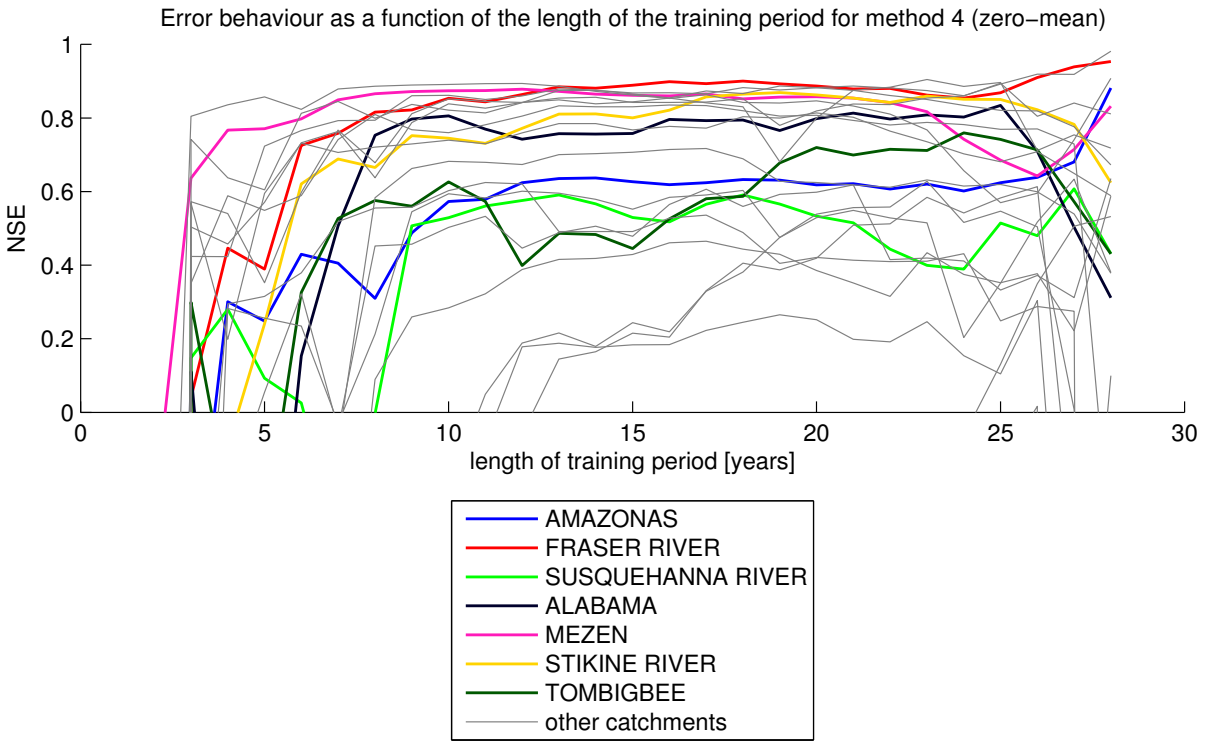

Figure 5.5: *Relationship between the length of the training period and the Nash-Sutcliffe coefficient for all catchments from the sample dataset, represented by the differently coloured lines, predicted by method 4. Colorado River (Pacific Ocean) is not featured in this plot because* NSE *is negative for all possible lengths of the training period. The training period always begins in 1980 and the validation period starts directly afterwards and lasts until 2008.*

tion methods.
Some special attention needs to be paid to the Colorado River (Pacific Ocean) catchment which scores very bad results in table 5.1. The values of the various Nash-Sutcliffe coefficients all indicate that a prediction is about 40 times worse than any form of the mean value of runoff for the catchment. A prediction does not yield any feasible results at all in this case. When looking at the runoff data from Colorado River (figure 5.7) one can see that it does not follow a regular behaviour, but instead it is dominated by some major and long-lasting events, probably caused by human interaction. As these events likely only affect this catchment and definitely are not found in the runoff time series of other analysed catchments, no correlation information is able to predict these irregularities, in contrast to events with atmospheric origin, which usually affect not only one catchment, but a whole group of catchments. For these reasons predicting the runoff during the 2000s out of covariances generated from the 1980s is an erratic process resulting in a very bad model quality and a very low Nash-Sutcliffe coefficient. However, predicting the 2000s from a training period between 1988 and 1992 might be perfectly viable, but this problem is not enlarged upon in this study.

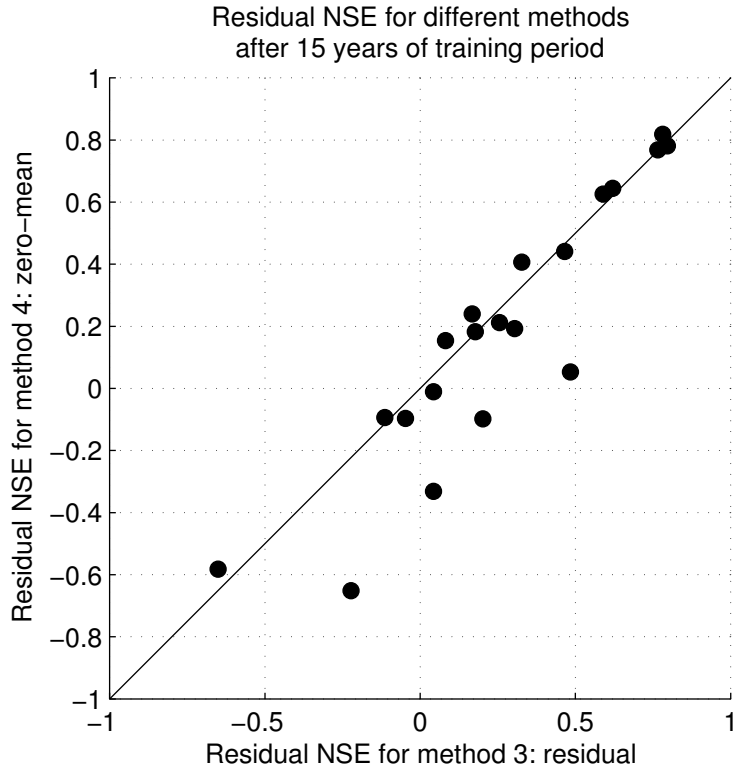

Figure 5.6: *Comparison of the residual Nash-Sutcliffe coefficient* $\widetilde{\mathrm{NSE}}$ *for methods 3 and 4. Only catchments with* $\widetilde{\mathrm{NSE}} > -1$ *for both methods are shown.*

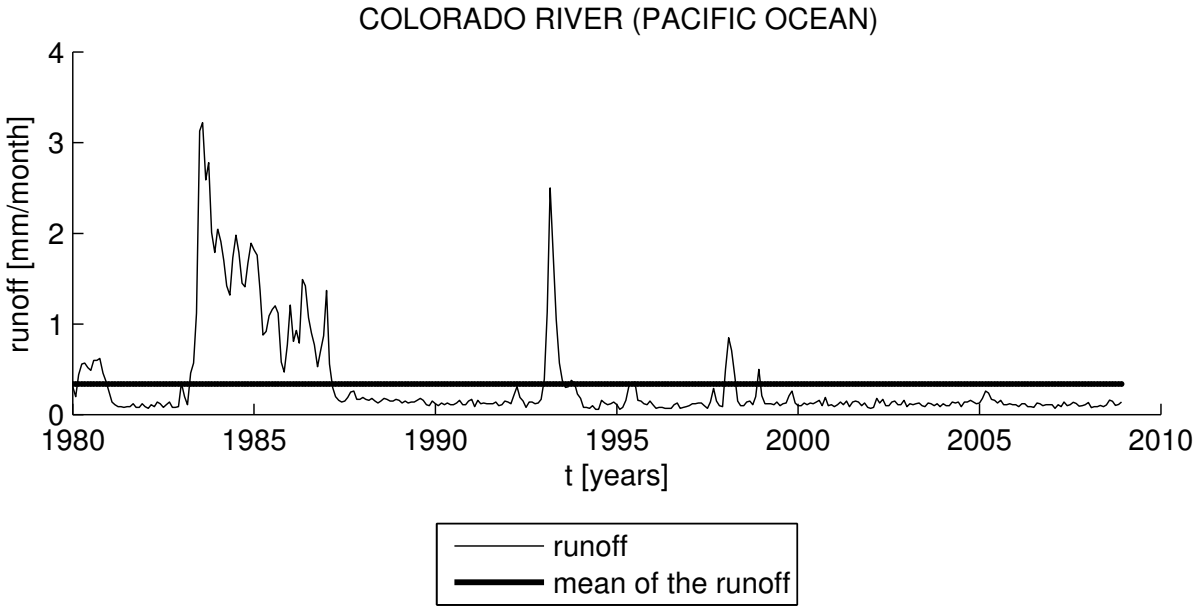

Figure 5.7: *Runoff for Colorado River (Pacific Ocean).*

5.2 Catchments used as observations

Until now, all prediction processes presented in this study have been using the maximum amount of catchments as observations. Out of the sample dataset of 25 catchments, 24 have

been observations and one has been predicted. In the following, the impact of the configuration of catchments chosen as observation is analysed.
Changing the amount of observed catchments means that $\boldsymbol{l}_{tr}$ (equation 4.6) becomes a $T_{tr} \times n_{\text{obs}}$ matrix and $\boldsymbol{l}$ becomes a $T_{val} \times n_{\text{obs}}$ matrix correspondingly, where n_{obs} is the number of catchments serving as observations. Not only variations in the amount of observations are analysed, but also the impact of the decision which catchments are used for this purpose. To access this problem systematically the correlation (see section 3.1.3 and figure 3.1) between catchments is used. The relationship between correlation and distance can also be analysed. To calculate the distance d between catchments 1 and 2, the spherical cosine formula (Todhunter, 1871) and the location information provided in table 2.1 are used:

$$d = \arccos(\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos(\lambda_1 - \lambda_2)\ \cos\phi_2)\ R_E \tag{5.1}$$

where $R_E = 6370\,\text{km}$ is used as the radius of the Earth.

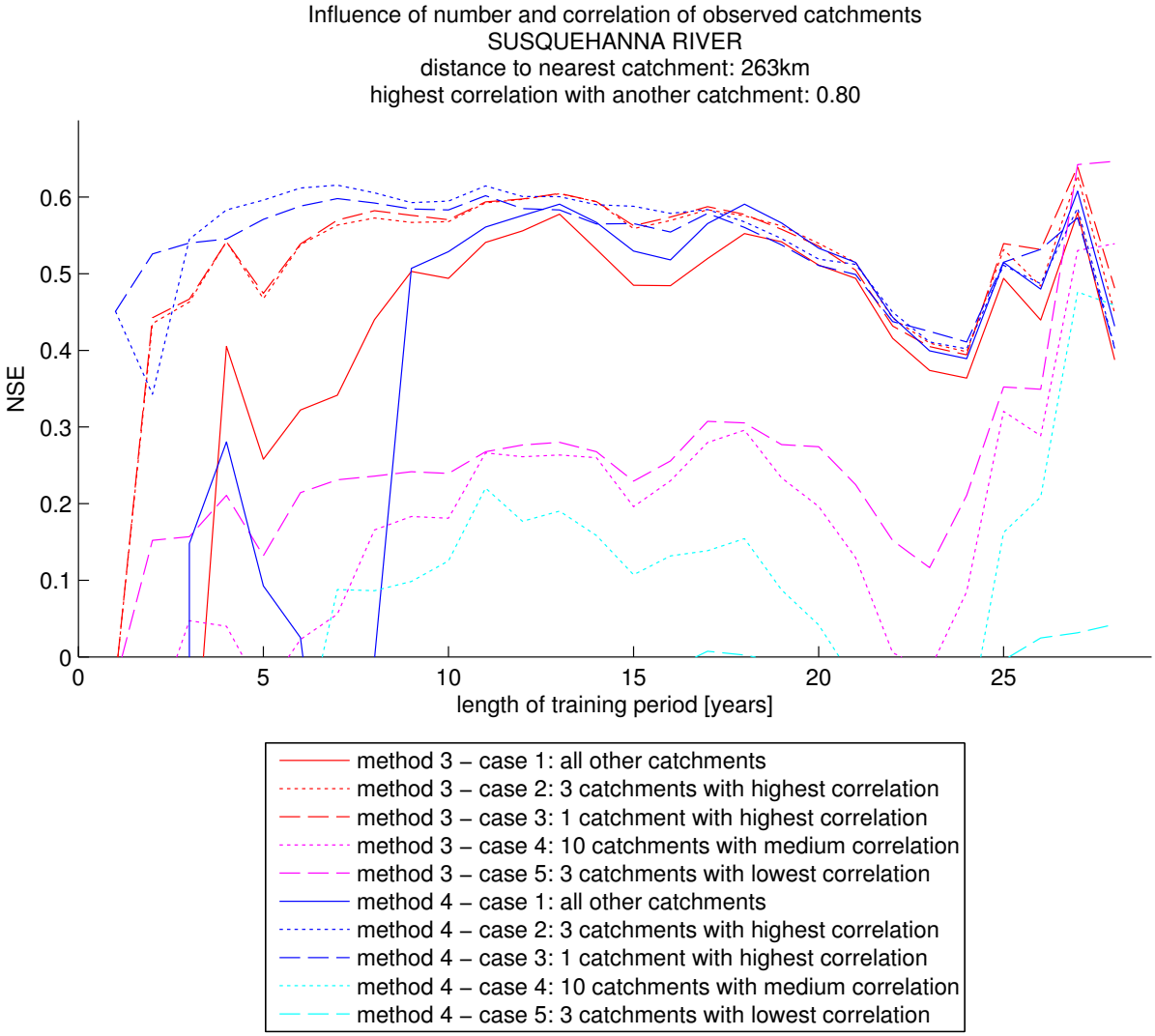

Figure 5.8: *Susquehanna River: Nash-Sutcliffe coefficient as a function of the length of the training period for different configurations of observed catchments and both analysed prediction methods.*

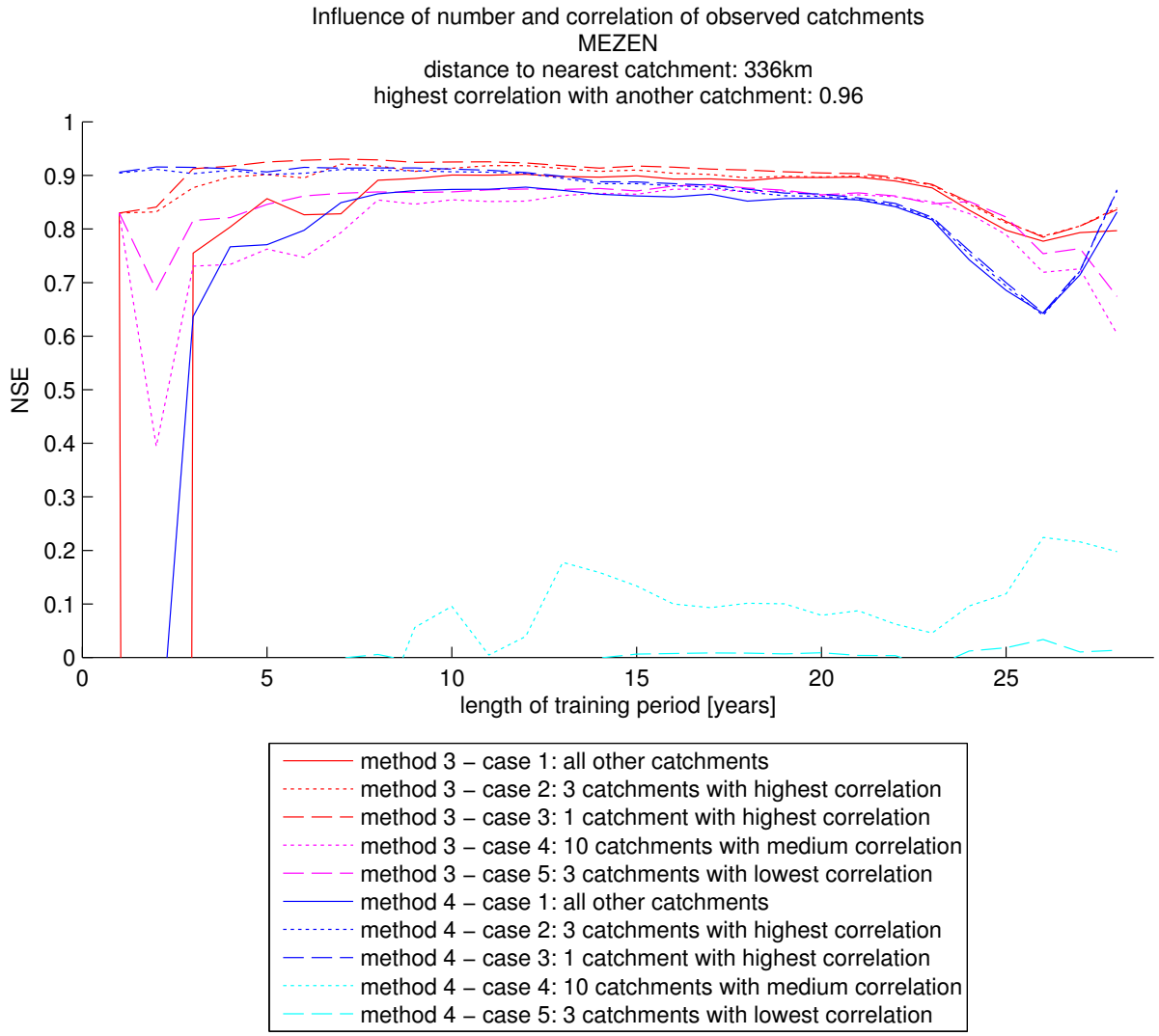

Figure 5.9: *Mezen: Nash-Sutcliffe coefficient as a function of the length of the training period for different configurations of observed catchments and both analysed prediction methods.*

For all the catchments in the sample dataset, the correlation with the examined catchment, as a first example Susquehanna River (figure 5.8), is computed. Then the catchments are sorted from highest to lowest correlation using the absolute values because the assumption is made that predicting a catchment out of observations which have a correlation value of about 0 is the most difficult. As a first step, only the 3 catchments which have the highest correlation values with respect to the predicted catchment are used as observations (case 2). The result of this prediction is rather surprising: Its efficiency is clearly greater for short training periods, even though the amount of information used in the prediction process is decreasing, compared to the usage of all available catchments as applied in this study until now (case 1). In fact, method 4 can already reach a Nash-Sutcliffe coefficient of over 0.4 after only 1 year of training period with 3 catchments observed, while it needs 9 years to do so with 24 catchments observed. Prediction using all other catchments as observations has previously not yielded any feasible results with training periods shorter than 3 years. The catchments with the highest correlations are Potomac River, Mississippi River, and Altamaha River. When only Potomac River

with a correlation of 0.80 is used as observation (case 3), the quality of the resulting prediction is still about similar. Essentially, one observed catchment is enough to predict the runoff of Susquehanna River. Still, the quality never surpasses a value of NSE $= 0.7$ which indicates a catchment that is rather hard to predict. However, the prediction quality can become worse than this: When only ten catchments from the middle of the list sorted by correlation (case 4) or even the 3 least correlated catchments (case 5) are chosen as observations, the Nash-Sutcliffe coefficient rarely surpasses 0.3, and in the latter case even struggles to rise over 0 with method 4. However, from Susquehanna River no clear preference for one of the two prediction methods (prediction from residuals, see section 4.3.3; and prediction from zero-mean time series, see section 4.3.4) can be deduced. The behaviour of Susquehanna River is very characteristic for most catchments which reach a low to medium amount of cyclostationarity ($\gamma < 0.6$) but have a high correlation with some other catchments. In the sample dataset, most of these catchments are located in the eastern part of North America.

Looking at Mezen (figure 5.9) it becomes clear that method 3 is very superior to method 4 when it comes to prediction with poorly correlated observations. While method 3 can model the runoff for both case 4 and case 5 nearly as well as for any case, reaching values for the Nash-Sutcliffe coefficient averaging around 0.85 for most of the training durations, method 4 can barely surpass the arithmetic mean's efficiency ($\text{NSE}_4 \approx 0$). Again, predictions using fewer, but highly correlated observations (cases 2 and 3) prove more successful than predictions using all other catchments (case 1), especially when only very short training periods are granted.

Mezen reaches the highest correlation value of 0.96 with its neighbouring catchment, Northern Dvina. The next catchments in the list are Fraser River and Skeena River with values already below 0.6. These are catchments which are located on a different continent, but share a similar boreal climate. Mezen's behaviour, including a high value of the Nash-Sutcliffe coefficient of about 0.9 which is nearly constant for most configurations and established after a short training period, is characteristic for those catchments which have a high cyclostationarity ($\gamma > 0.7$) and large correlation values with some other catchments. In the sample dataset, these properties are often found in the boreal areas.

Amazon River (figure 5.10) is another example for the superiority of method 3 over method 4. All configurations reach a significantly higher value for the Nash-Sutcliffe coefficient with method 3, with the difference averaging at about 0.2 for case 1, at about 0.4 for the cases 2 and 3, and even larger differences for the configurations using poorly correlated observations. The Nash-Sutcliffe coefficient of the prediction using method 4 with the 3 least correlated catchments as observations never surpasses a value of 0 and therefore it is not shown in the plot.

Amazon River is a very isolated catchment, the nearest other catchment being at a distance of over 5000 km, and therefore no very high correlations exist. Even though Amazon River's highest correlation value of 0.67 which is reached with Fraser River is lower than the values of the previously analysed catchments a prediction using only Fraser River as an observation still yields the best results. The behaviour of Amazon River which especially showcases a vast superiority of method 3 over method 4 is typical for those catchments in the sample dataset which are isolated both geographically and with respect to their highest correlations, and have a medium or high amount of cyclostationarity.

However, there are also a couple of catchments where only one observation does not produce the best prediction quality. Alabama (figure 5.11) is generally quite similar to Susquehanna River (figure 5.8) in their connection to other catchments, their integration into the whole dataset, and the behaviour of the quality of their predictions for the different methods and cases. However, Alabama's prediction quality greatly improves when the three most correlated catchments are chosen instead of only the one most correlated. In fact, case 3 even produces

Table 5.2: *Impact of the correlation and distance on the quality of observation configurations. C_{max}: Maximum correlation reached with any other catchment; d_C: The distance to this catchment of maximum correlation; d_{min}: The distance to the nearest other catchment; $\mathrm{NSE}_{case,T}$: Nash-Sutcliffe coefficient for cases 1 (all other catchments as observations) and 2 (only the 3 most correlated catchments as observations) after a training period of T years.*

Catchment	$C_{\max}$	d_C[km]	$d_{\min}$[km]	$\mathrm{NSE}_{1,3}$	$\mathrm{NSE}_{2,3}$	$\mathrm{NSE}_{1,20}$	$\mathrm{NSE}_{2,20}$
AMAZONAS	0.67	9073	5269	−0.15	0.70	0.82	0.88
MISSISSIPPI RIVER	0.52	1233	912	−0.69	0.30	0.55	0.44
YENISEI	0.83	2297	2258	0.42	0.88	0.89	0.91
LENA	0.84	5084	2297	0.66	0.91	0.91	0.92
ST LAWRENCE	0.42	1483	515	−1.26	−0.05	−0.50	−0.40
ORANGE	0.15	15886	7983	−0.08	0.01	−0.30	−0.12
DANUBE	0.43	10147	1755	−3.54	0.16	0.51	0.44
COLORADO RIVER (PACIFIC OCEAN)	0.27	3501	1777	−0.28	0.02	−138.02	−74.81
NORTHERN DVINA	0.96	336	336	0.80	0.86	0.85	0.85
FRASER RIVER	0.92	565	565	0.63	0.88	0.91	0.94
COLORADO RIVER (CARIBBEAN SEA)	0.81	90	90	−0.84	0.51	0.46	0.46
BRAZOS RIVER	0.84	90	90	0.09	0.50	0.81	0.81
SUSQUEHANNA RIVER	0.80	263	263	−0.24	0.46	0.51	0.54
VUOKSI	0.70	131	131	−2.21	−0.00	0.39	0.54
APALACHICOLA RIVER	0.92	303	248	0.20	0.79	0.78	0.80
ALABAMA	0.88	64	64	−1.43	0.61	0.76	0.77
MEZEN	0.96	336	336	0.75	0.88	0.90	0.90
STIKINE RIVER	0.84	5084	337	0.62	0.91	0.96	0.96
TOMBIGBEE	0.88	64	64	−0.76	0.32	0.73	0.74
TRINITY RIVER (TEXAS)	0.84	90	90	−0.91	0.42	0.62	0.64
GEORGE RIVER	0.55	3855	1312	−1.28	0.30	0.41	0.43
ALTAMAHA RIVER	0.92	303	303	−0.47	0.76	0.78	0.82
KYMIJOKI	0.70	131	131	−3.29	0.02	0.52	0.58
SKEENA RIVER	0.92	565	337	0.50	0.92	0.93	0.94
POTOMAC RIVER	0.80	263	263	−0.31	0.40	0.48	0.49

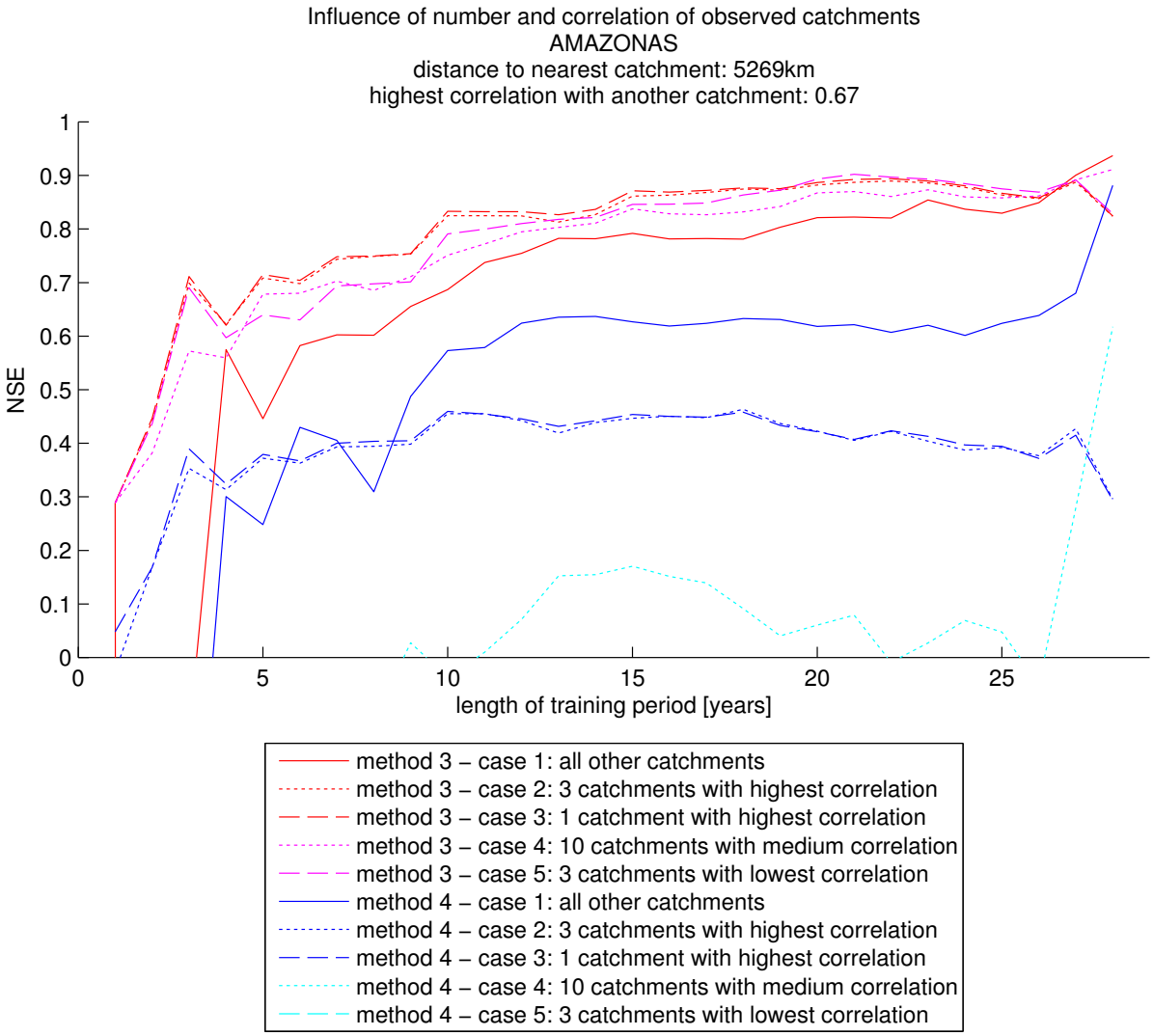

Figure 5.10: *Amazon River: Nash-Sutcliffe coefficient as a function of the length of the training period for different configurations of observed catchments and both analysed prediction methods.*

clearly worse results than using all other catchments as observations, except if the training period is short. A look into the actual correlation values delivers an explanation: While Susquehanna River has one highly correlated catchment with a value of 0.80 and then continues with values of about 0.52, Alabama actually features three catchments with values of 0.80 or higher. The largest part of the catchments in the sample dataset reaches their highest correlation value with a nearby catchment, if a catchment exists within a distance of less than about 600 km (table 5.2). If no other catchments are close, correlation sometimes peaks with very distant catchments. It is obvious that two catchments which are close to each other are predestined to have a high correlation. The most notable exception to this rule is Stikine River which is only 337 km away from Skeena River, but has their correlation of 0.80 surpassed by Northern Dvina at a distance of over 5000 km. This shows again that the catchments in the boreal areas are very similar, irrespective of the continent they are situated on. Except for St Lawrence, all catchments which have a neighbouring catchment within 600 km, and almost exclusively these catchments, reach a maximum correlation value of 0.70 or above. It is evident that correlation

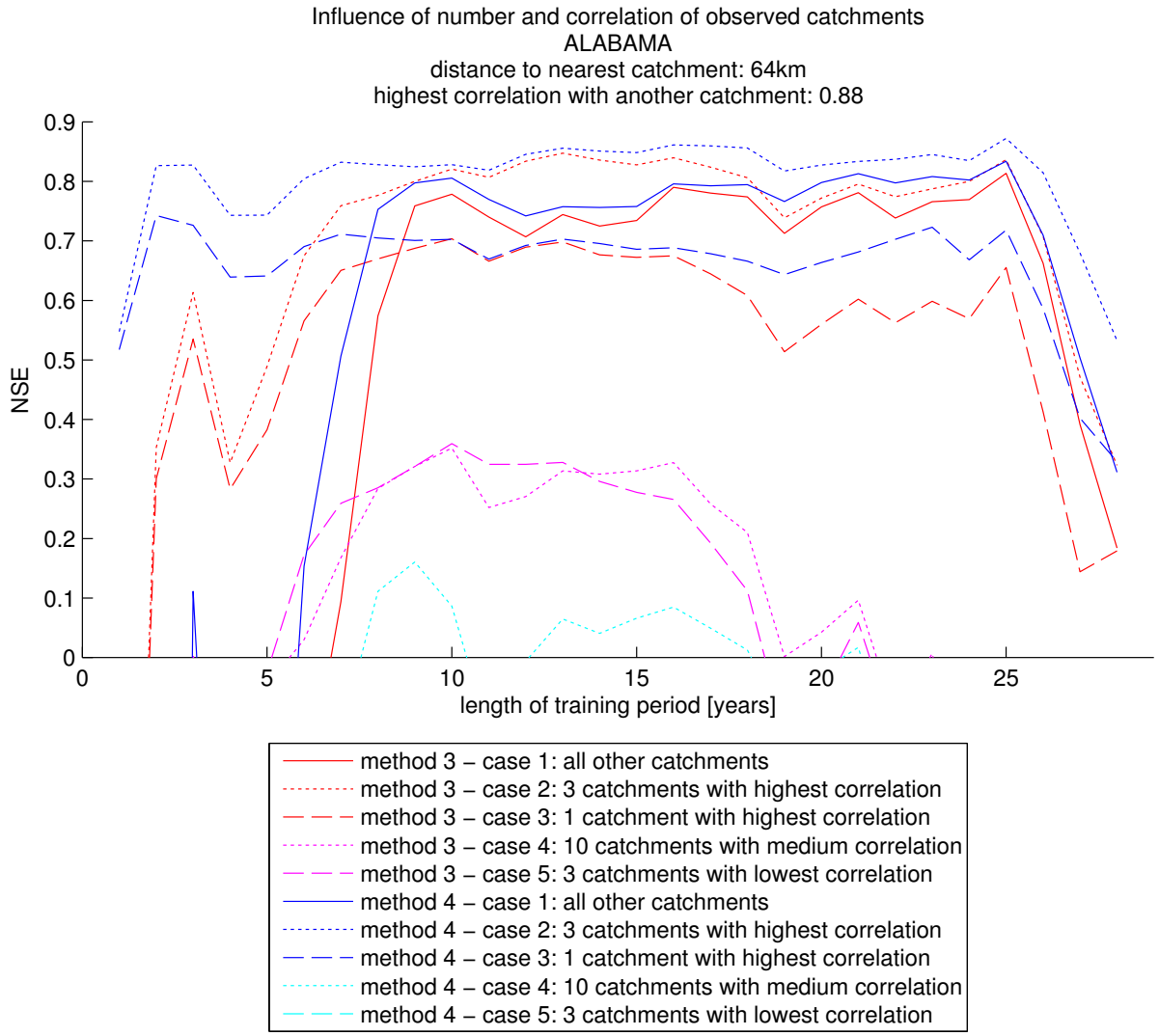

Figure 5.11: *Alabama: Nash-Sutcliffe coefficient as a function of the length of the training period for different configurations of observed catchments and both analysed prediction methods.*

and distance are strongly depending on each other: Proximity entails high correlation.
When only a short training period of 3 years is available, using few, but highly correlated catchments as observations (case 2) always exceeds using as many catchments as possible (case 1): $NSE_{2,3} > NSE_{1,3}$ (figure 5.12). Still, longer training periods generally benefit the prediction quality, also when few catchments are used as observations ($NSE_{2,20} > NSE_{2,3}$). With a long training period of 20 years, case 2 is often a slightly better configuration than case 1, but the difference is usually very small (figure 5.13).

5.3 Location and area of a catchment and its prediction quality

It has already been observed that geographically isolated and therefore badly correlated catchments tend to achieve lower qualities concerning the results of a prediction. To access this

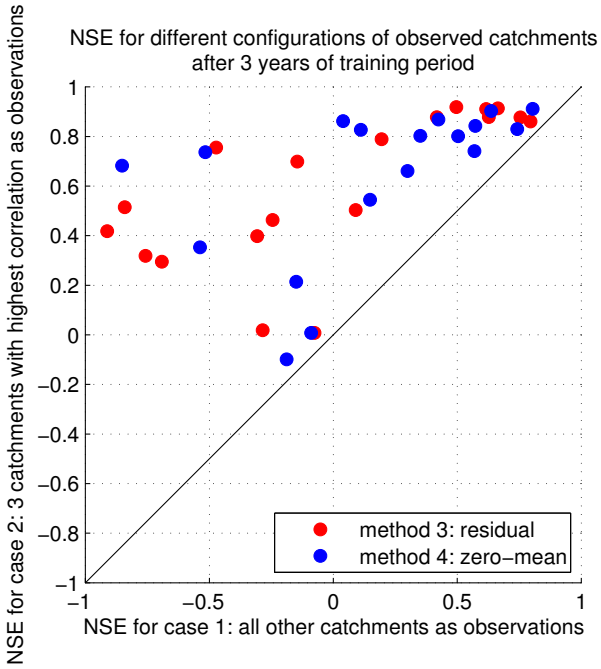

Figure 5.12: *Comparison of Nash-Sutcliffe coefficients after 3 years of training period. Only catchments with* $\mathrm{NSE} > -1$ *for both configurations of observed catchments are shown.*

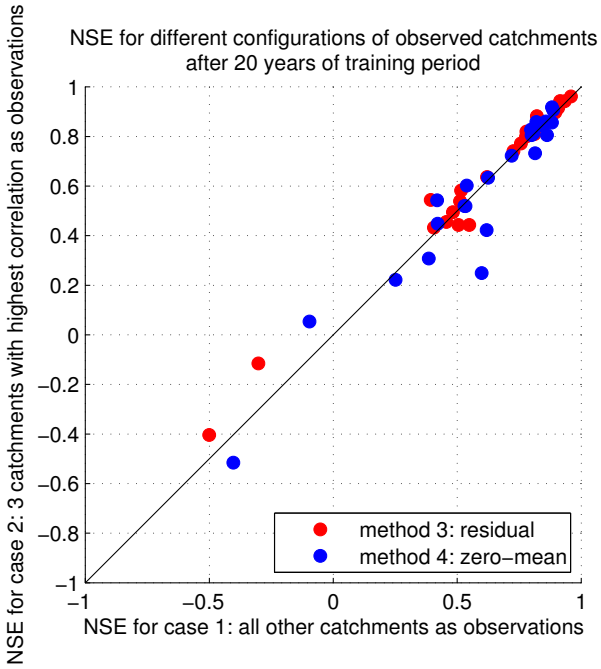

Figure 5.13: *Comparison of Nash-Sutcliffe coefficients after 20 years of training period. Only catchments with* $\mathrm{NSE} > -1$ *for both configurations of observed catchments are shown.*

problem more systematically the values for NSE are plotted into their respective catchments in a map (figure 5.14). This reveals that even large and isolated catchments like Amazonas can achieve good results. However, the residual Nash-Sutcliffe coefficient $\widetilde{\mathrm{NSE}}$ is more relevant to describe the prediction quality which can be achieved because it compares the predicted values to a model generated by using the mean runoff value for each month of the year. This model is easily computed and can therefore serve as a basic value which every prediction should surpass at least.

Figure 5.15 shows that the residual Nash-Sutcliffe coefficient $\widetilde{\mathrm{NSE}}$ indeed delivers a distribution matching the initial assumption that large catchments receive worse results. Amazonas and Yenisei catchments reach qualities clearly inferior to the monthly mean ($\widetilde{\mathrm{NSE}} < 0$). None of the larger catchments obtains a coefficient far better than zero, but on a closer look, one can see that actually most of the smaller catchments do, like for example on the East coast of the USA and in western Canada. This behaviour is also shown by figure 5.16 where it can clearly be seen that catchments with an area of more than $500000\,\mathrm{km}^2$ can never be predicted very successfully, whereas smaller catchments achieve good results in most cases.

A possible explanation for this phenomenon might be that correlation works well over a couple of hundreds of kilometres distance between the centre of two catchments but loses effectiveness after that. As for a large catchment the distance between its centre and its border may already surpass 1000 km or 2000 km, the central area is poorly correlated in any case and its influence on the runoff prediction degrades the whole result.

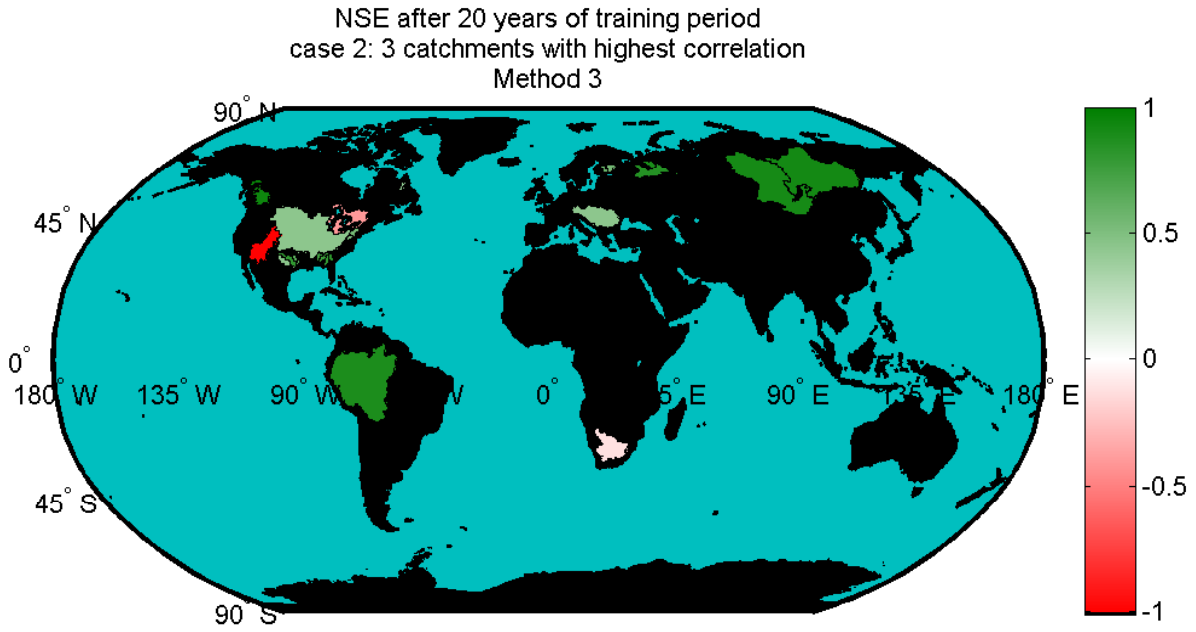

Figure 5.14: *Map showing the value achieved for the Nash-Sutcliffe coefficient* NSE *in the respective catchment. Catchments with* $\mathrm{NSE} < -1$ *are shown in red.*

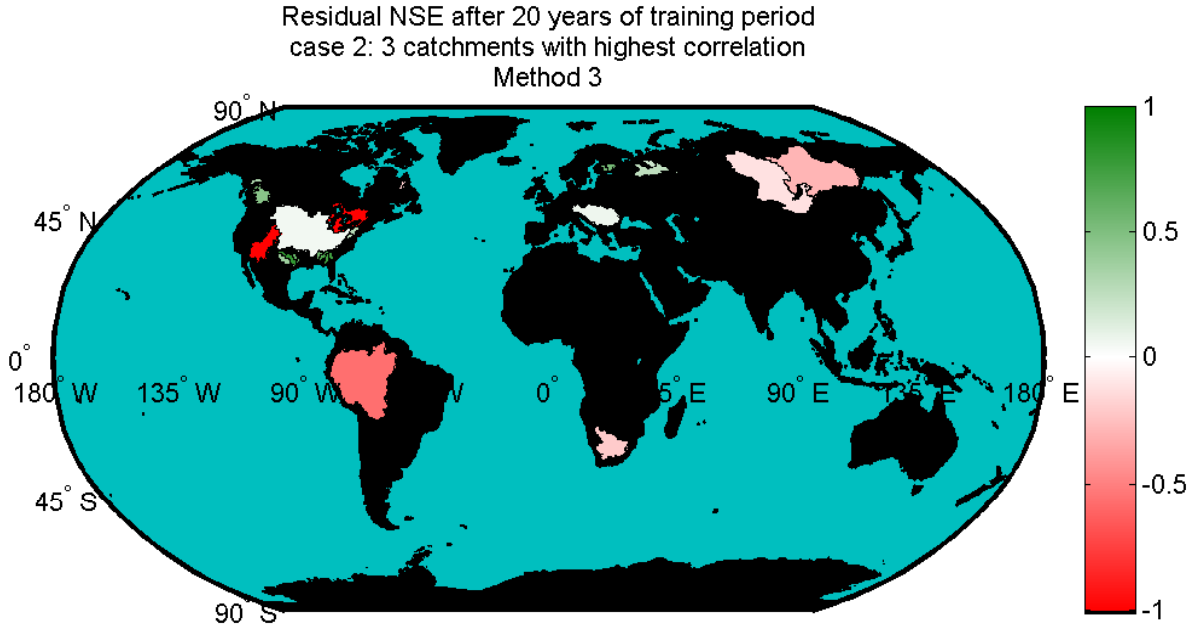

Figure 5.15: *Map showing the value achieved for the residual Nash-Sutcliffe coefficient* $\widetilde{\mathrm{NSE}}$ *in the respective catchment. Catchments with* $\widetilde{\mathrm{NSE}} < -1$ *are shown in red.*

5.4 Cyclostationarity

It is already evident from table 5.1 that there is no clear relationship between the cyclostationarity of a catchment and the potential quality of a prediction. However, not only the model

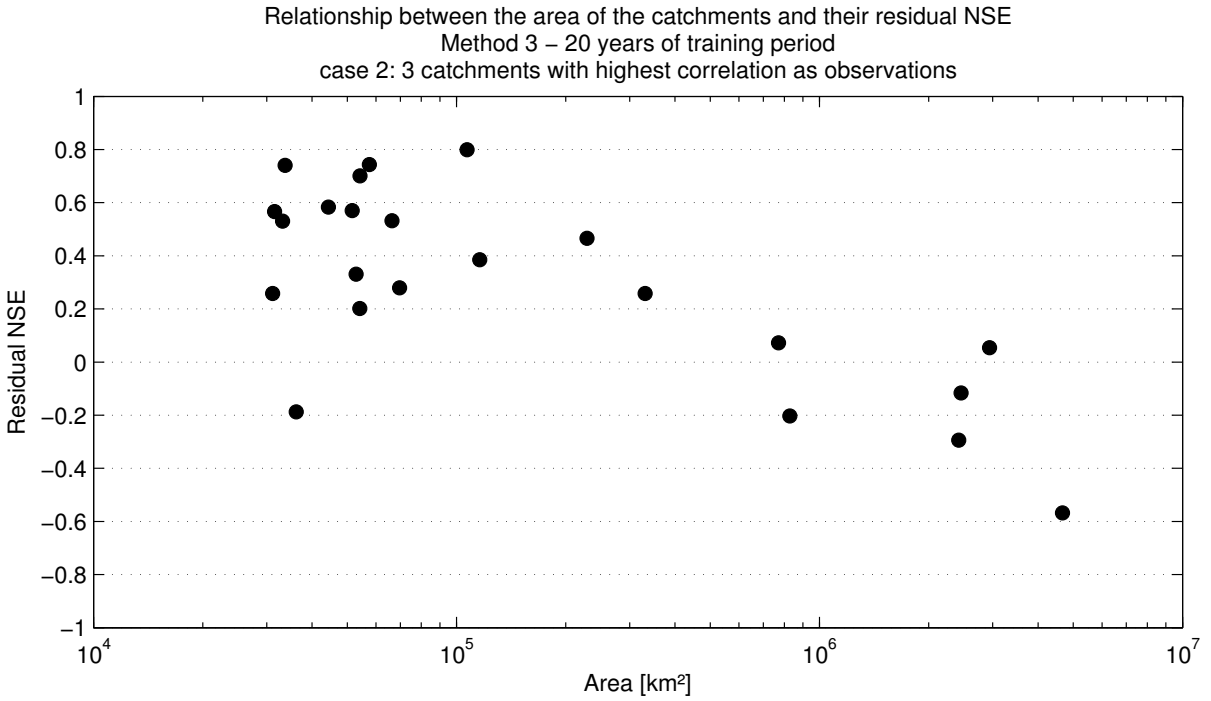

Figure 5.16: *Relationship between the area of the catchments and their* $\widetilde{\mathrm{NSE}}$. *Catchments with* $\widetilde{\mathrm{NSE}} < -1$ *are not shown.*

efficiency which is relating to the amount of errors, but also the distribution of the errors is important. Figure 5.17 shows that the cyclostationarity does have an impact on the noise performance of the errors when using method 4. On the left side the true values $\boldsymbol{s}$, the predicted values $\hat{\boldsymbol{s}}$ and the errors $\boldsymbol{e}$ are graphed like it has already been done in previous figures, but this time the errors are in a separate plot to be able to see more details. In the time domain no unusual properties are detectable, so the three signals are transformed into the frequency domain by a fast Fourier transform.

It can be seen that yearly oscillations and frequencies that are a divisor of the annual oscillation make up a great part of the original values and the prediction. Clear peaks are typical for a very cyclostationary behaviour. The spectral analysis of the errors (bottom right plot) reveals a similar tendency in the performance of the errors, namely three strong peaks at the $\frac{2}{a}$, $\frac{3}{a}$, and $\frac{4}{a}$ frequencies. This cyclostationarity of the errors entails an undesired loss of cyclostationarity in the behaviour of the predicted signal $\hat{\boldsymbol{s}}$. Therefore a frequency spectrum resembling white noise at frequencies of $\frac{1}{a}$ and higher is preferable.

Such can actually be reached by the utilisation of method 3. Figure 5.18 illustrates the absence of a cyclostationarity in the behaviour of the inconsistencies $\boldsymbol{e}$ when predicting out of monthly residuals under otherwise unchanged conditions. The residual methods' quality to conserve the cyclostationary properties of the true signal originates from the way the covariances are formed for this method (see equation 4.11). The used covariance matrices do not store any information about the cyclostationarity, and thus these informations are directly transferred by subtracting $\bar{\boldsymbol{R}}_{\mathrm{cyc}}$ (see equation 4.10) before the prediction process and adding them again later on.

The loss of cyclostationarity through the prediction process which can be observed when using method 4 can be prevented by using method 4. This problem mainly appears for catchments

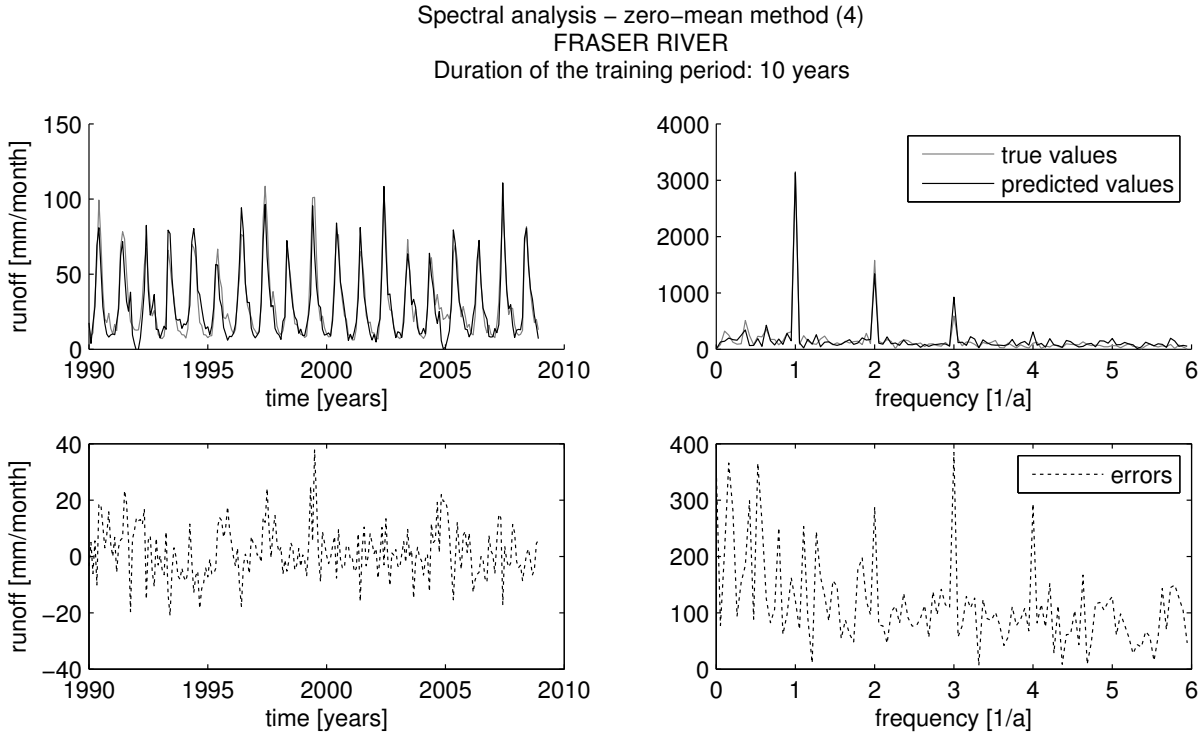

Figure 5.17: *Spectral analysis of the signal and the model error for a prediction for Fraser River (method 4) after a training period of 10 years. All other catchments are used as observations. True values* $\boldsymbol{s}$*, predicted values* $\hat{\boldsymbol{s}}$*, and errors* $\boldsymbol{e}$ *are plotted in the time domain (left side) and in the frequency domain (right side).* $\mathrm{NSE} = 0.85$*,* $\widetilde{\mathrm{NSE}} = -0.09$*.*

with a high cyclostationarity ($\gamma > 0.75$) which often occurs for those catchments in the sample dataset located in boreal and tropical climates.

5.5 Trend

The long-term trend of the runoff of a catchment as described in section 2.4.3 and its impact on a prediction has to be analysed. It gives information about the long-ranging deviations of the runoff from its regular course. These deviations may be caused by a variety of effects which are subject to further hydrological researches, among those climate change and anthropogenic influences.
As the covariance matrices used in a prediction are generated in a training period to predict values at another period of time, usually a later time, a strong trend can have a negative impact on the usefulness of these covariances at the time of the prediction. The actual covariances between the true signal and the observations may become too different from those generated in the training period causing the quality of the prediction to suffer.
However, careful analysis of the relationship between the trend values (see table 2.2) and various prediction results has not yielded hints at any meaningful effects caused by the trend. The influence of the trend is possibly overlapped by other parameters affecting the prediction quality, for example cyclostationarity. Also, a duration of 29 years is rather short in the context of

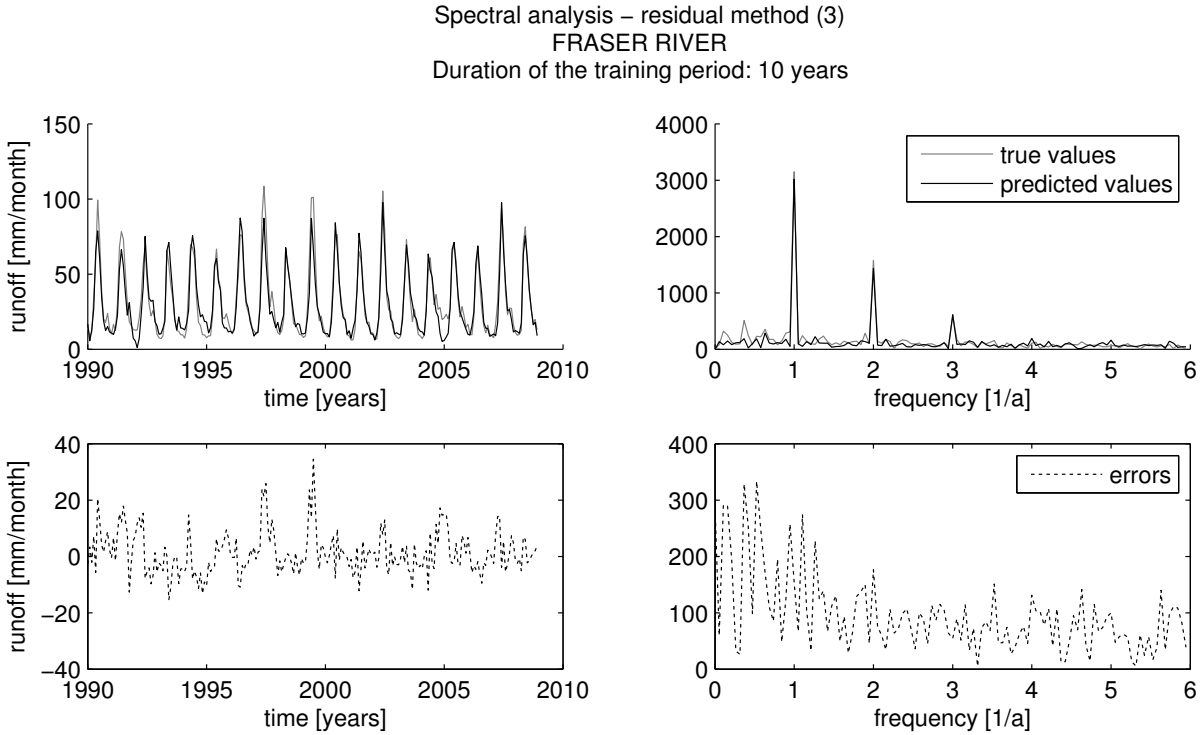

Figure 5.18: *Spectral analysis of the signal and the model error for a prediction for Fraser River (method 3) after a training period of 10 years. All other catchments are used as observations. True values* $\mathbf{s}$*, predicted values* $\hat{\mathbf{s}}$*, and errors* $\mathbf{e}$ *are plotted in the time domain (left side) and in the frequency domain (right side).* $\mathrm{NSE} = 0.89$*,* $\widetilde{\mathrm{NSE}} = 0.20$*.*

climatic changes, potentially degenerating the trend into insignificance for the configurations analysed in this study.

Chapter 6

Conclusion

6.1 Summary

Least-squares prediction has been chosen as a way to model runoff data whose availability from measurements has declined lately. The effectiveness of this process has been exemplified on a sample dataset of 25 catchments described in chapter 2. Five different prediction methods have been compared by an evaluation of the results on an exemplary catchment, resulting in the decision to focus on two of them for a more detailed analysis: Prediction out of monthly residuals and prediction out of zero-mean time series.
In the previous chapter, the results of predictions using different configurations have been compared. A main topic of this study is the relation between the two analysed prediction methods using residuals and zero-mean time series, as described in section 4.3.3 and section 4.3.4, respectively. Chapter 5 has shown that method 3 is generally preferable.
For a training period of 15 years and using all other catchments as observations, method 3 generally provides slightly better results than method 4, as it has been displayed in table 5.1. Figure 5.6 has led to the finding that, while prediction results are often similar in quality for both methods, method 3 is vastly superior for certain catchments. This advantage over method 4 intensifies when the observations are poorly correlated, be it because inapplicable catchments have been chosen intentionally like in figure 5.9 or because suitable observations are not available like for the Amazon River catchment. In practical applications, the need to predict with suboptimal observed catchments may arise from a false estimation of the correlation properties between observed and predicted catchment or from a lack of disposable observations with a high correlation. In this study, the correlations have been computed using the whole time series. When incomplete time series are to be used as observations, the correlations can only be evaluated from those parts of the time series which are available for the observed and the predicted catchment. Under the realistic assumption of a very short training period of for example 2 years, for which reasonable prediction results can be achieved, a reliable estimation of the correlation properties is not necessarily ensured. Another aspect favouring method 3 is the preservation of the cyclostationary properties of the original signal during the prediction process which is especially important for catchments with a high cyclostationarity.
While method 3 is generally preferable, there are individual cases occurring which argue for an application of method 4 under special configurations. A better prediction quality is sometimes achieved for certain catchments when few observations and a very short training period are used, for example for the Alabama catchment (see figure 5.11), where method 4 generally yields slightly better results than method 3, but especially for training periods which are shorter than 8 years. However, these are exceptions which only better the prediction efficiency by slight amounts in the most cases. On the contrary, there are many cases which reveal a

large advantage of method 3 over method 4, for example the Amazon River catchment (see figure 5.10). Therefore, for the further analysis of runoff using least-squares prediction, a usage of the residual method by default would certainly not be a bad decision, as long as the possibility of an application of the prediction method using zero-mean time series is kept in reference for special configurations.
When the runoff of a catchment is predicted, it is in any case important to make a considered choice of the catchments which are used as observations and the length of the training period. Contrary to the initial assumption, fewer observations usually benefit the quality of the prediction. Using all available catchments is generally not a reasonable option. The exact optimal amount of observations varies for each catchment. The testing conducted in section 5.2 suggests that only catchments with a high correlation to the predicted catchment should be used as observations if such are available. Declaring three catchments to be observations normally results in a prediction quality which is at least similar to that reached for an optimal amount of observations. The length of the training period is not as important as the choice of observed catchments, but it can still make a difference. Raising the length of the training period over the mark of 10 years does not yield any significant improvement, but below this value the choice of the prediction method (see above) and the observed catchments becomes increasingly important. It is even more crucial to optimise the choice of observations for the respective predicted catchment. If procurable, the conservative way of using long training periods is a sound alternative.

6.2 Discussion

The quality of the results which are achieved by least-squares prediction can compete with those of other methods. Using a long training period of 20 years and a small choice of well correlated catchments produces Nash-Sutcliffe coefficients $NSE_{2,20}$ (see table 5.2) which surpass a value of 0.4 in about 90% of the analysed catchments and 0.75 for about half of the 25 catchments. Larger catchments are generally harder to predict. One likely reason for this is that some of those are quite isolated geographically (Amazon River, Orange, Danube, Colorado River (Pacific Ocean)) within the sample dataset, but there are also catchments like Mississippi River and St Lawrence for which this is not the case. It can be assumed that the large area of a catchments entails a variety of climates influencing it. Therefore a few neighbouring catchments may not be sufficient to grant enough information for a high quality prediction which needs correlation data affecting the whole area. The strength of the least-squares prediction rather lies in the smaller catchments which is not displeasing because the availability of in situ measurements of these catchments is more limited anyway and they are also harder to replace by remote sensing monitoring methods (Fekete and Vörösmarty, 2007).
The study by Sneeuw et al. (2013) compares different approaches of estimating runoff. For the Amazon River catchment Nash-Sutcliffe coefficients of NSE $= 0.91$ and NSE $= 0.92$ are reached for a method using satellite altimetry with quantile function based stage-discharge relationships and a runoff-storage relationship, respectively, slightly higher values than with least-squares prediction (NSE $= 0.88$). For the Danube catchment, the altimetric method (NSE $= 0.76$) outperforms the results of this study (NSE $= 0.44$). In the Yenisei catchment however, the least-squares result of NSE $= 0.91$ is clearly better than a 0.75 reached by a runoff-precipitation relationship which is the best of the methods analysed in the study by Sneeuw et al. (2013). Milzow et al. (2011) reach a maximum Nash-Sutcliffe coefficient of 0.63

when estimating the Okavango catchment which is not examined in this study by satellite radar altimetry, SAR surface soil moisture and GRACE total storage changes. Bjerklie et al. (2003) indicates an accuracy of 20% which can be reached by measuring river discharge from space, a value equally comparable to the results of least-squares prediction, compare for example figure 5.18.
When matching least-squares prediction against other methods aiming at the estimation of runoff data, its limitations have to be considered: Catchments which have never been gauged before cannot be predicted because no covariance information at all is available. Events which only concern one catchment, like many anthropogenic influences and some natural incidents as well, will always lead to lower prediction qualities because the observed catchments are not affected by them.

6.3 Outlook

As the comparison of the results of other studies with least-squares prediction's outcomes is quite promising, a continuation of the evaluation of this approach is desirable. Certainly, an improvement of the predictive power of the method presented in this study can be achieved by pursuing certain further approaches. Many catchments are rather isolated within the sample dataset, as there is for example only one catchment located in South America, Africa, and Central Europe, respectively. The prediction results for these catchments are rather limited in comparison to those of other catchments because high correlation values benefiting the prediction quality usually occur with nearby catchments as it has been shown in section 5.2. This deficit could be eliminated by using the whole GRDC dataset which has been provided, meaning that catchments with temporarily available runoff measurements could be used as additional observations. This would entail some problems for which solutions have to be found: As the additionally introduced catchments are incomplete, the prediction could also only be carried out for the available parts of the time series. An application of multiple observations creating a prediction assembled out of several smaller parts might be a solution, but the behaviour at the juncture points has to be considered. Using incomplete time series as predicted catchments seems less useful because the validation periods might become very short.
Until now, training and validation periods have not been regarded in steps of less than a year. Incomplete time series may often contain short gaps of only one or a few months' length which do not justify discarding the rest of the data for the respective year. Therefore an implementation of the prediction enabling the use of monthly segments is necessary. In doing so, special attention has to be paid to the conservation of the cyclostationarity which might be affected by different observation conditions during the course of the year.
The influence of the number of observations on the quality of a prediction has only been analysed in a rather superficial way in this study. It has been shown that fewer observations are generally beneficial by comparing the results for one, three, and 24 observed catchments. Also, the usage of highly correlated catchments is important when using few observations. A more systematic approach could be applied here to find a best number of observations for each predicted catchment. Section 5.2 has already suggested that this optimal amount might change depending on the availability of highly correlated observations. The existence of a correlation threshold separating catchments which are advantageous to the prediction from potentially hindering catchments could be investigated.
In the context of correlations, the influence of cyclostationarity has to be regarded, as naturally

catchments with highly cyclostationary behaviours reach higher correlations so that it might be useful to apply the monthly mean $\bar{r}_{cyc}$ in the correlation (equation 3.2) as well, similar to how it has been done for the formation of covariance matrices for method 3. If using correlation information to optimise the configuration of the prediction proves valuable, generating this information out of the training period only will have to be considered because a calculation using the whole time series implies its availability.
All of the analysed prediction methods use covariances generated in a training period to predict runoff in a validation period. However, it can be useful to generate the covariance information in a different way, especially when potential training periods are very short or not available at all. The latter is the case when catchments are predicted which have never been gauged before. Such covariance information can be created by models, including MERRA (Reichle et al., 2011) and GLDAS (Rodell et al., 2004).
A final step would certainly be the application of least-squares prediction to actually predict runoff values for ungauged catchments for later use in hydrological modelling and calibration of satellite based measuring methods.

Bibliography

Alsdorf, D., I. Melack, T. Dunne, L. Mertes, L. Hess, and L. Smith. 2000. Interferometric radar measurements of water level changes on the amazon flood plain. *Nature* 404:174–177. doi:10.1038/35004560.

Bárdossy, András. 2005. Hydrology I. Lecture Notes, University of Stuttgart.

Birkett, Charon M. 1998. Contribution of the topex nasa radar altimeter to the global monitoring of large rivers and wetlands. *Water Resources Research* 34(5):1223–1239. doi:10.1029/98WR00124.

Bjerklie, David M., Lawrence S. Dingman, Charles J. Vorosmarty, Carl H. Bolster, and Russell G. Congalton. 2003. Evaluating the potential for measuring river discharge from space. *Journal of Hydrology* 278:17–38. doi:10.1016/s0022-1694(03)00129-x.

Bjerklie, David M., Delwyn Moller, Laurence C. Smith, and S. Lawrence Dingman. 2005. Estimating discharge in rivers using remotely sensed hydraulic information. *Journal of Hydrology* 309:191–209. doi:10.1016/j.jhydrol.2004.11.022.

Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione, and H. Savenije. 2013. *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*. Cambridge University Press.

Fekete, Balázs M., and Charles J. Vörösmarty. 2007. The current status of global river discharge monitoring and potential new technologies complementing traditional discharge measurements. *IAHS Publ* 309.

Hendriks, M.R. 2010. *Introduction to physical hydrology*. Oxford University Press.

Maidment, D.R. 1993. *Handbook of hydrology*. Civil engineering, McGraw-Hill.

Milzow, C., P. E. Krogh, and P. Bauer-Gottwein. 2011. Combining satellite radar altimetry, sar surface soil moisture and grace total storage changes for hydrological model calibration in a large poorly gauged catchment. *Hydrology and Earth System Sciences* 15(6):1729–1743. doi:10.5194/hess-15-1729-2011.

Moritz, H. 1980. *Advanced physical geodesy*. Wichmann.

Nash, J., and J. Sutcliffe. 1970. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology* 10:282–290. doi:10.1016/0022-1694(70)90255-6.

Reichle, Rolf H., Randal D. Koster, Gabrielle De Lannoy, Barton A. Forman, Qing Liu, Sarith P. P. Mahanama, and Ally Toure. 2011. Assessment and enhancement of MERRA land surface hydrology estimates. *Journal of Climate* 24(24):6322–6338. doi:10.1175/JCLI-D-10-05033.1.

Rodell, M., P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J. Entin, J. P. Walker, D. Lohmann, and D. Toll. 2004.

The global land data assimilation system. *Bulletin of the American Meteorological Society* 85:381–394. doi:10.1175/BAMS-85-3-381.

Schlittgen, R., and B.H.J. Streitberg. 1995. *Zeitreihenanalyse*. Lehr- und Handbücher der Statistik / Lehr- und Handbücher der Statistik, Oldenbourg Wissensch.Vlg. doi:10.1524/9783486710960.

Smith, Laurence C., Bryan L. Isacks, Arthur L. Bloom, and A. Brad Murray. 1996. Estimation of discharge from three braided rivers using synthetic aperture radar satellite imagery: Potential application to ungauged basins. *Water Resources Research* 32(7):2021–2034. doi:10.1029/96WR00752.

Sneeuw, Nico, Christof Lorenz, Mohammad J. Tourian, Balaji Devaraju, Johannes Riegger, Harald Kunstmann, and András Bárdossy. 2013. Estimating runoff using hydro-geodetic approaches: Status and challenges. *Manuscript submitted for publication* .

Sun, W. C., H. Ishidaira, and S. Bastola. 2010. Towards improving river discharge estimation in ungauged basins: calibration of rainfall-runoff models based on satellite observations of river flow width at basin outlet. *Hydrology and Earth System Sciences* 14(10):2011–2022. doi:10.5194/hess-14-2011-2010.

Todhunter, I. 1871. *Spherical Trigonometry, for the Use of Colleges and Schools*. Macmillan and Company.

Vörösmarty, Charles J., Cort J. Willmott, Bhaskar J. Choudhury, Annette L. Schloss, Timothy K. Stearns, Scott M. Robeson, and Timothy J. Dorman. 1996. Analyzing the discharge regime of a large tropical river through remote sensing, ground-based climatic data, and modeling. *Water Resources Research* 32(10):3137–3150. doi:10.1029/96WR01333.