wawi

INSTITUT FÜR HYDROLOGIE UND WASSERWIRTSCHAFT

MitteilungenHeft 106

A. FANGMANN

Low flow prediction in time and space
An adaptive statistical scheme for regional climate change impact assessment

Low flow prediction in time and space

An adaptive statistical scheme for regional
climate change impact assessment

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades
Doktor-Ingenieurin
- Dr.-Ing. -

genehmigte Dissertation

von

Anne Fangmann, M.Sc.
geboren am 26.07.1986 in Lohne

2017

Referent: Prof. Dr.-Ing. Uwe Haberlandt
Korreferent: Assoc. Prof. Dipl.-Ing. Dr.techn. Gregor Laaha

Tag der Promotion: 01.12.2017

Danksagung

Mein Dank richtet sich vor allem an Prof. Dr.-Ing. Uwe Haberlandt. Ohne seine Ideen, sein Wissen und die gute Betreuung während meiner Zeit am Institut für Hydrologie und Wasserwirtschaft wäre diese Arbeit nicht zustande gekommen. Vielen Dank für die Unterstützung, die ständige Motivation und neuen Anregungen, und dafür, dass ich so viel bei Ihnen lernen durfte.

Außerdem möchte ich mich bei den Mitgliedern meiner Promotionskommission bedanken: bei Assoc. Prof. Dipl.-Ing. Dr.techn. Gregor Laaha für die Übernahme des Korreferats, sowie bei Prof. Dr. rer. nat. Insa Neuweiler und Prof. Dr.-Ing. Karl-Heinz Rosenwinkel als Mitglieder des Promotionskollegiums.

Ein großer Dank gilt außerdem Hannes Müller und Bastian Heinrich, die bei allen arbeitsbezogenen und privaten Sorgen zur Stelle waren und mir von Anfang an eine äußerst angenehme Zeit am Institut beschert haben.

Erklärung

Hiermit versichere ich eidesstattlich, dass ich

1. die Regeln der geltenden Promotionsordnung kenne und eingehalten habe und mit einer Prüfung nach den Bestimmungen der Promotionsordnung einverstanden bin,

2. die Dissertation selbst verfasst habe, keine Textabschnitte von Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen in meiner Arbeit angegeben habe,

3. Dritten weder unmittelbar noch mittelbar geldwerte Leistungen für Vermittlungstätigkeiten oder für die inhaltliche Ausarbeitung der Dissertation erbracht habe,

4. die vorliegende Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe,

5. die gleiche oder eine in wesentlichen Teilen ähnliche Arbeit bei keiner anderen Hochschule als Dissertation eingereicht habe und auch keine andere Abhandlung als Dissertation eingereicht habe,

6. damit einverstanden bin, dass die Dissertation auch zum Zwecke der Überprüfung der Einhaltung allgemein geltender wissenschaftlicher Standards genutzt wird, insbesondere auch unter Verwendung elektronischer Datenverarbeitungsprogramme.

Hannover, den 01.09.2017

Anne Fangmann

Abstract

A comprehensive regional assessment of climate change impacts on low flows requires prediction of streamflow characteristics in time and space under consideration of non-stationarity. Three aspects of low flow estimation are analyzed, i.e. prediction in time, in space and simultaneously in time and space. For all three types, various methods are analyzed that all pose basic statistical approaches. The aim is the provision of fast and simple tools for inclusion into model ensembles for regional climate change impact assessment. Target variables for all methods are annual low flow index values, extracted from daily average discharge series measured all over Lower Saxony.

Temporal low flow prediction via extrapolation of trends and patterns of low flow time series without any climatic input data appeared to be of high uncertainty and highly dependent on the period used for model calibration. Their application is therefore discouraged. Strong relationships could be found between meteorological and low flow indices at the individual catchments, which were successfully used to build multiple linear regression (MLR) models that appear promising for prediction of future low flows based on input of climate model data.

For prediction of average low flow values in space, MLR models that relate the respective flow indices to catchment characteristics appeared to outperform geostatistical methods. The set of external variables primarily consisted of temporally aggregated meteorological indices rather than physiographic variables. The index flood method proved suitable for regional low flow frequency analysis and outperformed direct MLR regionalization for quantile prediction.

A combined approach for spatiotemporal estimation of low flows via panel data regression was tested on three levels, i.e. direct prediction of annual low flow values, prediction of flood index values and fitting of stationary regional distribution functions, and spatiotemporal estimation of L-moments. It turned out that the direct approach gave the best results. The L-moment regionalization exhibited major issues with prediction of higher order moments and the index flood approach bases on too many assumptions that cannot hold under non-stationary scenarios. The performance of spatiotemporal modeling of annual low flow values surpassed station-wise temporal prediction and subsequent regionalization.

Keywords: low flow, statistical modeling, spatiotemporal prediction

Kurzfassung

Eine umfangreiche regionale Folgenabschätzung des Klimawandels auf Niedrigwasser-abflüsse bedarf der Schätzung von Abflusscharakteristiken in Zeit und Raum unter besonderer Berücksichtigung von Instationaritäten. Drei Aspekte der Niedrigwasser-prognose werden hier analysiert, nämlich die zeitliche, die räumliche und die kombinierte Schätzung in Zeit und Raum. Für alle drei Arten werden unterschiedliche Methoden untersucht, die auf einfachen statistischen Ansätzen basieren. Ziel ist die Bereitstellung schneller und einfacher Werkzeuge, die in Modellensembles bei der Klimafolgenforschung berücksichtigt werden können. Zielvariablen für alle Methoden sind jährliche Niedrigwasser-indizes, abgeleitet aus Tageszeitreihen des Abflusses verteilt über ganz Niedersachsen.

Zeitliche Niedrigwasserprognose durch Extrapolation von Trends und Mustern in Niedrigwasserzeitreihen ohne Hinzunahme klimatischer Eingangsdaten ergab große Unsicherheiten und hing stark von der Wahl des Kalibrierungszeitraumes ab. Die Anwendung in der Praxis wird deshalb nicht empfohlen. Es konnten straffe Zusammen-hänge zwischen meteorologischen Kenngrößen und Niedrigwasserindizes für die einzelnen Einzugsgebiete festgestellt und zur Formulierung multipler linearer Regressionsmodelle (MLR) genutzt werden. Die Anwendung dieser Modelle auf Basis von Klimamodelldaten scheint aussichtsreich.

Für die Schätzung von mittleren Niedrigwasserkenngrößen im Raum erzielten MLR-Modelle wesentlich bessere Ergebnisse als geostatistische Methoden. Die wichtigsten Einflussgrößen sind dabei zeitlich aggregierte meteorologische Größen. Physiographische Größen stellten sich als eher unwichtig heraus. Die Index-Flood-Methode eignete sich für die regionale Niedrigwasseranalyse und übertraf MLR bei der räumlichen Prognose von Niedrigwasserquantilen.

Ein kombinierter Ansatz zur räumlich-zeitlichen Prognose von Niedrigwasserkenngrößen mithilfe der Paneldatenregression wurde auf drei unterschiedliche Arten angewandt: zur direkten Prognose jährlicher Indexwerte, zur Schätzung von Indexwerten mit Anschließender Anpassung regionaler Verteilungsfunktionen und zur räumlich-zeitlichen Prognose von L-Momenten. Die direkte Prognose lieferte die besten Ergebnisse. Bei der Regionalisierung von L-Momenten stellte sich heraus, dass es Schwierigkeiten bei der Regionalisierung der Momente höherer Ordnung gab. Die Annahmen der Index-Flood-Methode können im Fall von instationärem Niedrigwasserverhalten nicht standhalten. Die Leistung des räumlich-zeitlichen Modells zur Schätzung jährlicher Indexwerte übertraf die stationsweise temporäre Schätzung mit anschließender Regionalisierung.

Schlagworte: Niedrigwasser, statistische Modellierung, räumlich-zeitliche Prognose

Table of contents

List of Tables

List of Figures

Abbreviations

AIC	Akaike information criterion
ANFIS	Adaptive neuro-fuzzy inference systems
ANN	Artificial neural networks
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
BIC	Bayesian information criterion
CD	Catchment descriptors
D_{max} / D_{mean}	Maximum / mean low flow duration
DSD_{max} / DSD_{mean}	Maximum / mean dry spell duration
EDK	External drift kriging
EMD	Empirical mode decomposition
ETP_{mean} / ETP_{max}	Mean / maximum evapotranspiration
FGLS	Feasible generalized least squares
GAMLSS	Generalized additive models for location, scale and shape
GCM	Global climate model
GEV	General extreme value distribution
GLO	General logistic distribution
GLS	Generalized least squares
GLS-R	Restricted generalized least squares regression
G_{mean} / G_{max}	Mean / maximum global radiation
GRDC	The Global Runoff Data Centre
HHT	Hilbert-Huang transform
HWD_{max} / HWD_{mean}	Maximum / mean heat wave duration
IAHS	International Association of Hydrological Sciences
IMF	Intrinsic mode function
i_M	Meteorological index
i_{NQ}	Low flow index
LOOCV	Leave-one-out cross validation
MAE	Mean absolute error
ME	Mean error
MLR	Multiple linear regression
NLS	Niedersächsisches Landesamt für Statistik

NLWKN	Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz
NM7Q / NM7q	Lowest average 7-day flow / specific flow
NM30Q	Lowest average 30-day flow
NRMSE	Normalized root mean square error
NSE	Nash-Suttcliffe efficiency
obs_{stat} / obs_{reg}	Local / regionalized observed flow
OK	Ordinary kriging
OLS	Ordinary least squares
OLS-R	Restricted ordinary least squares regression
OOWV	Oldenburgisch-Ostfriesischer Wasserverband
Pbias	Percent bias
PC	Principal components
PC-R	Restricted principal component regression
$P-ETP_{mean}$ / $P-ETP_{max}$	Mean / maximum climatic water balance
P_{mean} / P_{max} / P_x	Mean / maximum / quantile precipitation
PSBI	Physiographical space-based interpolation
PUB	Predictions in ungauged basins
Q80	Flow exceeded 80% of the time
Q95 / q95	Flow / specific flow exceeded 85% of the time
R^2	Coefficient of determination
RCM	Regional climate model
SARIMA	Seasonal autoregressive integrated moving average
sim_{stat} / sim_{reg}	Local / regionalized simulated flow
SPEI	Standardized precipitation evaporation index
SPI	Standardized precipitation index
ST-1 / ST-2 / ST-3	Spatiotemporal model variant 1 / 2 / 3
SVM	Support vector machines
SVR	Support vector regression
TK	Topological kriging
T_{mean} / T_{max} / T_{min}	Mean / maximum / minimum temperature
V_{max} / V_{mean}	Maximum / mean deficit volume
VIF	Variance inflation factor
WSD	Wasser- und Schifffahrtsdirektion
WSD_{max} / WSD_{mean}	Maximum / mean wet spell duration

1 Introduction

1.1 Background

Low flow periods are part of any hydrological regime, as they represent runoff conditions that are below the normal and thus occur in every river of the world as natural part of the hydrograph. The magnitude of low flow events naturally varies strongly across climatic regions but even in temperate humid areas, like the North of Germany, low flows pose a major issue with regard to availability of water resources, navigability of rivers, and ecosystem health. Proper management and design are essential for mitigation and averting of adverse effects caused by reduced water levels and rely strongly on detailed prognosis of future dimensions and characteristics of low flow events, like timing, duration and deficit.

Low flow characteristics are directly related to weather. A summer low flow period is usually caused by a lack of rain and possibly high evapotranspiration caused by high temperatures, while a winter event is aggravated through low temperatures and retention of water as snow. Other than for floods, the causing meteorological mechanisms are usually observable over extended periods of time, so that the magnitude of a low flow event can be considered as a summary of the preceding weather conditions. The reaction time and thus the degree of alleviation of an event thereby naturally depends on factors like slope and available storage within a belonging catchment. Low flow behavior consequently varies significantly between base flow and quick flow dominated areas. Spatial examination of low flow characteristics therefore needs to account for specific catchment characteristics that affect the rainfall runoff transformation. In time, however, low flow characteristics appear to be much more readily modeled using average weather conditions than are floods.

Nonetheless, temporal prediction of low flow events poses a major challenge. Short-term forecasts based on recent weather conditions and meteorological prognoses appear feasible but prediction becomes increasingly demanding in the long-term, especially under non-stationary conditions. Non-stationarity implies that the characteristics of low flow events change over time. Depending on the order of non-stationarity, this change can affect means, variances or entire distributions; factors that cannot be ignored during planning.

Next to anthropogenic interference, climate change can be considered a primal cause for non-stationary behavior of a low flow time series. Changes in climate entail alterations of meteorological states and events, which in turn affect the processes of low flow formation. The integration of climate change effects into prediction approaches thus is of paramount importance, but simultaneously brings on some challenges to the modeler.

A comprehensive regional climate change impact analysis requires consideration of various aspects, including the identification of the relationship between meteorology and system response and transformation mechanisms specific to the catchment, investigation of spatial patterns and similarities in streamflow behavior and its controls, assessment of the stationarity of model parameters in space and time and eventually prognosis of future regional streamflow based on all previous findings.

The complexity of the models used for prognosis can range from mere extrapolation of observed low flow trends into the future, over data-driven black box models and conceptual rainfall-runoff models that require parameter adjustment to highly complex physically-based models. The application of simple models lacks the physical basis and denies full understanding of underlying processes, may however be capable of capturing occurrences that are not readily explained by conceptual or even physically-based models and pose a practical alternative due to their low data requirement and easy model set-up.

1.2 Motivation and objectives

At the beginning of this work low flows had been vaguely analyzed in the northern parts of Germany, in great contrast to floods, which have been of scientific and administrative interest for quite some time. In view of the vast number of regional climate change impact studies all over the world, a demand for analyses concerned with the development of low flow behavior arose, even for the rather humid region of Lower Saxony. Knowledge about low flow development in the region is of particular importance for reservoir operators who depend on planning for water level elevation, but also for agriculture and industry who rely on stable water supply. Additionally, environmental concerns that arise from low water levels are of high relevance for the authorities.

Analyses that had been carried out prior to this study (Fangmann, 2012; Fangmann et al., 2013) have shown that changes in meteorological variables and low flow indices have already been observable for the past and present. Especially for summer low flows significant trends and breakpoints toward greater dryness could be detected for the entire region, following shifts in precipitation amounts, which appeared to increase in the winter and significantly decrease during summer. The efforts of this work base on these initial results. The primary question was how low flows in all their aspects will change with a changing climate, which directly entailed questions about existing types of non-stationarity and their incorporation into low flow prediction. Furthermore, since the previous studies revealed great similarities in trends between annual or seasonal meteorological indices and low flow indices, it was questioned whether prediction would be possible on mere basis of these very relationships rather than on the classical hydrological modeling approach. The previous results also showed that changes in low flows can be highly variable in space, indicating that an exhaustive regional impact assessment would additionally call for proper regionalization in order to assess changes at ungauged locations.

The ultimate goal of this work is the identification and provision of a framework comprising of simple statistical modeling approaches that allow for a regional climate change impact analysis, tailored specifically to the problem of low flows. The objective is not to outperform common impact analysis practice but rather to support it by extending the model ensemble through alternative model types. The selected methods should be readily applicable due to low input data requirement, basically no necessary catchment-specific model set up, and minimal processing time. Thus, the framework will ideally be able to produce regional low flow prognoses through input of locally observed discharge and meteorological variables in little time with minor effort. Additionally, the study should enable the investigation of low flow peculiarities in time and space, potential non-stationarity and temporal variability, as well as identification of the specifics of low flow prediction.

1.3 Overview

This thesis will deal with the prognosis of low flow indices in time and space. Spatial and temporal predictions will be attempted separately and combined. For each type of prediction, several methods will be compared and analyzed. The work is accordingly structured in the following way: after an overview of the state of the art in the field of temporal and spatial low flow prediction in chapter 2, the study area and base data utilized for analysis will be described in chapter 3. The three major parts of this thesis will be dedicated to the modeling approaches made for the three prognosis types. Chapter 4 will deal with temporal low flow estimation, subdivided into mere time series extrapolation and modeling using external variables, chapter 5 is designated to regionalization and chapter 6 is concerned with combining spatial and temporal prognoses within single models. For all three chapters, the applied methods, data and results will be presented and discussed in detail. In the subsequent chapter 7 the three main types of prognosis will be compared, i.e. regionalization of temporal prognoses will be contrasted to direct spatiotemporal modeling. The final chapter 8 will give a brief summary of the findings and discuss potential further advancement and shortcomings.

2   State of the art

The analysis of droughts and low flows has become a major discipline in hydrology and a number of reviews and guides about low-flow generating mechanisms, low flow estimation, description via indices, low-flow frequency analysis and regionalization have been published to date (e.g. Institute of Hydrology, 1980; Smakhtin 2001; Tallaksen and van Lanen, 2004; WMO, 2008; DWA, 2009). The following chapter presents a short literature review, featuring but a selection of many relevant studies for the three main parts of this study.

2.1   Temporal aspects of streamflow prediction

Depending on the anticipated use and field of study temporal prediction of streamflow can occur on a range of different time scales. Forecasting yields streamflow predictions from a couple of hours up to months ahead in time, while climate change impact analysis requires long-term streamflow projections under consideration of non-stationarity. The respective types of methods used for each type of analysis differ quite substantially.

Forecasting

Based on complexity and structure one can distinguish two major groups of forecasting models. The first group comprises a variety of stochastic approaches. There are stochastic time-series models, like autoregressive integrated moving average (ARIMA, e.g. Modarres, 2007) or seasonal autoregressive integrated moving average models (SARIMA, e,g. Fernandéz et al., 2009), which elongate the observed streamflow into the future via modeling of the underlying stochastic structure of the time series. Probability models like Markov chains base on the probability of a process being in a certain state given its previous state (e.g. Ochola and Kerkides, 2003). For these types of models forecasts can be obtained by mere evaluation of the past observations. Other models allow input of external data, like multiple linear regression (MLR). These models can be used to e.g. relate flow characteristics to a set of climatic variables (e.g. Liu and Juarez, 2001). Artificial neural networks (ANNs) have proven useful for short-term forecasts (e.g. Zealand et al., 1999), since they are capable of pattern detection and can handle non-linear relationships between input and target variable. When based exclusively on past data, Lin et al. (2006) have shown that support vector machines (SVM) outperform both ARMA and ANN models for long-term forecasts. Most recent forecast methods base on hybrids of machine learning techniques and stochastic methods, often in combination with soft computing elements.

The second group of forecasting models consists of physically based conceptual models. Their application is more expensive than that of the stochastic approaches and requires necessary input of climatic data, ideally of entire ensembles. The forecast ranges can be small to long

range and base on weather forecasts (e.g. Cloke and Pappenberger, 2009) or climate model forecasts (e.g. Wood et al., 2002).

Climate change impact analysis

Long-term streamflow prediction aims at providing projections of possible alterations, which is especially relevant for climate change impact analysis. The need for assessment of climate change impacts on low flows has been validated in a large number of studies worldwide. Douglas et al. (2000), for example, who carried out analyses on both flood and low flow time series over the entire USA found upward trends in the minimum 7-day average in the central northern regions. McCabe and Wolock (2002) who also analyzed streamflow in the USA over the period from 1941 to 1999 found increases in annual minimum daily streamflow, predominantly in the eastern parts of the country. Burn et al. (2010), who conducted a study on trend behavior in time series of several flood and low flow indicators for Canadian watersheds showed that low flow magnitude appeared to be both increasing and decreasing throughout the study area. Hisdal et al. (2001), who focused their investigation on deficit volumes found increasing deficits in Spain and the eastern parts of Eastern Europe, but also in the UK. Decreasing deficits were detected in Central and western Eastern Europe. Renard et al. (2008) detected decreasing floods and droughts in the Alps, caused by earlier snow-melt and glacier melting. Stahl et al. (2010) investigated low flow trends over entire Europe and detected decreasing streamflow between April and August in nearly all European catchments. One of their general observations was that low flows decrease where they occur in summer, as opposed to regions with lowest flows in other seasons. Special focus on mean and low flow for all German streams was put for example by Koehler (2008). He found that annual specific NM7Q (NM7q) appeared to decrease in the northern and increase in the southern half of Germany.

The classical chain of climate change impact analysis on streamflow begins with a global climate model (GCM) run on a certain emission scenario. The large-scale output needs to be downscaled, either dynamically or statistically, to a scale appropriate for regional analysis. The regional climate model (RCM) serves as input for the impact model, usually a conceptual rainfall-runoff model. In order to evaluate the uncertainty of prediction, ensembles of various climate models and scenarios should be applied. Analyses of this kind have been carried out in great numbers, also for low flow. De Wit et al. (2007) analyzed climate change impacts on critical low flows in the Meuse river basin using an RCM-driven hydrological model. They projected decreasing streamflow during the low flow period. Forzieri et al. (2014) used ensemble projections based on the IPCC SRES A1B scenario according to which droughts will intensify for almost all of Europe, except for northern and northeastern parts. Dai (2013) compared observed trends in drought and streamflow indices with model-simulated data and found discrepancies between the two based on the coupled models' inability to consider sea temperature variations. In an evaluation of uncertainties in the classical model chain for climate change impacts on low

flows Wilby and Harris (2006) found largest errors due to uncertainty of the climate change scenario and through downscaling of the GCMs, especially for statistically downscaled summer precipitation.

In view of these shortcomings Laaha et al. (2016) proposed the combination of information sources for assessment of climate change impacts on low flows. They named three sources, namely extrapolation of observed trends, rainfall-runoff model projections using climate scenarios, and extrapolation of trends in stochastic rainfall characteristics for use in rainfall runoff models. Tested on an Austrian data set, the extrapolation yielded high uncertainty, rainfall-runoff projections appeared to be sensitive to the calibration strategy and climate scenario, and numerous assumptions needed to be made for extrapolation of stochastic rainfall characteristics. Nevertheless, the authors strongly recommend the combination of various independent sources of information for a robust assessment of climate change impacts on low flows.

Approaches that differ from the classical impact assessment strategy include attempts to establish scenario-neutral impact studies like that of Yao et al. (2009), who used extrapolated long-term trends of temperature and precipitation as input for their water balance model. Prudhomme et al. (2010) based their impact study on sensitivity analyses of catchment responses to a plausible range of climate changes rather than directly on temporal model output for various scenarios. Vogel et al. (2011) proposed a non-stationary exponential trend model for annual maximum floods and determined decadal flood magnification factors, which represent the ratio of the magnitude of a future T-year flood to the current one. Ling et al. (2011) used a periodicity-trend superposition model to predict future changes in climatic variables and runoff without application of an impact model.

Another reference point for alternative modeling approaches may pose statistical relationships between streamflow metrics and climatic phenomena. Several studies have aimed at explaining observed peculiarities in flows by climatic phenomena, like Mosley (2000), who analyzed the influence of El Niño and La Niña effects among others on monthly lowest 7-day flow in New Zealand. He determined a major deviation of low flows from the normal in La Niña years. Also Ward et al. (2014) statistically analyzed El Niño effects on a global data set of annual peak discharges and found non-stationary relationships in many regions. Cunderlik and Burn (2004) investigated linkages between regional trends in monthly maximum flows and selected climatic variables in Canada. They found that increasing spring flows are related to shifts in snowmelt timing resulting from increasing spring temperatures, and that decreasing flows in fall are caused by increased evapotranspiration and changes in precipitation behavior in summer. Labat et al. (2004) correlated global annual runoff to temperature and found that a 1°C increase in the latter leads to a 4% increase in the former. In a different study Labat (2010) used a cross wavelet analysis to link continental streamflow variability to large-scale oscillations. Van Vliet et al. (2011) assessed climate impacts on daily water temperatures via a non-linear regression model

under input of discharge and air temperature. Haslinger et al. (2014) analyzed relationships between monthly streamflow anomalies and meteorological drought indices in 47 Austrian catchments. They found that correlations between streamflow anomalies and meteorological indices are high, especially within the low flow period. Van Loon and Laaha (2015) analyzed the dependence of streamflow drought duration and deficit volume on climatic indicators and catchment descriptors also in Austria. They found that the BFI is the main determinant of drought duration, indicating the dominance of catchment control, while deficits were primarily related to climatic factors. Liu et al. (2015) applied a climate-informed low-flow frequency analysis, i.e. they estimated the parameters of the GEV distribution fitted to the low-flow time series at their gauge in China as functions of a set of climatic indices. Consideration of all these climate-streamflow relationships could - under input of climate model data – yield meaningful prognoses of future flows.

All in all, the assessment of future changes in streamflow due to climate change and other factors, like anthropogenic interference, remains challenging. Assessment of the whole range of possible changes is important and can only be achieved through application of various modeling types and sources of information. Even though the uncertainty of trend extrapolation and analysis of climate-streamflow relationships is high, simple approaches may support or counteract findings based on coupled modeling and are worthy of consideration in the field of climate change impact analysis.

2.2 Streamflow regionalization

Regionalization of streamflow has long been identified as an important but challenging task in hydrology. Through the predictions in ungauged basins (PUB) initiative, launched by the International Association of Hydrological Sciences (IAHS) in 2003, awareness was raised toward the need for detailed process understanding for successful regionalization. A number of objectives and recommendations indispensable for constructive advancement in the field were listed, which include, among others, the identification of patterns in flow causing and forming factors on the basin scale, the investigation of links between processes in nested catchments, and the need for data driven approaches as the first step to model complex heterogeneous systems and thereby identify spatiotemporal patterns that cannot be understood from "a purely deterministic point of view" (Sivapalan, 2003).

Regression approaches

For a direct regionalization of flow statistics, multiple linear regression (MLR) approaches, which relate streamflow characteristics to climatic and physiographic catchment characteristics, have proven successful and readily applicable in many areas of the world over various conditions of data availability and sizes of study area. Hence, up to this date, regression probably remains the most common approach applied for regionalization of streamflow metrics of any kind. Sanborn

and Bledsoe (2006), for example, computed an entire 84 ecologically relevant streamflow metrics for 150 natural streams in the northwestern USA and related them to a large set of catchment characteristics via MLR. They found that performance of the regionalization is high for indices related to magnitude, timing and rate of change of flow, while prediction of streamflow variability was poorer. Hrachowitz et al. (2010) analyzed the effect of scale on the prediction of mean transit times and found that MLR models based on catchment characteristics remained unchanged over four orders of magnitude of catchment size.

Also for the specific case of low flow regionalization MLR has been applied and adjusted in a large amount of studies. Demuth (1993), for example, has given an extensive regional analysis of the suitability of MLR for low flow prognosis in Western Europe. He observed that a global regression approach is meaningful but recommended the inclusion of locally relevant catchment characteristics for smaller regions. Gustard et al. (1992) used MLR to relate the average annual NM7Q and the low flow discharge exceeded by 95% of all days (Q_{95}) to soil type classifications and obtained coefficients of variation of 0.61 and 0.57, respectively. Their analyses based on 633 catchments in the UK. Marechal and Holman (2005) used the same soil classification system to successfully predict low flow statistics exclusively on soil map data. Vogel and Kroll (1992) found multiplicative regression models for estimation of the NM7Q with 2 and 10 year return periods in Massachusetts. They identified strong relationships to area, slope, drainage density and baseflow recession for both target variables and obtained coefficients of determination of 0.97 and 0.95, respectively. Laaha and Blöschl (2006c) carried out regionalization of the specific Q_{95} (q_{95}) over 325 Austrian catchments using seasonality indices as predictors. Beyond the observation that seasonality is a major predictor for low flow in the study area, their main finding indicated that different regions require fitting of individual regression models in order to adapt to respective conditions. Another study by Laaha and Blöschl (2006b) suggests that exploitation of all information sources, i.e. streamflow records of variable lengths and combination of regionalization approaches, yields the best low flow predictions.

Instead of using single low flow metrics Yaeger et al. (2012) analyzed entire flow duration curves and the influence of external factors on their individual parts in the USA. They observed that the lower tails, i.e. low flow, are generally controlled by catchment characteristics, like landscape, soils and geology, while more average flows are primarily influenced by climate, as they represent the average catchment response. This finding supported the results of Gudmundsson et al. (2011) who determined in a spatial cross-correlation analysis that catchments are primarily linked to climatic variations during wet conditions and that the influence of local catchment properties becomes increasingly relevant with decreasing flow.

Regional frequency analysis

Regional frequency analysis extends the approach of spatial prediction from specific streamflow metrics to the estimation of entire streamflow distributions. The methods require regional

approximation of the distribution of flow. The regional distribution is adjusted to the local catchments via scaling by a certain index value or through local parameter estimates. For ungauged basins, index values and parameter estimates are obtained via regionalization. The methods have the major advantage of providing direct estimation of desired flow quantiles. A large number of studies have been carried out, where regional frequency distributions have been fitted in different ways, including maximum likelihood estimation, moment, and probability weighted moment estimation. Nathan and Mcmahon (1990b) found problems with all three methods for low flow regionalization, especially related to sample size. Since Hosking and Wallis' work on regional frequency analysis (1997), L-moments have probably become the most common method for distribution fitting. The goodness-of-fit is in many cases determined via L-moment diagrams or by the goodness-of-fit measure, also introduced by Hosking and Wallis. Independent of the method of induction the distributions identified as suitable for regional low flow analysis are manifold and differ greatly between study areas. Vogel and Kroll (1990) assessed the merit of regional frequency analysis for NM7Q regionalization in Massachusetts and fitted a 2-parameter lognormal distribution. Clausen and Pearson (1995) analyzed annual maximum streamflow drought in New Zealand and found that a 3-parameter log-normal distribution gave the best fit to both severities and durations. Durrans and Tomic (1996) found the log-Pearson III distribution as best fit for regional low flows in Alabama. Kroll and Vogel (2002) identified the 3-parameter lognormal distribution for low flows of non-intermittent regimes also in the USA. Zaidman et al. (2003) analyzed annual minimum series in the U.K and found that the type of the most suitable distribution depended on storage. For high storage catchments the General logistic (GLO) and General extreme value distribution (GEV) were identified, for low storage catchments the Pearson III and the GEV. Modarres (2008) found the Generalized Logistic suitable for NM7Q distribution in Iran and Shi et al. (2010) used the GLO distribution for NM7Q estimation in a karst region in China.

For extensive regional low flow analysis, also the regionalization of flow duration curves is useful, as done e.g. in Longobardi and Villani (2013). Pugliese et al. (2016) developed a top-kriging approach for the regionalization of index values for flow duration curves. They found that their method outperformed direct quantile regionalization via regression for very small quantiles.

Catchment grouping

For application of regional regression and for the formation of hydrologically homogeneous regions catchment classification is an important step. Spatial predictability increases significantly if models are applied to regions in which the hydrological processes are similar. Additionally, and most importantly, classification aids process understanding through identification of factors that determine differences in flow behavior. Ideally, classification is done on basis of catchment descriptors that are obtainable at any point of an area, allowing a straightforward assignment of ungauged catchments to a group. A large number of such classification schemes have been proposed, of which many have been tested in regional studies. Nathan and Mcmahon (1990a)

compared several clustering techniques for the identification of homogeneous sub regions for regionalization of low flow characteristics in 184 catchments in Australia. They identified a procedure that includes variable selection and weighting by stepwise MLR, clustering using several algorithms, and identification of the best cluster distribution via Andrews' curves. Also Hosking and Wallis (1997) suggested the use of cluster analysis of catchment characteristics for the identification of homogeneous regions from large data sets. Like Nathan and McMahon Hall and Minns (1999) found that the identification of homogeneous regions has to go beyond geographical proximity and contiguity in order to represent local hydrological processes. They tested Kohonen networks and fuzzy c-means on flood data in England and Wales, both methods resulting in similar clusters.

Laaha and Blöschl (2006a) compared various catchment grouping methods with respect to their performance in low flow regionalization using 325 Austrian sub-catchments. They found that for prediction of the specific q_{95} low flow seasonality is the ideal factor for catchment classification. Second best performance was achieved by a regression tree approach, third best by a residual pattern approach and worst by a weighted cluster analysis. Also Ouarda et al. (2006) found that seasonality is an important factor for delineation of homogeneous regions. Their study dealt with flood frequency analysis in Canada. The same methods as in Laaha and Blöschl (2006a) were tested for low flow regionalization in Italy by Vezza et al. (2010). They identified the regression tree as the best classification method for their Alpine study area.

Other methods for delineation of homogeneous regions include self-organizing maps (e.g. Lin and Chen, 2006; Ley et al., 2011; Toth, 2013; Farsadnia et al., 2014) and combinations thereof with other methods (e.g. Srinivas et al., 2008; Agarwal et al., 2016), boosted regression trees (e.g. Snelder et al., 2009), and canonical correlation analysis (e.g. Tsakiris et al., 2011). Instead of determining a fixed set of regions for regionalization, so-called region of influence approaches may be applied, determining a dynamic hydrologically similar neighborhood for every unobserved point. Region of influence methods have been applied by e.g. Burn (1990), Zrinji and Burn (1994), Holmes et al. (2002) and Eng et al. (2007).

Geostatistics

Classical geostatistical methods for regionalization are highly effective for point data continuous in space. Due to the nature of flows, being restricted to the stream network and having observations that pose an accumulation of processes over the belonging catchment area, the application of common distance-based kriging approaches for their regionalization is bound to fail. Nevertheless, special consideration of the peculiarities of streams and respective adaptation of the approaches provides a whole new array of methods for streamflow regionalization. One adaptation of the geostatistical methods to the stream network problem is the inclusion of external information via catchment descriptors. Haberlandt et al. (2001) tested ordinary kriging (OK) and kriging with external drift (EDK) for simulated baseflow index regionalization on 114 subbasins of the Elbe and compared it with MLR. They found that EDK performs slightly better

than MLR using the same catchment descriptors as input and that both methods outperformed OK by far. Merz and Blöschl (2005) found that a regional flood frequency analysis via regionalization of flood moments in 575 Austrian catchments functioned best when EDK was applied with catchment descriptors as external drift and consideration of record lengths of the observations. For their study, OK showed even better performance than MLR.

Still, the inclusion of catchment descriptors does not resolve the issues related to increased similarity of catchments due to their connectedness. Skøien and Blöschl (2007) proposed a topological kriging method that adapts OK to the problem of streams. They account for nestedness and differences in spatial extent of catchments by using a block rather than a point kriging and by regularization of the underlying variogram. The method has been applied by Laaha et al. (2014) for low flows in Austria with great success, i.e. a coefficient of determination of 0.75 for q_{95} prediction. He found that downstream river stretches can be predicted much more successful via top-kriging than via MLR. For headwater catchments the performance of both methods was comparable. Laaha et al. (2013) extended the top-kriging method by incorporation of an external drift for regionalization of stream temperatures and recommended the inclusion of external drift for spatially non-stationary conditions. Merz et al. (2008), who used top-kriging for flood moment regionalization, additionally took several regional controls and uncertainty of data into account.

A different approach, termed physiographical space-based interpolation (PSBI), was introduced by Chokmani and Ouarda (2004). This method applies interpolation in physiopgraphical rather than geographic space. This very space is made up by multivariate combinations of climatic and physiographic catchment descriptors. Castiglioni et al. (2009) applied the method to low flow indices in 51 catchments in central Italy. They built the physiographic space using principal component analysis and tested deterministic and geostatistical techniques for interpolation therein. They found that geostatistical interpolation in physiographic space outperforms a standard regression approach. Deterministic techniques were less successful than regression. In a subsequent study Castiglioni et al. (2011) used the same data base for comparison of PSBI and top-kriging and found that top-kriging performed better for larger river branches, while PSBI was more suitable for prediction of headwater catchments. Archfield et al. (2013) compared PSBI and top-kriging to MLR for flood data regionalization in the southeastern USA and found that top-kriging substantially outperforms MLR and PSBI, especially for large catchments. A combined approach of top-kriging and PSBI lead to a minor improvement of the spatial predictions. They identified the high spatial correlation between neighboring gauges as major issue for regression-based approaches. Castellarin (2014) extended the PSBI regionalization by a third dimension representing streamflow duration in order to predict flow duration curves in ungauged catchments.

Other approaches

The set of other possible methods for streamflow regionalization is unlimited and a variety of approaches have been proposed. Promising is, for example, the application of artificial neural networks (e.g. Dawson et al., 2006; Ouarda and Shu, 2009). Due to their non-linearity and generality they tend to outperform linear regression models. Shu and Ouarda (2008) proposed an adaptive neuro-fuzzy inference system (ANFIS) for flood quantile estimation as an extension of the ANN approach. Castellarin (2007) used probabilistic envelope curves for estimation of design floods and found results comparable to the index flood approach. Other approaches include kriging-based map-correlation methods (Archfield and Vogel, 2010), pair-wise empirical copula densities (Samaniego et al., 2010), and model tree ensembles (Schnier and Cai, 2014). A comparative world-wide study of flood and low flow regionalization methods by Salinas et al. (2013) revealed that predictions by any method are usually poorer in arid climates than in humid regions and more accurate in larger catchments. They also found that regional regression performs better than global regression but that regression approaches perform worse than other approaches, especially geostatistical ones.

Of course there exist a variety of simpler approaches for estimation of streamflow at ungauged sites, especially within the same river, including simple transfer functions or power models. Also, as opposed to direct regionalization of streamflow metrics and regional frequency analysis, there is continuous hydrological modeling. Regional prediction with these models usually requires transfer of parameters to the unobserved locations (e.g. Merz and Blöschl, 2004). Also, simultaneous multivariate regionalization of various drought or low flow characteristics, like duration and volume, e.g. via copulas has become popular (Zhang et al., 2015).

2.3 Spatiotemporal considerations

According to Ouarda et al. (2008), who carried out a detailed review on statistical models for local and regional low flow estimation it is "no longer realistic to ignore change signals in low-flow regimes when carrying out local and regional estimation activities". This is especially valid for regional frequency analysis. Intrinsically, the approach brings on many advantages when compared to single-site analysis, the most prominent one being the concept of "trading space for time", which allows assessment of streamflow distributions at unobserved or poorly recorded locations based on well-observed neighboring stations that exhibit similar hydrologic behavior. However, the concept of transferring information from one stations' record to another can only be successful under stationary conditions or within the same period of time. Gathering information from pooled streamflow records of different times with different underlying distributions will not yield reliable predictions.

A range of studies have dealt with this very problematic. Leclerc and Ouarda (2007) carried out a non-stationary regional flood frequency analysis. They fitted non-stationary general extreme

value distributions at observed sites and grouped the study area in southeastern Canada according to the respective numbers of parameters required. Locally derived flood quantiles were then regionalized using MLR including climatic and physiographic external information. Their non-stationary approach gave better predictions during cross validation than a comparative stationary procedure. Before, Cunderlik and Burn (2003) have proven that ignorance of non-stationarity can heavily bias quantile prediction even for near time horizons of 0-20 years ahead. Their approach for assessment of non-stationarity based on separation of regional quantile functions into time-dependent local components and time-independent regional components. Cunderlik and Ouarda (2006) later adapted a flood-duration-frequency analysis to non-stationary conditions using regional trend analyses for identification of time varying model parameters. Lima and Lall (2010) applied a hierarchical Bayesian model to assess parameter uncertainty and non-stationarity in a regional flood frequency analysis in Brazil. They used the method to identify scaling relationships of maximum flows with area and estimated non-stationary parameters for the Gumbel distribution.

Instead of merely assessing the time dependence of model parameters, approaches have been developed that link parameter values to external variables. Lopez and Frances (2013) proposed a non-stationary framework for flood frequency analysis based on generalized additive models for location, scale and shape (GAMLSS), in which distribution parameters are modeled as functions of time and of climate indices. Delgado et al. (2014) directly linked a non-stationary flood-frequency model based on a non-stationary log-normal distribution to downscaled climate model data by estimating the scale parameter as a function of a derived circulation index.

Even though the above studies all deal with the spatiotemporal estimation of floods, the application of non-stationary frequency analyses provides the opportunity to also assess future low flow distributions in space and time. The important steps include the identification of proper regional distributions, assessment of the required degree of non-stationarity, i.e. the number of time varying parameters, effective modeling of the time-dependence of the parameters, either as a function of time, climatic variables or other external information, and analysis of the stationarity of the similarity criteria, i.e. potential time dependence of homogeneous regions.

3   Study area and data

This chapter will give a general overview of the whole study area and the base data used for all methods. An exact apportionment of the area and required data pre-processing steps will be explained in detail in the respective chapters of this work, as data requirements and availabilities differ.

3.1   Study area

The area under investigation is the federal state of Lower Saxony, in the northwest of Germany. Its total area of 47,634 km^2 extends from the North Sea in the northwest to the Harz mountains in the south east. The largest portion of Lower Saxony is part of the Northern German Plain, where the terrain is generally shallow. In the south follow the Weser Uplands and eventually the secondary mountains of the Harz. Accordingly, elevations range from 2.5 m below sea level in the north to 971 m above sea level in the Harz. The topography is depicted in Figure 3.1.
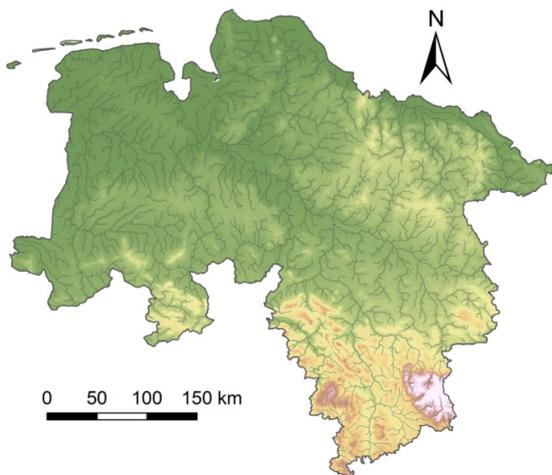
Figure 3.1: Study area.

Lower Saxony lies within the temperate zone and is subject to a transition from the coastal maritime to the eastern European continental climate. Accordingly, temperatures and precipitation depths show corresponding gradients. The annual average temperature is 8°C, while temperature amplitudes are far lower in the northwestern parts of the state that are affected by the Atlantic than for the eastern, more continental areas. A similar pattern holds for precipitation. While annual rainfall depths in the northwest range around the Lower Saxon mean of 754 mm, less than 550 mm are found in the very east. Highest rainfall amounts, however, occur due to luv effects in the Upper Harz, with more than 1,300 mm per year. On a monthly basis, precipitation amounts are generally smallest in February, with an average 46 mm, and

highest in June with a mean of 79 mm. All values are based on the normal period from 1961 to 1990 (Nds. MU, 2012).

The predominant land use in Lower Saxony is agriculture with about 60% of its total area. Approximately 22% are made up by forest. In recent years, there has, however been a decreasing trend in both agricultural and forest area towards an increase in settlements and open space (NLS, 2004). Primarily, sandy soils are found in Lower Saxony, especially in the Northern and Central parts. Only in the southeast and at the coast, loamy soils prevail.

The main rivers in Lower Saxony are, from east to west, the Ems, the Weser and the Elbe. The river basin of the Weser thereby makes up the largest portion of the state with 29.440 km². Significant parts of the Weser watershed are constituted by its largest tributary, the Aller, and the Leine in the southeastern lobe of the state. Further larger rivers include the Hase as a tributary of the Ems, the Hunte and the Wümme, both flowing into the Weser, and the Oste in the northeast, joining the Elbe.

3.2 Data

3.2.1 Discharge data

Basis for all analyses pose daily time series of average discharge. In total, there are 353 flow records available for gauges in and around Lower Saxony. These gauges are under operation of the Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz (NLWKN), the Harzwasserwerke, the Hamburger Wasserwerke, the Oldenburgisch-Ostfriesischer Wasserverband (OOWV) and the Wasser- und Schifffahrtsdirektion (WSD). Record lengths are highly variable and range from 7 to 193 years. The positions of all available stations can be found in Figure 3.2.
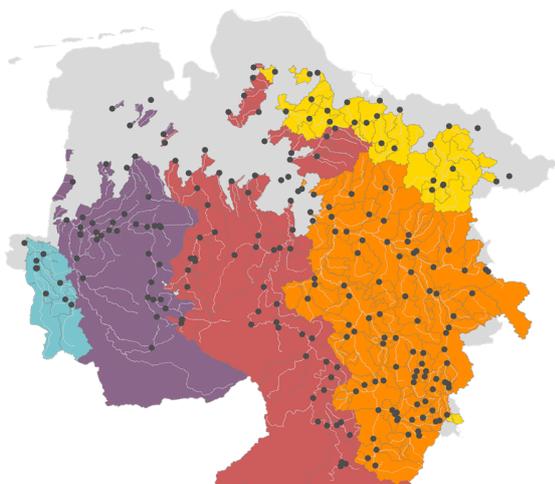
Figure 3.2: Available discharge gauges and belonging catchments within the main watersheds.

Screening of the data for continuity has been conducted. Series with periods of missing data (at least 2 consecutive days within the low-flow relevant period) have been rejected. Since low flows are highly sensitive toward human interference, detailed research has been carried out to identify gauges whose low flow behavior is rather unnatural. The analyses include literature research, plausibility checks of the hydrographs, homogeneity analyses, outlier checks before and during model fitting, and qualitative assessment of anthropogenic influence via maps. The numbers of stations selected for the particular analyses as presented later are those that remain after screening. A list of all stations used in the different parts of this work can be found in Appendix A1.

In order to increase the data set for specific analyses requiring long time series data of more than 100 years, discharge gauge data from the Global Runoff Data Base of the Global Runoff Data Centre (GRDC, 2016) for entire Europe has been included. The data has been subjected to equal screening procedures as the data set for Lower Saxony. The available stations are shown in Figure 3.3. A list of the stations in the data set can be found in Appendix A2.
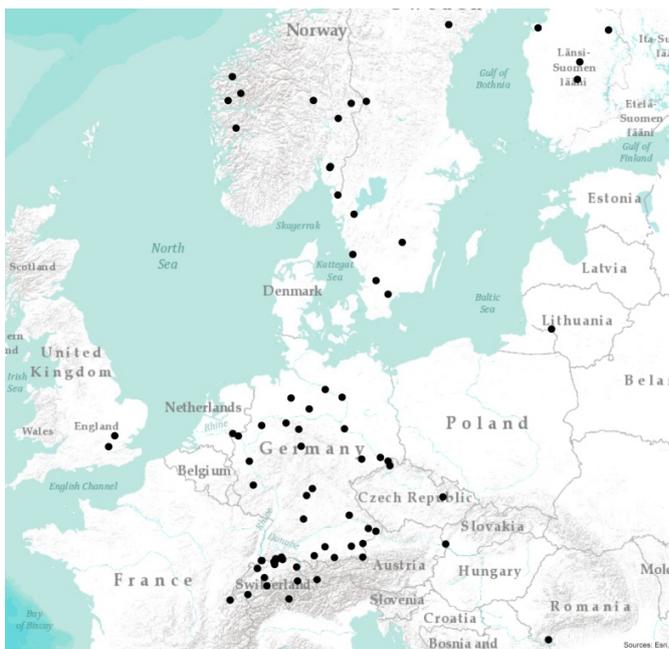
Figure 3.3: GRDC gauges with record lengths for daily discharge equal to or above 100 years.

3.2.2 Climate data

The meteorological data used in this work comprises of three main variables: precipitation, temperature and global radiation. For all variables daily time series are considered. In order to be able to make the data available continuously in space, interpolated data is utilized for which all three variables are available on a 1x1-km raster for the total area of Lower Saxony. Input for

regionalization is station data of 771 precipitation gauges, and 123 stations for other climatic variables. Regionalization had been carried out for a period of 1951 to 2010. The regionalization process is described in detail in NLWKN (2012).

3.2.3 Physiographic catchment characteristics

A wide range of physiographic catchment descriptors (CDs) has been employed in the analyses. These are primarily derived from raster data of various resolutions, as well as from Polygon shape files. Topographic descriptors are derived from a 1x1-km DEM (Jarvis et al. 2008), soil properties from the Soil Map of the Federal Republic of Germany 1:1000000 (BÜK1000; BGR, 2013), landuse information is obtained from the CORINE land cover map (CLC2006; Umweltbundesamt, 2009), permeability of near-surface rock layers and hydrogeological conductivity of aquifers from the Geological Map 1:500000 (GÜK500; LBEG) and variables related to the river network from shapefiles. The considered CDs are listed in Table 3.1.

Table 3.1: List of catchment descriptors.

Catchment descriptor	Unit	Catchment descriptor	Unit
Catchment area	km²	Percentage of urban area	%
Minimum, mean, maximum elevation	m	Percentage of agricultural area	%
Minimum, mean, maximum slope	%	Percentage of forest	%
Effective slope	m/km²	Percentage of porous aquifers	%
Usable field capacity	mm	Percentage of karst aquifers	%
Field capacity	mm	Percentage of poorly conductive aquifers	%
Air capacity	mm	Hydraulic conductivity	m/s
Total pore volume	mm	River network density	-
Clay content	%	Circumference/area	-
Silt content	%	Position in main river	km
Sand content	%		

3.3 Indices

Basis for all spatial and temporal models are so-called indices. Indices are values calculated for a time series that represent a certain characteristic of this series within a specific period of time (e.g. a year or a season). In the process of index calculation, the temporal resolution of the original time series is thus broken down to an index time series of much smaller resolution, facilitating a superior focus on the relevant parts of the series for a specific purpose.

In this study, the purpose is the analysis of low flows and their behavior. The selected low flow indices therefore represent different quantities of interest for low flow analysis (e.g. intensity,

duration and timing of low flow periods). Additionally, meteorological indices will be computed that potentially influence the low flow behavior (e.g. average rainfall, dry spell durations, heat wave durations) and can be used as model input for low flow prediction. Basis for index calculation pose the daily time series of discharge and climatic variables.

3.3.1 Low flow indices

Target variables for all temporal and spatial models are low flow characteristics or low flow indices. These are annual values calculated from the daily series of average discharge. The low flow indices used in this study are shown in Table 3.2. The low flow indices are computed for the summer half year only in order to exclude winter low flows, which are subject to different underlying processes and may thus disturb proper model calibration. Figure 3.4 shows the average number of days over all stations on which the discharge is below three different long term quantiles. It can clearly be seen that the majority of days on which the thresholds are not exceeded lies between June and October.
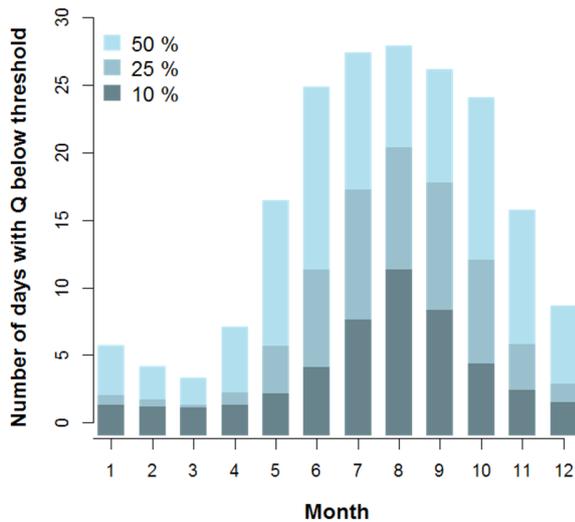
Figure 3.4: Number of days per month with discharge below the long-term 10-, 25- and 50-% quantile averaged over all stations.

Calculation of the indices is straightforward. For every consecutive summer half year, the indices are extracted individually. The 5- and 20-percent non-exceedance quantiles (Q_{95}, Q_{80}) are computed empirically, according to

$$Q_{1-x} = \begin{cases} \frac{1}{2}\left(y_{n*x} + y_{n*x+1}\right) & \text{, if } n*x \text{ integer} \\ y_{floor(n*x)} & \text{, if } n*x \text{ not integer} \end{cases} \text{,} \qquad (3.1)$$

where x represents the desired non-exceedance percentage, y is the ordered time series, and n is the sample size. The lowest average flows of 7 or 30 days (NM7Q, NM30Q) are simply obtained as

$$NMxQ = min\left(\frac{1}{x}\sum_{i=j}^{j+x-1}y_i \ , for \ j=1,...,n\text{-}x+1\right).$$

(3.2)

For computation of low flow duration and deficit volume, a fixed threshold approach is selected. For the entire flow record the 20-percent non-exceedance quantile is calculated according to equation (3.1). This quantile is treated as a threshold. All days with flows below this threshold are counted. The maximum low flow duration D_{max} simply equates the maximum number of immediately consecutive days within a year. Discontinuities of up to two days are neglected. The average low flow duration D_{mean} corresponds to the total number of days below the threshold divided by the number of events. An event is defined as a period of consecutive days below the threshold, given the same constraints as above. An event needs to consist of at least three days to be counted. The maximum and mean deficit volume (V_{max}, V_{mean}) are calculated in the exact same way but instead of counting days, the difference between the threshold and a day's actual flow are summed up.

Table 3.2: List of low flow indices.

Index	Unit	Description
Q_{95} Q_{80}	m³/s	5- and 20-percent non-exceedance quantile of the daily average discharge
NM7Q NM30Q	m³/s	Lowest 7- and 30-day average flow
V_{max} V_{mean}	m³	Maximum and mean deficit volume: sum of daily discharge below the long term 20-percent non-exceedance quantile
D_{max} D_{mean}	d	Maximum and mean low flow duration: number of days with discharge below the long term 20-percent non-exceedance quantile
timing	-	Day of the year at which smallest daily flow occurs

The low flow indices selected for this study represent a small fraction of possible indices but cover quite a range of relevant low flow quantities for water resources management and planning. Non-exceedence quantiles represent the overall low flow situation in a year without specific relation to a certain event. The same holds for average volume and duration. NMxQ, maximum volume and duration, as well as timing, on the other hand, focus on the greatest events per year in their respective terms and characterize them accordingly. The consideration of variations within the individual index types (e.g. Q_{95} and Q_{80}) is not solely owing to different management requirements but used to investigate the various models' capability to model both

more and less extreme low flows. Since detailed consideration of each variable in every model approach will not be feasible in the scope of this work, main investigations will be carried out using the NM7Q. Comparison with other indices will be made in each case for the best performing model and when relevant.

3.3.2 Meteorological indices

Input for most of the temporal and spatiotemporal models pose meteorological indices. These are, like their low flow counterparts, annual values calculated from daily time series of precipitation sums, mean and maximum temperature, as well as global radiation. Table 3.3 summarizes the applied indices and gives a short description. The computation of the meteorological indices will not be described in detail here, but can be found in Fangmann (2012).

Table 3.3: Meteorological indices based on precipitation, temperature and potential evapotranspiration.

Index	Unit	Description
P_{mean}	mm/d	Average daily precipitation
P_x	mm/d	Non-exceedance quantiles of daily precipitation sums
SPI	-	Standardized Precipitation Index: standardized deviation of accumulated precipitation sums from the long term normal (Mckee et al., 1993)
DSD_{mean} DSD_{max}	d	Mean and maximum dry spell duration: number of days with precipitation <1mm/d
WSD_{mean} WSD_{max}	d	Mean and maximum wet spell duration: number of days with preciptation >= 1mm/d
T_{min} T_{mean} T_{max}	°C	Minimum, mean and maximum daily average temperature
HWD	d	Heat wave duration: Number of days above 90-percent non-exceedance quantile of maximum temperature calculated for each specific day of the year
ETP_{mean}	mm/d	Average daily potential evaporation calculated according to Turc-Wendling
$P\text{-}ETP_{mean}$	mm/d	Average climatic water balance: precipitation minus potential evapotranspiration
SPEI	-	Standardized Precipitation Evaporation Index: standardized deviation of accumulated climatic water balance from long term normal (Vicente-Serrano et al., 2010)

While the low flow indices are calculated for a fixed period within the year (the summer half year), the period for computation of the meteorological quantities is varied in length and position. Figure 3.5 shows the scheme according to which every meteorological index has been computed. I_{NQ} represents the low flow index, calculated for its fixed period within a given year. I_M denotes the meteorological index. The first number in the index notation represents the length of

the base period for calculation (eg. 3, 6 or 12 months), while the second one indicates the lead time, i.e. the number of months the period is shifted back in time relative to the low flow index. A 0-shift indicates that the calculation period ends simultaneously with the low flow period.
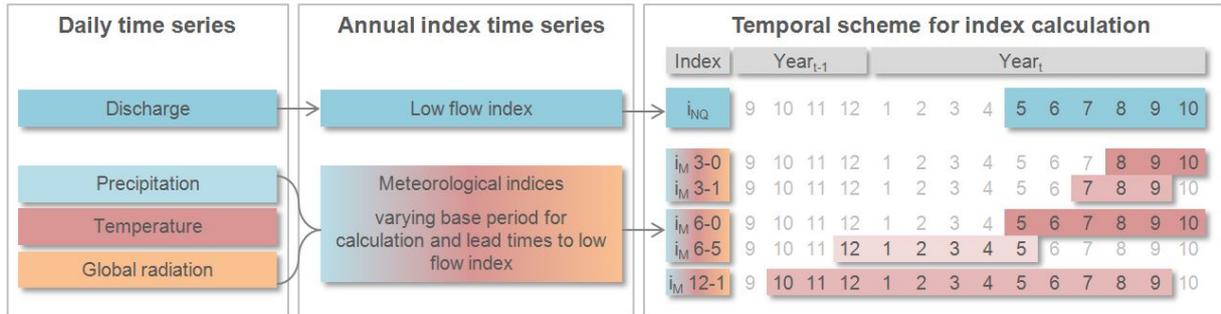
Figure 3.5: Calculation scheme for low flow indices with fixed base period and meteorological indices with varying base period and lead times relative to the low flow calculation period.

4 Temporal low flow modeling

This chapter focuses on the temporal long-term prediction of low flows. Two main approaches will be analyzed: a) extrapolation of patterns recognizable in the observed low flow time series to the future and b) modeling of the low flow indices based on input of meteorological indices. For both approaches, several methods have been tested, a selection of which is presented and discussed in this chapter.

4.1 Methods

4.1.1 Time series extrapolation

The first approach at predicting future low flow behavior aims at making use of observable patterns in past low flow time series and extrapolating these into the future. An exact prognosis of yearly values will not be possible in this way, but changes in the mean or different quantiles of low flow indices may potentially be mapped with certain accuracy. The major advantage of the methods presented in this section is their independence from external input data, like climate model data, which are of unknown uncertainty.

4.1.1.1 Linear trend extrapolation

The first method for time series extrapolation is also the simplest one. A linear trend is estimated for the observed annual low flow values and extrapolated into the future. The trend is estimated non-parametrically according to Theil-Sen (Theil, 1950 a-c, Sen, 1968). Its slope b is calculated as the median of the differences between all pairs y_i and y_j at times t_i and t_j:

$$b = median\left(\frac{y_i - y_j}{t_i - t_j}\right) \quad with \ t_i < t_j.$$

(4.1)

The intercept a can now be computed as

$$a = median\left(y_i - bt_i\right).$$

(4.2)

The reason for choosing a non-parametric estimator is simple: the only explanatory variable in this case is time, which is not capable of explaining changes in the variance of the target variable. Consequently, heteroscedasticity in the residuals will be a major issue and thus restricts the use of parametric methods like ordinary least squares regression. The Theil-Sen estimator is simple and conveniently applicable over a wide range of time series without previous or subsequent testing and modification.

In order to assess whether a fitted trend is actually significant, a Mann-Kendall trend test with prior trend-free pre-whitening is applied to the time series. Only significant trends are extrapolated into the future, otherwise, a stationary mean is assumed.

Trend free pre-whitening (Yue et al., 2002) of the data is recommended due to the nature of trends. Trends mimic autocorrelation in time series, while actual autocorrelation increase the probability of trend detection by tests (von Storch, 1995). Thus, the best possible way to test for a trend in a time series is to firstly estimate the trend slope, which is done here according to Theil-Sen (equation (4.1)), and substract it from the time series

$$y'_t = y_t - bt \ .$$

(4.3)

The residual time series can then be tested for lag-1 autocorrelation by calculating the serial correlation coefficient

$$r_1 = \frac{\frac{1}{n-1} \sum_{t=1}^{n-1} (y'_t - \overline{y'})(y'_{t+1} - \overline{y'})}{\frac{1}{n} \sum_{t=1}^{n} (y'_t - \overline{y'})^2} \ ,$$

(4.4)

where $\overline{y'}$ is the mean of the trend-free time series and n the number of observations. Significance of r_1 is tested based on the 95% confidence interval

$$\pm \frac{1.95}{\sqrt{n}} \ .$$

(4.5)

If the coefficient lies outside of these bounds, it is removed from the trend free time series via

$$y''_t = y'_t - r_1 y'_{t-1} \ .$$

(4.6)

Finally, the trend is added back and a pre-whitened series is obtained, which is tested using the non-parametric Mann-Kendall trend test (Kendall, 1975). This test bases upon the null hypothesis that no trend is present in a time series. The test statistic S is calculated as the sum of signs between all chronologically ordered data pairs

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(y_i - y_j) \ .$$

(4.7)

For determination of the p-value of the test, the variance of S needs to be known. It can be computed as

$$s^2(S) = \frac{n(n-1)(2n+5) - \sum_{j=1}^{n_t} t_j(i)(i-1)(2i+5)}{18},$$

(4.8)

where $t(i)$ represents the number of ties of extent i. With help of the variance the test statistic is standardized

$$Z_S = \begin{cases} \dfrac{S-1}{\sqrt{s^2(S)}} & for\ S > 0 \\[2mm] \dfrac{S+1}{\sqrt{s^2(S)}} & for\ S < 0, \\[2mm] 0 & for\ S = 0 \end{cases}$$

(4.9)

and p is obtained as

$$p = 2\left(1 - \Phi\left(|Z_s|\right)\right),$$

(4.10)

where Φ denotes the cumulative distribution function of the normal distribution.

4.1.1.2 Support vector regression

Support vector regression (SVR) is selected as an extension of the linear trend extrapolation, in order to be able to model potential non-linear changes over time.

SVR is a modification of the so-called Support Vector Machines (SVM), a machine learning technique initially developed for classification (Vapnik and Chervonenkis, 1974). The principle of this method is the separation of objects within a multidimensional vector space by a hyperplane. This hyperplane is fitted to a set of training data with known membership of its data points to one of two classes. The class separating plane is positioned in a way that its distance to all vectors of both classes is maximized, making the technique a Large Margin Classifier. This enables a reliable general assignment of objects beyond the training data set to the respective classes. Conveniently, in order to achieve a maximum distance separation only the vectors closest to the hyperplane, the so-called support vectors need to be considered.

Since in most cases data is not linearly separable, the input data is mapped onto a higher-dimensional space where a linear hyperplane suffices for proper separation of the training data. This is done using kernel functions. Through backtransformation to the original feature space, the linear hyperplane becomes non-linear.

In SVR (Vapnik, 1995), which extends the SVMs from classification to regression, the hyperplane is positioned to match the training data x_i. In linear terms the function is described via

$$f(x) = \langle w, x \rangle + b \; , \tag{4.11}$$

where w is its normal and b a bias term. Identification of the optimal function is subject to two opposed constraints: a) the deviation of the function to the target values y_i should be small and b) the generality of the fitted function, i.e. its capability to predict future data should be large (avoidance of overfitting). The constraints can be accounted for by setting a value ε up to which a deviation of the function to the target values is ignored. Within this defined range, the function is fitted - as in SVM - with maximum possible margin, i. e. the deviation to the target values is maximized. Tuning of ε can thus create the desired balance between accuracy and generality of the model.

Mathematically, the optimization problem for a given ε can be formulated as follows: maximization of the margin is equivalent to maximization of the flatness of the fitted function, which is achieved by minimizing its norm $||w||^2$. Simultaneously, it needs to be assured that deviations of the function to the target values y_i remain in the range of ε. The problem can be written as

$$\text{minimize} \quad \frac{1}{2} ||w||^2$$

$$\text{while} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} . \tag{4.12}$$

In case the optimization problem is not solvable under the given constraints, one may allow for errors by introducing slack variables ξ_i, ξ_i^* to the problem. These variables take the values of the so-called ε-insensitive loss function $|\xi|_\varepsilon$, i.e.

$$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{4.13}$$

represent the deviation of the function from the training data larger than ε. As these errors should be as small as possible, their sum is also subjected to minimization. The optimization problem becomes:

$$\text{minimize} \quad \frac{1}{2} ||w||^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

$$\text{while} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} . \tag{4.14}$$

The constant C balances the amount of acceptable error and the flatness of the fitted function. Proper adjustment of both parameters ε and C is hence crucial for the tradeoff between fit and generality of the model.

The convex optimization problem can be solved in its dual form, which can be derived using Lagrange multipliers and Karush-Kuhn-Tucker conditions. For details see e.g. Smola and Scholkopf (2004).

Eventually, the optimal function can be described as

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b \qquad (4.15)$$

with α_i and α_i^* being Lagrange multipliers. It should be noted that w no longer needs to be computed for function evaluation and can be fully represented by the dot products between the training data $\langle x_i, x \rangle$.

As stated above, the algorithm is capable of fitting non-linear functions by mapping data onto higher-dimensional feature spaces, which can be achieved using kernel functions. The kernel function is defined as a linear dot product of a non-linear mapping of the training data

$$k(x_i, x) = \langle \varphi(x_i), \varphi(x) \rangle. \qquad (4.16)$$

The kernel function needs to be capable of reproducing the dot product in some feature space, the mapping itself does not need to be known. Thus, the optimal function becomes

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b. \qquad (4.17)$$

Which kernels are suitable to represent dot products between training patterns in some feature space is discussed in detail in Smola and Kondor (2003). All analyses in this study will be carried out using radial basis function kernels of the following form:

$$k(x_i, x) = exp\left(-\gamma \|x_i - x\|^2\right). \qquad (4.18)$$

The γ-parameter thereby controls the spread of the kernel and $\|x_i - x\|$ denotes the Eucledian distance between the feature vectors.

Apart from offering a non-linear option and potentially greater generality than linear regression, SVR is far less restricted through statistical preconditions. All analyses regarding SVR are carried out using the "e1071" R package by Meyer (2017).

4.1.1.3 Empirical mode decomposition

Additionally to fitting above models to the original time-series data, extrapolation of individual time series components is tested. Decomposition of the series of low flow indices is thereby achieved via empirical mode decomposition (EMD).

EMD is an integral part of the Hilbert-Huang transform (HHT; Huang et al., 1998), a transform method analogous to the Fourier transform. Like the latter, the HHT aims at decomposing a signal into its different frequencies. The difference lies within the premises of the two methods: while the Fourier transform is applicable to stationary signals only, the HHT is specially designated to the analysis of non-stationary data by considering instantaneous frequencies, i.e. the final spectrum remains in the time domain.

In order to be able to obtain instantaneous frequencies, the original signal needs first to be separated into components, the so-called intrinsic mode functions (IMFs), which then can be converted into instantaneous frequencies via Hilbert spectral analysis. For the analyses in this work, only the first part of the HHT is required, namely the decomposition of an index time series into various IMFs via EMD.

An IMF has to fulfill specific criteria in order to be viable for spectral analysis. A detailed description of the HHT and derivation of the prerequisites for computation of instantaneous frequency can be found e.g. in Huang et al. (1998). Here, the two prerequisites will be explained briefly.

The instantaneous frequency ω

$$\omega = \frac{\partial \theta(t)}{\partial t} \tag{4.19}$$

with

$$\theta(t) = arctan\left(\frac{Y(t)}{X(t)}\right), \tag{4.20}$$

where $X(t)$ is the original time series and $Y(t)$ its Hilbert transform, is a single value function of time, i.e. at any given time only one frequency value exists. This entails that it can describe a monocomponent signal only. This precondition is met with the narrow band criterion. In order for a signal to be narrow band, the number of zero crossings and extrema needs to be equal, which excludes riding waves from the signal. The second prerequisite for deriving instantaneous frequency from a signal is that it only yields positive frequency values after transformation. This requires the signal to be locally symmetric over the mean. An eligible signal thus consists of only one mode of oscillation, but may still consist of various amplitudes and frequencies. Based on these criteria, the process of EMD is derived.

The process of deriving IMFs from the original signal is called sifting. Sifting is carried out as follows: at first, all local extremes of the original series are identified. The maxima are connected via interpolation. The same is done for the minima and an upper and lower envelope, respectively, are obtained. The mean line (m_1) between the envelopes is computed. In order to fulfill the preconditions of an IMF, the mean needs to be 0 at all times and the number of maxima and minima needs to be equal. If this is not the case, m_1 is substracted from the original series $X(t)$

$$h_1 = X(t) - m_1 \qquad (4.21)$$

and h_1 becomes the input data for the next iteration. The iteration is repeated k times until

$$h_k = h_{k-1} - m_k = IMF_1 \qquad (4.22)$$

fulfills the required criteria and thus constitutes the first IMF. Figure 4.1 gives a schematic overview over the sifting process for an IMF. It can be seen that with every step of the process the upper and lower envelope become more and more symmetric so that the mean approaches 0. Since perfect symmetry is impossible to achieve for an IMF and the sifting algorithm requires a stopping point to keep computation times low, several criteria can be considered for acceptance of an IMF. In this case, the algorithm is stopped, when the mean goes below

$$x_{tol} = s(X(t)) * 0.1^2 \qquad (4.23)$$

with s as the standard deviation. The sifting process for the next IMFs is conducted for the residual series, i.e. the original series minus all previous IMFs. The process is repeated until the residue becomes monotonic, i.e. no further local minima and maxima can be determined. The final residue is treated as the trend of the original time series. The resulting IMFs are ordered from smallest to largest period and are complete, i.e. adding up all individual components will yield the original time series.

There are some factors that need to be considered during EMD. One major issue is so-called end effects, i.e. fitting of the envelopes at the beginning and the end of a signal. An improper envelope leads to unrealistic IMF values. Furthermore the choice of boundary conditions for interpolation of the extremes may become an issue when creating unanticipated extremes. Both constraints need to be considered during sifting.

The individual IMFs uncover underlying processes in the investigated time series, which in combination with the above extrapolation methods may yield better prognoses than extrapolation of the original series itself. Ideally, white noise will be filtered out, revealing stochastic processes that can be appropriately modeled.

Empirical mode decomposition is carried out using the "EMD" package for R (Kim and Oh, 2014).
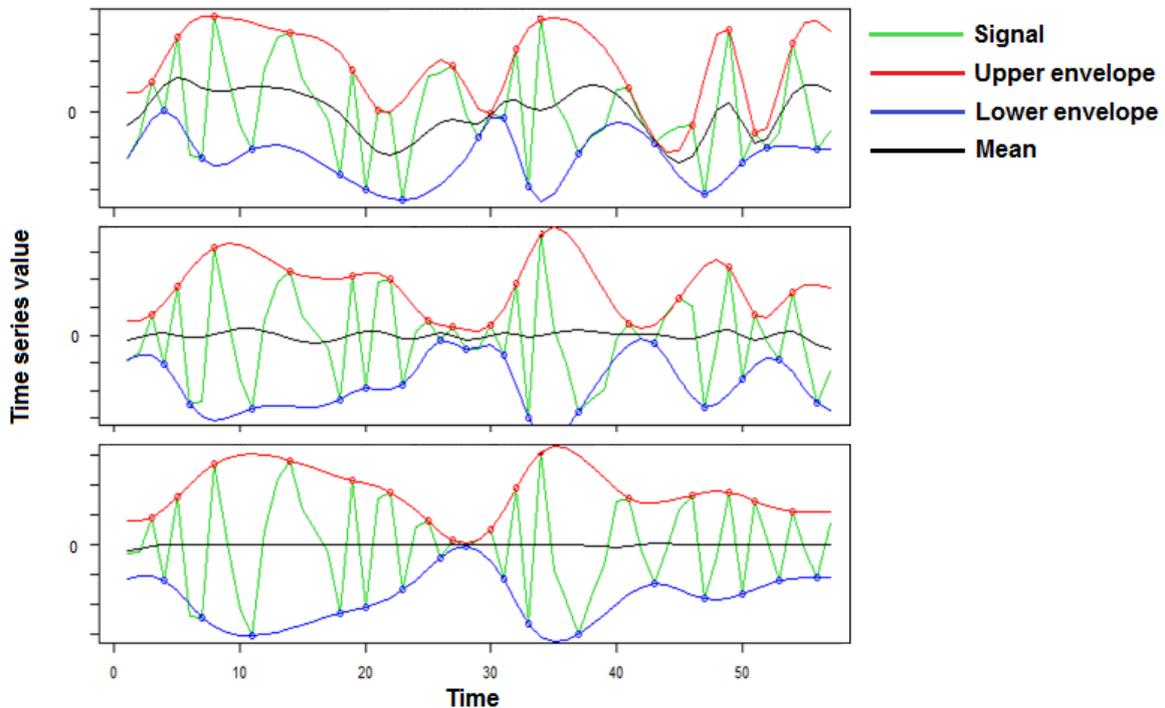
Figure 4.1: Schematic illustration of the sifting process for derivation of an intrinsic mode function.

4.1.1.4 Quantile Regression

The above methods perform a prognosis of changes in the mean of a time series without capturing further changes in its distribution. Quantile regression is an approach to address this problem, as it allows fitting of models to desired quantiles rather than the mean.

The concept of quantile regression, developed by Koenker and Bassett (1978), is closely related to ordinary least squares (OLS) regression, which is described in detail in section 4.1.2.1.

The basis of quantile regression poses the problem of defining quantiles of a sample, which is usually achieved via sorting. However, definition of a quantile may also be achieved via optimization. The median of a sample of size m, for example, can be found by minimizing a sum of absolute residuals

$$\min_{y_{med}} \sum_{i=1}^{m} (y_i - y_{med}), \tag{4.24}$$

analogous to least squares regression, where the sample mean is the solution of the minimization of squared residuals

$$\min_{\bar{y}} \sum_{i=1}^{m} (y_i - \bar{y})^2 \ . \tag{4.25}$$

For general quantiles $\alpha(\tau)$, the residuals to be minimized need to be adjusted. They need to be asymmetrically weighted by giving different weights to negative and positive values. This is achieved by introducing a loss function ρ_τ *into* the minimization problem

$$\min_{\alpha} \sum_{i=1}^{m} \rho_\tau (y_i - \alpha). \tag{4.26}$$

ρ_τ is defined as

$$\rho_\tau (u) = \begin{cases} v(\tau - 1) & for \quad v < 0 \\ u\tau & for \quad u > 0 \end{cases} \tag{4.27}$$

where u and v represent the positive and negative parts of the residuals.

A graphical example for some loss functions for different quantiles is depicted in Figure 4.2. The asymmetry becomes obvious for the 25- and 75-percent quantiles. The median, however yields a symmetric function, in accordance with equation (4.24).
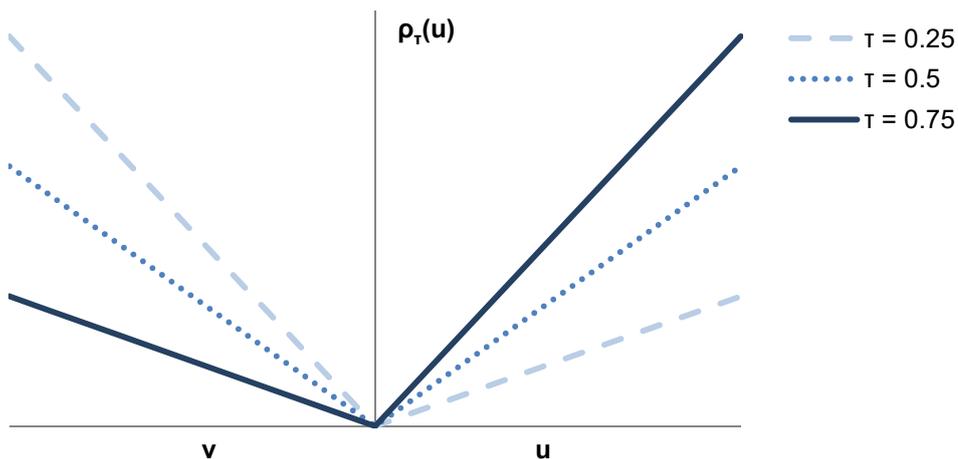
Figure 4.2: Loss functions for quantile regression for the three different quantiles.

In order to be able to estimate the quantiles from an external variable x, α can be expressed as βx^T, where β is the respective regression parameter

$$\min_{\beta} \sum_{i=1}^{m} \rho_\tau \left(y_i - \beta x_i^T \right).$$

(4.28)

Minimization then leads to estimation of β. The problem can be easily solved using linear programming. Therefore, the problem needs to be reformulated as

$$\min_{(\beta,u,v)} \left(\tau 1_m^T u + (1-\tau) 1_m^T v \mid \beta X + u - v = y \right),$$

(4.29)

where 1_m denotes a vector of 1 of length m.

A variety of algorithms exist to solve the optimization problem. The one applied here is a modified version of the Barrodale and Roberts algortithm. This algorithm will not be presented due to reasons of conciseness but can be found in Koenker and Dorey (1987) and Koenker and Dorey (1994).

Quantile regression is carried out using the R package "quantreg" by Koenker et al. (2016).

4.1.1.5 Non-stationary distribution functions

Another method tested to achieve prognosis of future low flow quantiles is the fitting of non-stationary distribution functions to the observed time series. The non-stationarity is thereby accounted for by assuming temporally variant distribution parameters. This approach has one major advantage over the quantile regression, as it allows modeling of the entire future distribution rather than single quantiles.

Due to its flexibility to fit extreme value data given its three parameters, the general extreme value (GEV) distribution has been selected for the analyses in this study. Its classical form is given by

$$F(y) = \exp\left(-\left(1 + \kappa \frac{y-\xi}{\alpha} \right)^{-\frac{1}{\kappa}} \right),$$

(4.30)

with $-\infty < \xi < \infty$, $\alpha > 0$, and $-\infty < \kappa < \infty$ as the location, scale and shape parameter. Estimation of the parameters can be achieved using maximum-likelihood estimation in case the time series y_t is independent and identically distributed. The likelihood function is given as

$$L_m(y \mid \xi, \alpha, \kappa) = \prod_{t=1}^{m} \frac{1}{\alpha} exp\left(-\left(1 - \kappa\left(\frac{y_t - \xi}{\alpha}\right)\right)^{\frac{1}{\kappa}}\right) * \left(1 - \kappa\left(\frac{y_t - \xi}{\alpha}\right)\right)^{-\left(1 - \frac{1}{\kappa}\right)},$$ (4.31)

where m is the number of observations. The corresponding log-likelihood is accordingly

$$l_m(y \mid \xi, \alpha, \kappa) = -m \ln(\alpha) - \sum_{t=1}^{m}\left(1 - \kappa\left(\frac{y_t - \xi}{\alpha}\right)\right)^{\frac{1}{\kappa}} - \sum_{t=1}^{n}\left(1 - \frac{1}{\kappa}\right)\ln\left(1 - \kappa\left(\frac{y_t - \xi}{\alpha}\right)\right)$$ (4.32)

for $\xi \neq 0$. Allowing the parameters to vary as a function of time extends the GEV model to

$$F(x) = exp\left(-\left(1 + \kappa(t)\frac{y - \xi(t)}{\alpha(t)}\right)^{-\frac{1}{\kappa(t)}}\right).$$ (4.33)

The time dependence of the parameters can be modeled in any desired way. In this case, they are chosen to be of simple linear, i.e.

$$\xi(t) = \beta_0 + \beta_1 t,$$ (4.34)

or quadratic form, i.e.

$$\xi(t) = \beta_0 + \beta_1 t + \beta_2 t^2.$$ (4.35)

The scale parameter α is thereby modeled using the exponential function, as in

$$\alpha(t) = exp(\beta_0 + \beta_1 t)^2$$ (4.36)

in order to not violate the positivity constraint during maximum likelihood estimation.

The time dependence is easily included into the optimization problem, increasing the number of equations to solve in proportion to the number of parameters included. The tested variants comprise the classical model, as well as linear and quadratic time dependence of all the parameters. This leads to minimization problems with 3 to 9 equations to solve. The algorithm selected for solution of the respective system is the one according to Nelder and Mead (1965).

How many or if any of the parameters are significantly non-stationary is tested via the likelihood ratio test. This parametric test is developed to test the superiority of nested models over their simpler forms, i.e. whether a model with a higher number of parameters performs significantly better than the same model with fewer parameters. Performance is thereby measured in terms of maximized log-likelihood function values. Assuming for example a model M_1 with $y_t \sim GEV(\beta_0, \beta_1, \alpha, \kappa)$ and its reduced form M_0 with $y_t \sim GEV(\xi, \alpha, \kappa)$, the models can be directly compared via

$$D = 2\big(l_1(M_1) - l_0(M_0)\big). \tag{4.37}$$

If the deviation is large enough for M_1 to perform significantly better than M_0, can be determined by comparison with a threshold c_α, which in this case is determined as the $(1 - \alpha)$ quantile of the X^2 distribution. The degrees of freedom of the distribution are chosen as the difference in number of parameters between M_0 and M_1, in this case 1. Where $D > c_\alpha$, M_1 is significantly better than M_0 on a significance level α (Coles, 2001).

4.1.2 Modeling low flow – climate relationship

The second set of methods tested for temporal low flow modeling base on input of climate data. The advantage of having external variables beyond time is the increase in accuracy compared to mere time series extrapolation. Aim is to directly model low flow values at desired resolution rather than just making rough predictions about changes in mean or quantiles.

4.1.2.1 Multiple linear Regression

Multiple linear regression is probably the most common method to model relationships between variables and will pose the basis for several methods in this work. It aims at reproducing a target variable y as linear combination of k explanatory variables $x_1, \ldots x_k$, termed regressors. The general shape of a multiple linear regression model is the following:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \tag{4.38}$$

where $\beta_0, \beta_1, \ldots \beta_k$ are regression parameters and ε is an additive error term. The error, the so-called residuum, arises for each observation as difference between measured and modeled value.

Model fitting occurs via estimation of the regression parameters. Given a set of k explanatory variables and m observations, the following system of equations can be obtained:

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\;\;\vdots \\
y_m &= \beta_0 + \beta_1 x_{m1} + \cdots + \beta_k x_{mk} + \varepsilon_m
\end{aligned}
\tag{4.39}
$$

in the following abbreviated as

$$y = \beta X + \varepsilon. \tag{4.40}$$

Ordinary least squares fitting

The parameters for the regression are sought to minimize the sum of squared residuals, i.e.

$$S(\beta) = \varepsilon^T \varepsilon = (y - \beta X)^T (y - \beta X) = y^T y - 2\beta^T X^T y + 2\beta^T X^T \beta X \rightarrow Min. \tag{4.41}$$

The vector of estimated regression coefficients can be obtained as

$$b = (X^T X)^{-1} X^T y \tag{4.42}$$

with a covariance matrix of

$$Cov(b) = \sigma^2 (X^T X)^{-1}. \tag{4.43}$$

The parameters in this work are estimated using maximum likelihood estimation. The likelihood function can be formulated as function of the sought parameters:

$$L(\beta, \sigma^2; y, X) = (2\pi\sigma^2)^{-\frac{m}{2}} exp\left(-\frac{(y - \beta X)^T (y - \beta X)}{2\sigma^2}\right). \tag{4.44}$$

The corresponding log-likelihood function is

$$l(\beta, \sigma^2; y, X) = -\frac{m}{2} ln(2\pi) - \frac{m}{2} ln(\sigma^2) - \frac{(y - \beta X)^T (y - \beta X)}{2\sigma^2}. \tag{4.45}$$

The estimator for the regression parameter vector b equates to the least squares estimator in equation (4.42). The variance can finally be estimated via

$$\hat{\sigma}^2 = \frac{(y - bX)^T (y - bX)}{m} \tag{4.46}$$

and the value of the log-likelihood function for the estimated parameters thus becomes

$$l(\beta, \sigma^2; y, X) = -\frac{m}{2} ln(2\pi) - \frac{m}{2} ln(\hat{\sigma}^2) - \frac{(y - bX)^T (y - bX)}{2\hat{\sigma}^2}. \tag{4.47}$$

The log-likelihood function value will be critical during model fitting for selection of suitable regressors, as explained in section 4.3.

The OLS-estimators for the regression coefficients are assumed to be best linear unbiased estimators (BLUE), when the following assumptions regarding the error terms are fulfilled: a) their expected value is equal to zero, b) they are uncorrelated, and c) they are homoscedastic. These assumptions need to be tested during model fitting. In order for parametric test to be valid and for calculation of standard errors of the model coefficients, the error terms additionally need to be d) normally distributed and e) independent.

Generalized least squares fitting

The above described multiple linear regression model is also referred to as ordinary least squares (OLS) model, which needs to fulfill all named assumptions to be effective. The fitting of a linear regression model, however, can be generalized, thereby adapting to constraints that cannot be met by the data at hand. The so-called method of generalized least squares (GLS) can address both dependence and heteroscedasticity of the residuals.

In GLS, the regression parameters are estimated as described in equation (4.42) under inclusion of the covariance matrix of the residuals:

$$b = \left(X^T \hat{\psi}^{-1} X \right)^{-1} X^T \hat{\psi}^{-1} y \,. \tag{4.48}$$

The covariance matrix ψ is denoted $\hat{\psi}$ because in practice it is unknown and must be estimated. In this study, especially potential serial correlation between the residuals will be considered, as heteroscedasticity did not prove a relevant issue for this model type. Therefore, autoregressive (AR) covariance structures of various orders are tested to describe the residual covariance. The elements of an AR(1) covariance structure, for example, can be described via

$$\sigma_{ij} = \sigma^2 \rho^{|i-j|} \,, \tag{4.49}$$

where ρ denotes the autocorrelation between terms with lag 1. The need for and order of the considered AR process is case wise determined using the likelihood ratio test described in section 4.1.1.5.

4.1.2.2 Principal component analysis

The fitting of multiple linear regression models is restricted by sample size. Fitting a model with a large number of regressors to a small data set will result in over-fitting, i.e. coefficients are estimated inefficiently to match the data the model has been calibrated on, whilst losing generality for prognosis outside of the calibration data set. Additionally, the problem of multicollinearity between the regressors in a model arises, as later on described in section 4.3. Consequently, for a small data set, only a small number of uncorrelated regressors should be selected to make up the ideal model. However, leaving out important explanatory variables may drastically lower the predictive power of the model. Simultaneously, variables that are to a certain extent correlated with regressors in an existing model may still contribute important information for prediction of the target variable. In order to overcome the restrictions given through the limited period of observation, a principal component analysis (PCA) is applied, merging many explanatory variables into a few uncorrelated components that pose the ideal basis for model fitting on limited data with maximum exploitation of information.

PCA is a method for reducing the dimensionality of a problem. A large set of variables, for example, may be linearly combined to form a smaller set of orthogonal, hence uncorrelated, components. In detail PCA is carried out as follows: in a first step, the variables of the matrix $X = (x_1, x_2, …, x_p)$ are centered by subtracting the respective means

$$x_i^* = x_i - \overline{x}_i. \tag{4.50}$$

Then, the ($p \times p$) covariance matrix Σ of X^* is computed and its eigenvalues $\lambda_1, \lambda_2, … \lambda_p$ and eigenvectors $\gamma_1, \gamma_2, …, \gamma_p$ are determined. The eigenvectors can be combined as the columns of a matrix Γ. Multiplication of X^* with Γ yields the following system of equations

$$
\begin{aligned}
y_1 &= \gamma_{11} x_1 + \gamma_{21} x_2 + … + \gamma_{p1} x_p \\
y_2 &= \gamma_{12} x_1 + \gamma_{22} x_2 + … + \gamma_{p2} x_p \\
&\quad\quad\quad\quad \vdots \\
y_p &= \gamma_{1p} x_1 + \gamma_{2p} x_2 + … + \gamma_{pp} x_p
\end{aligned}
\quad , \tag{4.51}
$$

where $y_1, y_2, …, y_p$ denote the principal components, sorted according to their contribution to the total variance of data at hand. The eigenvectors represent the respective loading of a variable x on a principal component y.

4.1.2.3 Support vector regression

Support vector regression, as described in section 4.1.1.2, is once more applied for index-based modeling, i.e. instead of time as the only covariate, external variables are included into the model. The aim is to address two flaws of the multiple linear regression approach, i.e. a) capture potential non-linearities in the low flow-climate relationship and b) improve generaltity in order to increase the prognostic capability.

4.1.2.4 Hydrological modeling

In order to evaluate the performance of the statistical approaches, hydrological modeling is applied as the benchmark for prognosis of future flow. The rainfall-runoff model used here is an adaptation of the Swedish HBV by Lindstrom et al. (1997), denoted HBV-IWW. The peculiarities of the model are described in detail e.g. in Wallner et al. (2013). A schematic overview over the structure of the model is presented in Figure 4.3. The model is semi distributed and can be applied on the sub-catchment scale. Input is daily precipitation, temperature and potential evapotranspiration. The model comprises of 5 routines and a variety of parameters controlling the translation between the various model components.

The parameters are automatically optimized using AMALGAM, an evolutionary multimethod algorithm. The sequence of this algorithm is as follows: at first, a random population is generated and each parameter set in this parent population is ranked. In the next step, so-called offspring

is created from the non-dominated members of the initial parent population through simultaneous application of various algorithms. The new population includes the non-dominated parent and all offspring elements and is ranked once more for creation of a new population until convergence is achieved.

Offspring creation via multiple algorithms refers to the simultaneous application and evaluation of a set of optimization strategies. In every step of producing a new offspring population, the success of each applied algorithm is assessed by accounting the number of offspring contributed by the respective algorithm. The algorithms are accordingly weighted, thereby reducing the contribution of weak algorithms that produce unsuccessful offspring and may cause overall premature convergence. A detailed description of the algorithm and the included optimization algorithms are given in Vrugt and Robinson (2007).
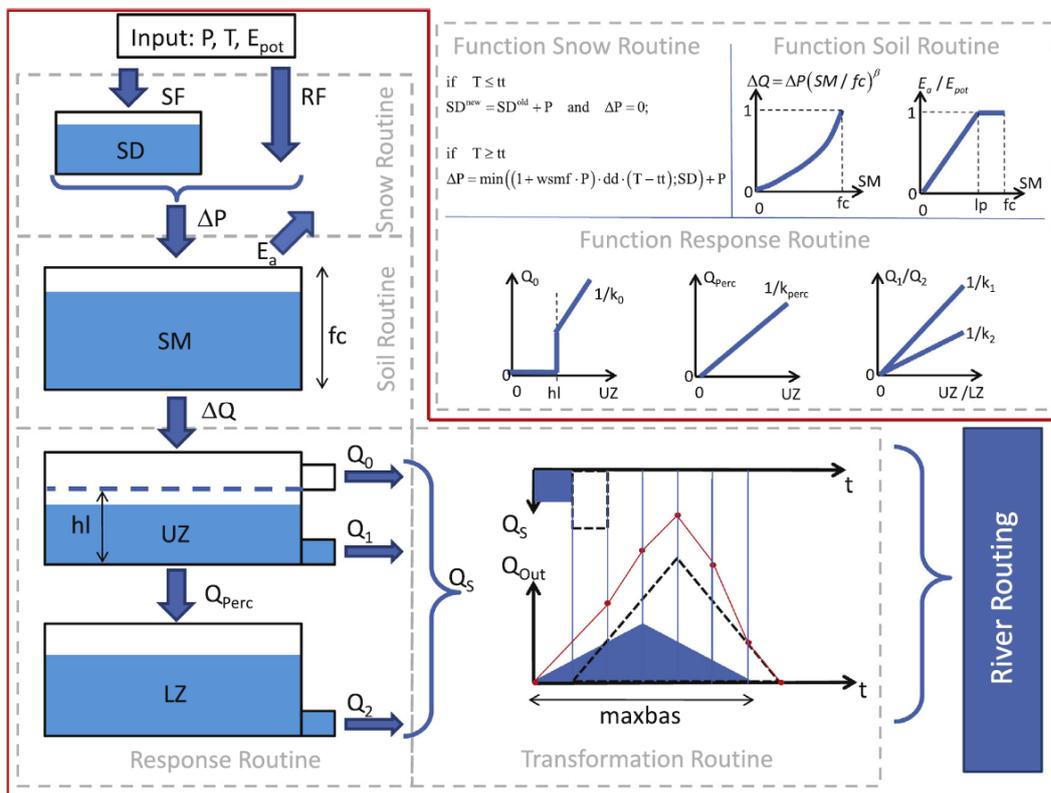
Figure 4.3: Vertical structure of the HBW-IWW model according to Wallner et al. (2013)

4.2 Data preparation

The data required for the temporal analyses are mere time series of low flow indices for time series extrapolation and low flow index time series and simultaneous time series of meteorological indices for the index-based modeling. For the evaluation of prognostic model

performance in time, the records of the selected gauges need to be as long as possible, in order to allow a proper separation into "past" calibration and "future" validation period. For time series extrapolation a minimum record length of 90 years has been set, 60 years for calibration, 30 for validation. A total number of 88 stations could be utilized for this purpose. For the index-based models, the gauge selection is further restricted by the limited availability of climate data. To ascertain a maximum record length for analysis, a mutual period from 1951 to 2010 has been chosen. This leaves 28 stations for analysis. The hydrological model could only be set up for 7 of these stations. Table 4.1 gives an overview over the number of available stations for the two temporal approaches, as well as for the other models applied in this work.

Table 4.1: Time periods and number of available stations for the applied methods.

Method	Time period	Criterion	Number of stations
Time series extrapolation	Full record with length > 90 years	Maximum length for assessment of temporal model quality	88
Meteorological index-based model	1951 – 2010	Maximum overlap with record of climate variables for assessment of temporal model quality	28
Regionalization	1988 – 2009 + artificial data set extension	Maximum number of stations for assessment of spatial model quality	137 + 84
Spatiotemporal models	1966 - 2005	Compromise between maximum of number stations and maximum overlap with climate record	51

The catchment sizes for time series extrapolation range from 74 to 576,232 km² for index-based modeling from 24 to 132,000 km².

4.3 Model fitting and evaluation of model performance

Data set separation

In order to allow for a proper assessment of the models' capabilities of predicting future low flows, the available record at each station is split into calibration and validation period. Both periods are continuous in time and the calibration consistently precedes the validation period. This set-up is chosen to test the ability of models fitted to a past period of time to predict "future", i.e. the validation period's characteristics.

For time series extrapolation the entire available record of annual low flow values is used for analysis. A period of 60 years is chosen for calibration for validation periods of 10, 20 and 30 years are selected. The respective periods are considered in a moving-window manner with an

offset of 10 years in between. Depending on the total time series length, a different number of possible model evaluations can be done for each station. For validation the mean or specific quantiles, respectively, are compared for all available 10-year blocks, in order to assess a potential decline of predictive power with increasing distance to the calibration period.

For the meteorological index-based models, the available time period of 1951 to 2010 is divided equally into 30 years of calibration and 30 years of validation. In order to double the available periods for validation, the time series are reversed and models are fitted to the former validation period and evaluated in the calibration period. The inversion is applied in order to still have a continuous series in time.

Fitting of time series extrapolation models

Fitting of any model is done explicitly on the calibration period only. The validation period is used merely for assessment of model quality and not for model selection. In order to prevent overfitting, however, the opportunity of assessment through an independent data set needs to be accounted for otherwise. Fitting of models for time series extrapolation is straightforward, as time is the only "external" variable to be considered. The only method that requires restriction to avoid overfitting is SVR, namely in setting the parameters ε and C. In Figure 4.4 the impact of the parameters ε and C are demonstrated for a two-dimensional problem. Panel 1 shows a fitted model with the parameters $\varepsilon = 0$ and $C = 1$. In panel 2 the ε parameter is increased to 1 which results in a greater flatness of the fitted function resulting from maximization of the margin within the ε-intensive zone. In this specific case, the increased flatness leads to an overestimation of the lower flow values by the model. In panel 3, C has been set to 100, which results in greater complexity of the fitted function, as prediction errors are weighted higher within the optimization process, regulating the function towards minimization of the amount of values outside of the ε-intensive zone. C can theoretically be increased to infinity, leading to a perfect fit of the training data.
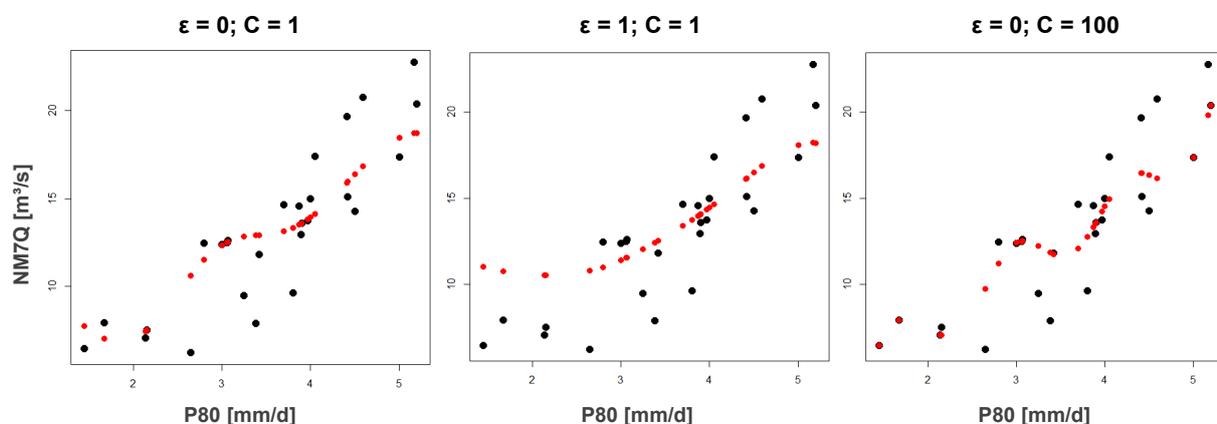
Figure 4.4: Effect of tuning the parameters ε and C in an SVR model.

In order to prevent an unrealistically perfect model fit to the training data and loss of generality for prediction of future values, tuning of ε and C is carried out using a 50:50 split validation on the training data, creating a large number of possible model outcomes. The parameter set showing best model performance in both subsets is selected for the final model. As relationships to important explanatory variables appear to be more or less linear, the range for tuning C has been set between 1 and 20 in order to speed up model fitting. Also, the kernel parameter γ needs to be evaluated during split validation. Its possible range has also been fixed to values between 0.1and 2.

For model fitting on EMD decomposed signals end-effects were accounted for through exclusion of the first and last value of the calibration period, as this appeared to have a major influence on the model quality. Models fitted to individual IMFs were integrated for validation by comparison to the original time series.

Selection of explanatory variables

Problematic is the variable selection for the index based multiple linear regression model, as the entire set of annual meteorological indicators with all lead times and base periods is used as input. In order to select appropriate indices and restrict the numbers of selected regressors, a two-way approach for variable selection is chosen that aims at minimizing the Bayesian information criterion (BIC) of the final model. The BIC of a fitted regression model, as suggested by Schwarz (1978) can be obtained as follows:

$$BIC = -2 * ln\widehat{L} + k * ln(m),$$

(4.52)

where $\widehat{L}$ is the maximum likelihood function value, obtained via maximum likelihood fitting of the proposed model (see equation (4.47)), m is the number of observations, and k is the number of estimated parameters, i.e. the regression coefficients and the intercept. The BIC is closely related to the Aikaike information criterion (AIC; Aikaike, 1974) but penalizes variable inclusion stricter for larger sample sizes, which is the reason for its selection in this study. The smaller the BIC, the better the predictive power of a model

The algorithm for variable selection starts by randomly selecting an initial variable, fitting a regression model via OLS and computing the BIC of the model. In a next step, all remaining variables are individually added to the model. The variable that leads to the strongest (if any) decrease in BIC is chosen as the second regressor. Before starting the process over and adding another variable to the model, the BIC is computed for the model without the initial variable. If the BIC is reduced, the variable is removed from the model. The algorithm is continued until no further BIC reduction can be obtained.

Since the meteorological indicators are in some cases very similar, due to only slight differences in calculation periods and/or lead times, a second criterion is included into the BIC-minimization

procedure in order to prevent multicollinearity. Multicollinearity arises when two or more regressors in a regression model are highly correlated. The major issue that results from multicollinearity is overfitting; an overfitted model lacks robustness, i.e. it is highly effective in estimating the target values of the data set it has been calibrated on but shows much lower predictive power for samples outside of this set. The stronger the relationship between target variable and predictors and the smaller among the predictors, the more robust is the model supposedly. In order to prevent multicollinearity the data set is first of all screened for perfectly correlated variables and duplicates are removed. Additionally, the variance inflation factor (VIF) of each variable added to the model is computed. The VIF is a measure of how much of the variance of a predictor in a model can be explained by the other predictors in the same model. Computation is done in the following way: a regression model is fitted to the variable X_{test} to be added to the existing model via OLS. Predictors are the n regressors of the existing model:

$$X_{test} = \alpha + \beta_1 * X_1 + \ldots + \beta_n * X_n + e.$$
(4.53)

The coefficient of determination is calculated for this model via

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$
(4.54)

where SS_{tot} denotes the total sum of squares of the data

$$SS_{tot} = \sum_i (X_{test\,i} - \bar{X}_{test})^2$$
(4.55)

and SS_{res} the sum of squares of the residuals

$$SS_{res} = \sum_i (X_{test\,i} - \hat{X}_{test\,i})^2 .$$
(4.56)

The VIF of X_{test} is then computed as:

$$VIF = \frac{1}{1 - R^2}.$$
(4.57)

The recommended limits for the VIF vary throughout literature. Since multicollinearity is a major issue for the analyses of this study, a low value of 5 has been chosen. Variables that show higher VIFs are not included into the existing model.

For the GLS models with AR-correlation structure, variable selection was carried out according to OLS model fitting, evaluating the necessity of the AR-structure for each added variable. The necessary order of the AR covariance structure is determined via pair wise comparison of higher versus lower order models using the likelihood ratio test introduced in 4.1.1.5. For variable

selection using principal components, the BIC-minimization strategy is extended, not adding each additional variable directly to the model but computing principal components first and testing those as potential regressors of the model. Variable selection for the SVR-model was not carried out. Rather was SVR applied to the variables selected for each model with the aim of improving the respective model through non-linearity and increased generality.

For the OLS, GLS and principal component model, a second variant with restricted variable selection is applied. This approach aims at filtering out variables that have a non-stationary relationship to the target variable. This filtering is achieved via bootstrapping. Randomly, 30 years are sampled with replacement from the calibration data set. A linear regression model is fitted between target variable and all potential explanatory variables. The random sampling disrupts the chronological order of the time series. If the regression is significant for all bootstrap samples, the relationship is considered continuously stationary, if, on the other hand, the slope of the regression line is time dependent, random samples taken from different periods will not yield a significant slope coefficient. Only the former variables are included in further variable selection. The models with restricted variable selection are denoted OLS-R, GLS-R and PC-R.

The hydrological model was calibrated in accordance with the statistical model approaches: only the annual low flow index values have to be met by the simulation, the fit of the rest of the daily hydrograph was of no interest. This approach appeared to be more effective for low flow simulation and was selected to not give an advantage to the statistical models.

Testing

Since independence of the residuals is a prerequisite for proper model fitting a Durbin-Watson test for autocorrelation of disturbances is carried out for each final model proposal (Durbin and Watson, 1950, 1951, 1971). The null hypothesis of the test is that the autocorrelation of the residuals of the fitted model is zero. Prerequisite for the test is a normal distribution of the residuals. The test statistic is calculated as follows:

$$d = \frac{\sum_{i=2}^{m}\left(\varepsilon_i - \varepsilon_{i-1}\right)^2}{\sum_{i=1}^{m}\varepsilon_i^{\,2}} \tag{4.58}$$

where ε denotes the residuals at time i. Positive autocorrelation causes the test statistic to approach 0, negative autocorrelation makes it tend to 4. No autocorrelation is indicated by a test statistic value of 2. Critical values may be obtained from the distribution of the test statistic.

Heteroscedasticity is tested with help of the Goldfeld-Quandt test (Goldfeld and Quandt, 1965). The test is based on simple comparison of variances between two subsamples of the residuals separated equally according to their respective observation values. The variances of the two samples are then compared with an F-test, i.e. a test statistic of

$$F = \frac{s_1^2}{s_2^2} \text{ with } s_1{}^2 > s_2{}^2 \qquad (4.59)$$

where

$$s_i{}^2 = \frac{1}{n_i - k} \sum_{j=1}^{n_i} \varepsilon_{ji} \cdot \qquad (4.60)$$

The test statistic is distributed as

$$F(n_1 - k, n_2 - k), \qquad (4.61)$$

where the degrees of freedom are dependent on the number of parameters k and the respective sample sizes n_1 and n_2.

Normality, as precondition for all other tests is tested using the Shapiro-Wilk test (Shapiro and Wilk, 1965). The test assumes a normal distribution of the tested data as its null hypothesis. The test statistic W is estimated via

$$W = \frac{b^2}{(n-1)s^2}, \qquad (4.62)$$

where b^2 describes the expected variance of the tested sample in case it came from a normal distribution. This value is estimated using expected normal order statistics. s^2 is the actual estimated sample variance.

All tests are carried out using the R package "lmtest" (Hothorn et al., 2011).

Goodness of fit

To compare observed with estimated means or quantiles, the error, which is simply the deviation of the estimated from the observed value, as well as the absolute error, i.e. the absolute value of this deviation, are calculated for each validation period. To make errors comparable between stations with different magnitudes of low flow, both values are given as percentages. The mean of the error over all considered periods corresponds to the mean percentage error (ME)

$$ME = \frac{x_{sim} - x_{obs}}{x_{obs}} * 100, \qquad (4.63)$$

the mean of the absolute errors to the mean absolute percent error (MAE)

$$MAE = \frac{\left| x_{sim} - x_{obs} \right|}{x_{obs}} * 100 \, . \tag{4.64}$$

Index based modeling allows prediction of annual low flow values and thus more detailed comparison between observed and simulated time series. In order to compare the overall fit of estimated low flow index values, a selection of goodness-of fit measures is computed. The first one is the normalized root mean squared error (NRMSE). The NRMSE is calculated as follows

$$NRMSE = \frac{\sqrt{\mu(sim - obs)^2}}{s(obs)} * 100 \, , \tag{4.65}$$

where *obs* and *sim* represent the observed and simulated series, μ and s denote the mean and the standard deviation. The NRMSE is given in %, the lower the error, the better. The second goodness-of-fit measure is the percent bias (pbias), defined as

$$pbias = \frac{\sum\limits_{i=1}^{N}(sim_i - obs_i)}{\sum\limits_{i=1}^{N} obs_i} * 100 \, . \tag{4.66}$$

A pbias of 0 is ideal, negative or positive values indicate the tendency of the model to under- , or respectively overestimate the observation. Another quality criteria is the Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970)

$$NSE = 1 - \frac{\sum\limits_{i=1}^{N}(obs_i - sim_i)^2}{\sum\limits_{i=1}^{N}(obs_i - \mu(obs))^2} \, , \tag{4.67}$$

whose values range between -∞ and 1. A value of 1 indicates a perfect fit, a value of 0 indicates that predictions are as accurate as the observed mean, and values < 0 indicate that the model is less accurate a predictor than the observed mean.

The final goodness-of-fit is the coefficient of determination, calculated as the squared Pearson correlation coefficient

$$r_{xy} = \frac{\sum\limits_{i=1}^{N}(obs_i - \mu(obs))(sim_i - \mu(sim))}{\sqrt{\sum\limits_{i=1}^{N}(obs_i - \mu(obs))^2} \cdot \sqrt{\sum\limits_{i=1}^{N}(sim_i - \mu(sim))^2}} \, . \tag{4.68}$$

The criteria are selected to show various aspects of model quality, i.e. information about the average fit, systematic errors, and similarity of the course of simulated and observed time series.

4.4 Results and discussion

4.4.1 Time series extrapolation

The graphics shown in the following for each method will compare predicted and observed means for the first, second and third 10-year blocks of a 30-year validation period. All prediction errors shown are based on models calibrated on 60 years of observation. The boxplots not only show the spread of the errors over stations but over all possible 90-year periods available in the data, which, due to differences in record length, can be several per station. For each period, the extrapolation error is compared with the error of assuming a stationary mean, i.e. the mean of the calibration period for validation. On top of the whiskers, the respective means of all deviations are shown.
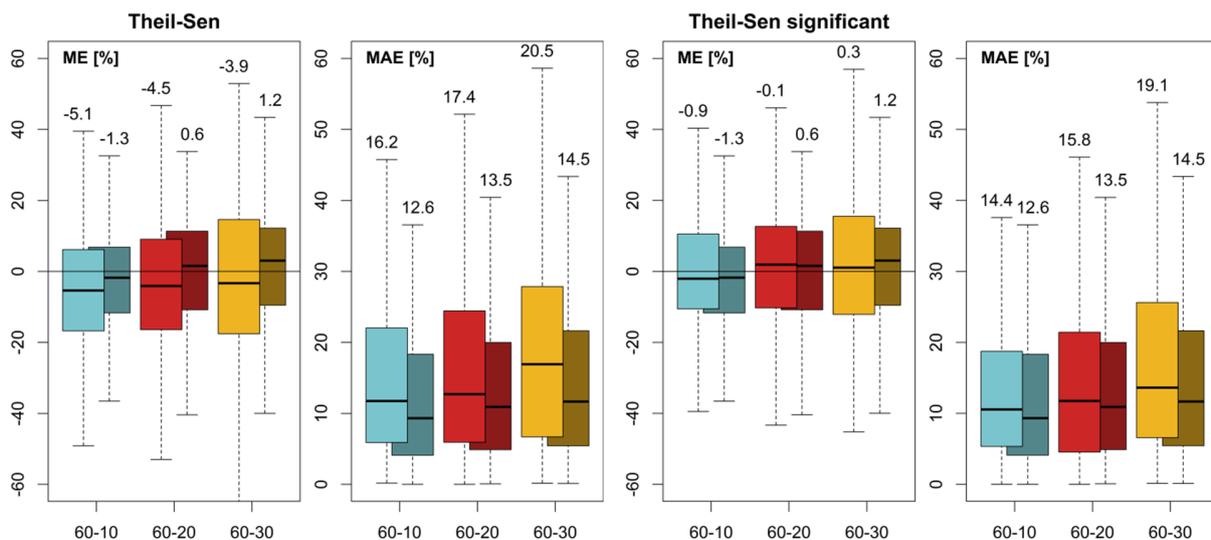
Figure 4.5: Mean and mean absolute deviation of Theil-Sen estimated means from the observation over all stations (light colors) in comparison to a stationary transfer of past means to the validation period (dark colors). The panels on the left show the extrapolation of all estimated trend slopes, the panels on the right show extrapolation of Mann-Kendall tested significant trend slopes.

Figure 4.5 shows the mean and mean absolute deviation of the estimated from the observed means for the Theil-Sen approach. On the left, the results for extrapolation of all fitted trend lines can be seen, while on the right, the results of extrapolating only significant (α = 5%) trend lines are depicted. The leftmost plot shows that extrapolation of all trends, significant or not, results in most of the cases in an underestimation of means in the validation periods. Correspondingly, a
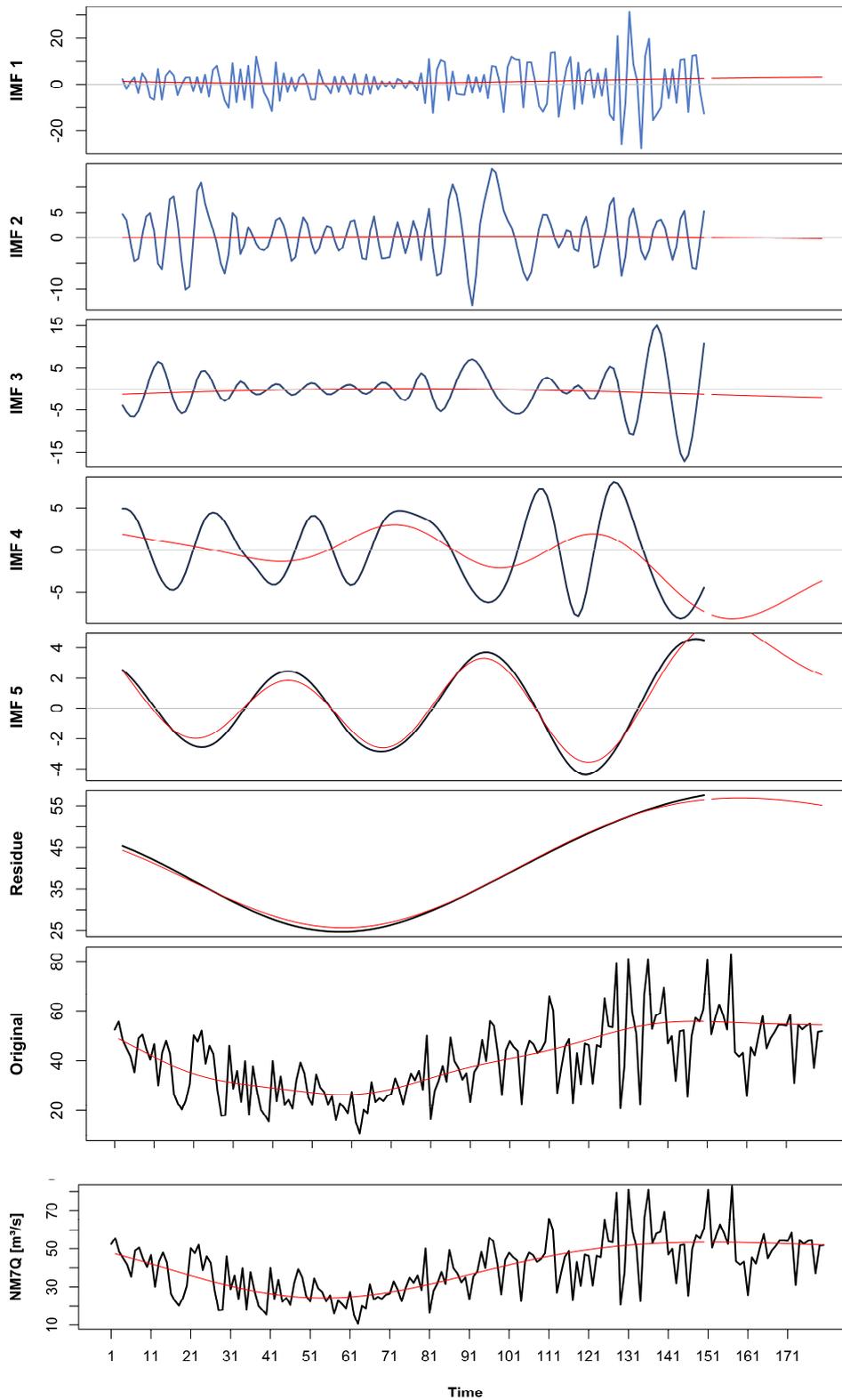
Figure 4.6: Decomposed annual NM7Q series (5 IMFs and residue) and SVR-models fitted to the individual components (red). The bottom two plots show the original signal with the integrated SVR-predicions and an SVR-model fitted directly to it.

significant amount of negative trends is detected during calibration that is falsely extrapolated to the future. The mean absolute error is accordingly much higher than that of assumed stationarity, up to an average 20.5 % for the final 10 years of the validation period. The exclusive extrapolation of significant trends shows much better results. The median ME is for all validation periods close to 0. The MAE, however, is still essentially higher than for the stationary case. In total for all 60 year calibration periods, 30.55% showed significant trends. Simple linear extrapolation of mean trends turns out to be a poorer choice when trying to assess future developments than simply assuming stationarity. The variability in the time series is too high and changes in the mean too low to be modeled by a linear function.
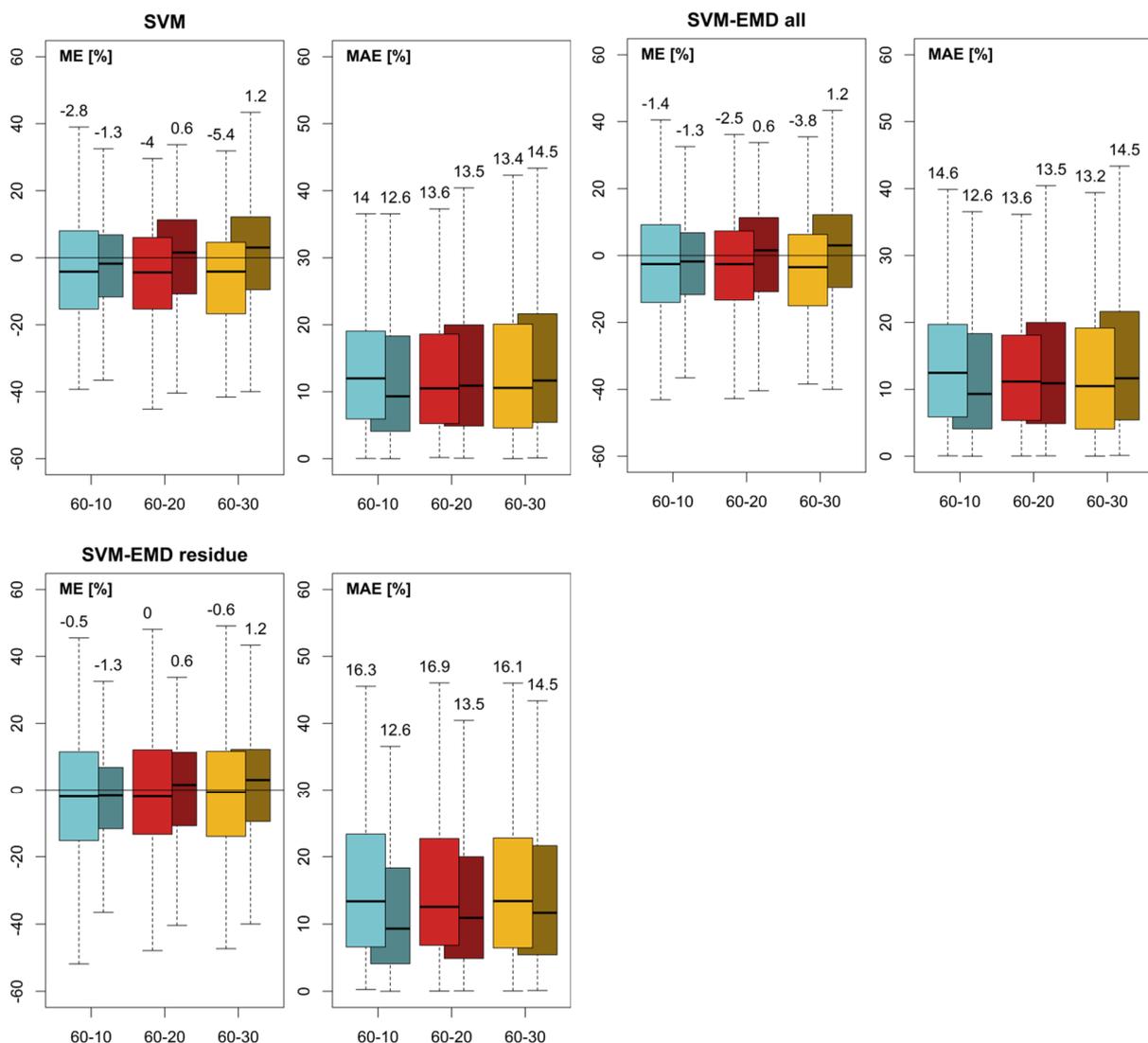
Figure 4.7: Mean and mean absolute deviation of SVR estimated means from the observation over all stations (light colors) in comparison to stationary transfer of past means to the validation period (dark colors). The first panels show the models fitted to the original time series, the second panels show the models fitted to individual IMFs and the final panels show the models fitted to the residue only.

For improvement of the prognostic power of time series extrapolation models, non-linear approaches eventually need to be considered. This is done here by applying SVR. Decomposition of the time series into a set of components with different frequencies is supposed to help the regression model by removing high-frequency noise components and allowing a separate fitting to each component individually, which is assumed to be more straightforward than for a complex signal.

Figure 4.6 shows an example of a decomposed series and the SVR models fitted to each component in red. The first IMFs show a high frequency that cannot be captured by the SVR model. The fifth IMF and the residue, on the other hand, appear to be reproduced quite well. Due to the inability to model the first IMFs, the integrated simulated signal, shown in red over the original signal, is not reproducing the annual variability. IMF 4 makes up a minor portion of the total signal so that its course is no longer visible in the integrated signal. For comparison, an SVR model is shown that was predicted directly on the original signal. The difference to the integrated signal is minor, indicating that decomposition via EMD may not contribute to a better model fit.
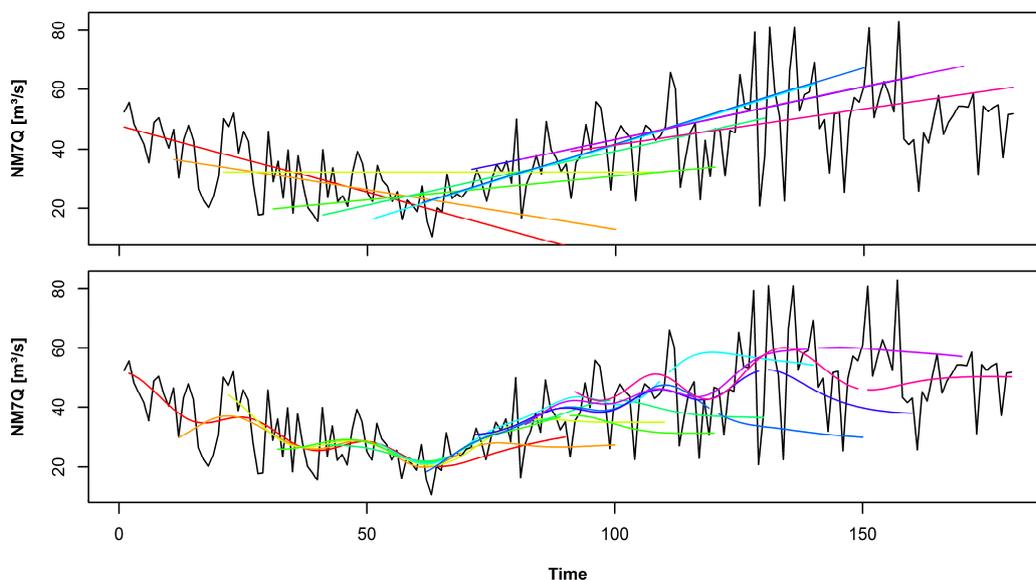
Figure 4.8: Fitted models and extrapolation of the significant Theil-Sen approach (top) and SVR fitted to individual IMFs (bottom). The differently colored lines represent different periods used for model fitting with 60 years of calibration and 30 years of validation.

Figure 4.7 shows the validation results for the EMD-SVR combinations. In the first case the original series is used for model fitting. The ME is on average below zero, indicating again an underestimation of future values. The MAE is higher than the stationary case for the first 10 years of the validation period, but slightly lower for the remaining periods. The results for the individually modelled time series components (SVM-EMD-all) are, as seen before, comparable

to modelling of the entire series. Underestimation is slightly reduced but overall MAE shows comparable quality. SVR modeling of only the residue of the decomposed series, as shown in the lower two panels of Figure 4.7, exhibits smallest mean bias but MAE is significantly higher. Inclusion of higher-frequency components into modelling thus appears to be meaniful.

The advantage of using non-linear methods for time series extrapolation is once more depicted in Figure 4.8. The top graph shows an example time series of the NM7Q and trend lines fitted according to Theil-Sen to 60 years of observation and extrapolated 30 years ahead in time. It becomes obvious that the linear approach is not suitable in case non-unidirectional trends are present in the series. Unforeseen breaks and change in direction make it impossible to predict future values via linear extrapolation. Non-linear approaches, on the other hand, as exemplified by SVR models in the lower panel, show much greater flexibility to adapt to breaks and directional changes. Thus, the dependence on length and position of the calibration period is reduced. Additionally, extrapolation is "smoother" than for linear trends, i.e. predictions are more damped and thus more realistic.
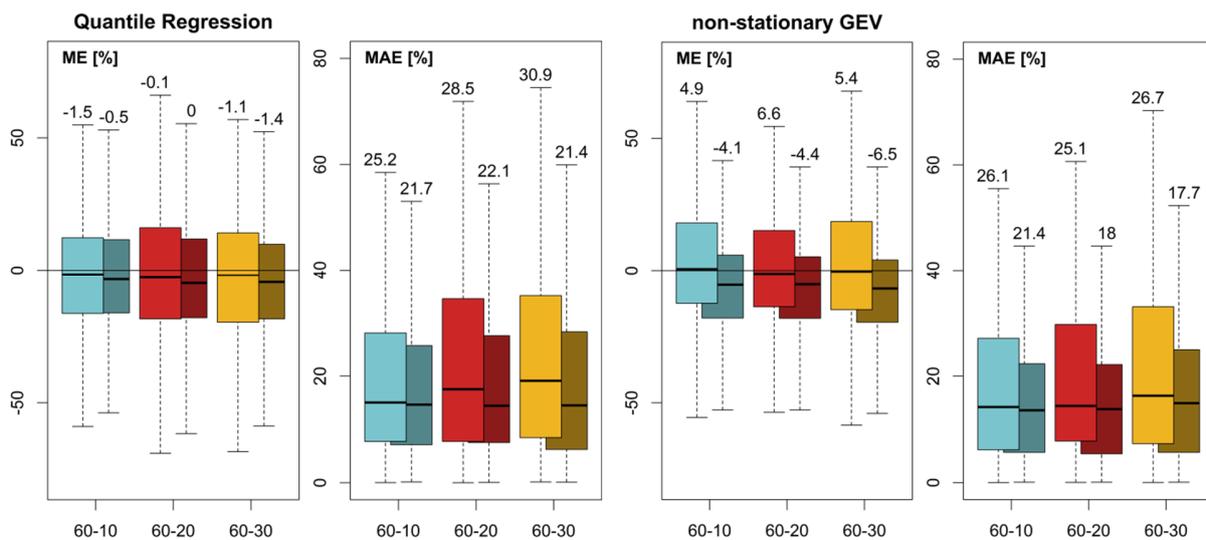
Figure 4.9: Mean and mean absolute deviation of the 10-percent quantiles of the NM7Q estimated via quantile regression (left) and non-stationary GEV functions (right) from the observation over all stations (light colors) in comparison to a stationary transfer of past quantiles to the validation period (dark colors).

The extrapolation of quantiles shows similar results as for the means. Figure 4.9 depicts the results of extrapolation of the 10-percent quantile of the NM7Q via quantile regression and non-stationary GEVs. The latter approach yields better predictions than the former but still does not prove better than mere stationary transfer of past quantiles to the future. The same picture is obtained for other quantiles. Again, these results show that non-linear approaches outperform linear ones for the purpose of extrapolating time series.

For all tested methods, the predictive power of the models did not improve with increasing calibration length. For the linear approaches, predictive power appears to drastically decrease with distance to the calibration period. Non-linear methods appear to have similar performance for all validation periods. The results apply similarly to the other low flow indices, i.e. NM30Q, Q_{95} and Q_{80} and low flow timing. The results for SVR and non-stationary GEVs for the individual indices can be found in Appendix B1 and B2. Fitting of non-stationary GEVs was not meaningful for the duration and deficit volume indices, as well as for timing, and is therefore not shown.

Based on these results it has to be concluded that simple time series extrapolation cannot be applied to even roughly estimate future developments and expected low flow magnitudes. The methods used here aimed at recognizing patterns and underlying processes of the time series which they should extend to the future. It was also tried to make use of potential autoregressive processes resulting from storage related internal factors, like vector autoregression for simultaneous extrapolation of different IMFs, as well as of several variables, hidden Markov models and ARIMA approaches. These models however showed even lower prognostic power and higher uncertainty than the approaches elaborated on.

Based on the findings, inclusion of the linear time series extrapolations into ensembles for climate change impact assessment is discouraged. Direction and magnitude of their projected changes depend too much on the calibration period. Non-linear approaches like SVM and non-stationary GEVs may be a good indicator of the future development and can pose a basis for comparison with other prognostic methods. It should be noted that the results do by no means indicate that the time series are stationary, they simply show that the error of the applied methods is too high and overshoots the changes observed in the time series.

4.4.2   Index-based models

Models including climate data as external explanatory variables show a much better goodness of fit than the models based exclusively on the time series itself. Figure 4.10 (left) shows the comparison of estimated vs. observed means in comparison to the stationary case, as in the previous section, for the non-restricted principal component model. The two pictures on the right show the direct comparison to the Theil-Sen approach with extrapolation of significant trends. The model including external variables appears to be a much better estimator for the mean than mere linear extrapolation and, as not shown here, any other extrapolation method tested, especially if the "future" is highly non-stationary, as seen in the 30-20 scenario.

The models that base on external meteorological variables are in fact not only capable of predicting changes in means but to reproduce the entire course of the time series. The advantage becomes obvious when predicting quantiles of the time series, where the stationary transfer of past quantiles into the future results in significantly higher errors than estimation via multiple linear regression models. An example is shown for the principle component model in Figure 4.11 for three quantiles with 10, 20 and 50-year return periods. The model is more

accurate in estimating all quantiles in the validation period than the stationary approach. Nevertheless, a deterioration of model performance with increasing return period is visible, indicating problems in modeling the lower extremes.
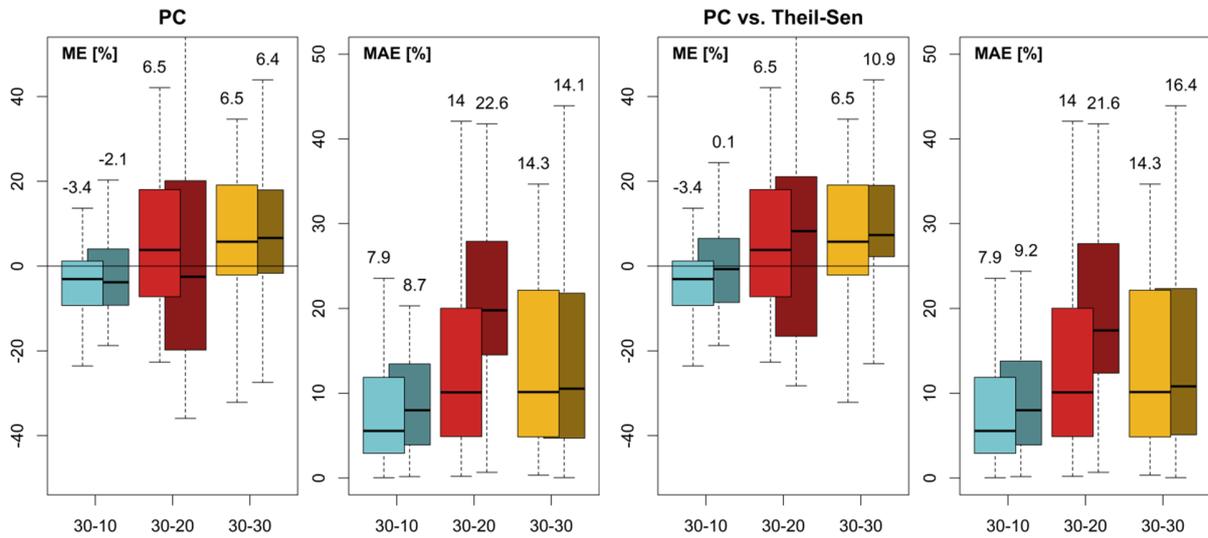
Figure 4.10: Mean and mean absolute deviation of the mean NM7Q estimated via principal component regression from the observation over all stations (light colors) in comparison to a stationary transfer of past means to the validation period (left) and Theil-Sen estimated means (right; dark colors).
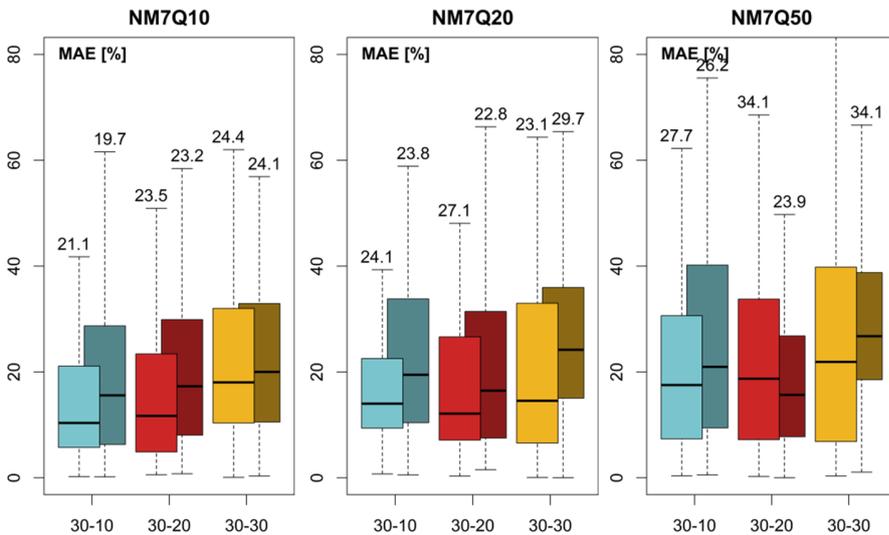
Figure 4.11: Mean and mean absolute deviation of different quantiles of the NM7Q estimated via principal component regression from the observation over all stations (light colors) in comparison to a stationary transfer of past quantiles to the validation period (dark colors).

Due to the high accuracy of the estimates, performance of the individual MLR model variants will be tested using additional goodness-of-fit measures. Figure 4.12 shows the model performance of all tested model configurations for NM7Q prediction in the calibration (top) and validation period (bottom). The left bars show the performance of the MLR models, the right bars the performance of the SVR models fitted using the same explanatory variables. For the validation period one can clearly observe an increase in model performance from left to right, i.e. an increasing predictive power with increasing model complexity. The non-restricted OLS-fitted model shows poorest performance with median NSE and R² of only 0.34 and 0.42, respectively. The removal of variables with a potential non-stationary relationship to the target variable increases model performance significantly, as seen for the OLS-R model. NSE and R² increase in median to 0.50 and 0.55 and NRMSE and pbias drop from 80.5 to 70.8 % and 2.0 to 0.0 %.
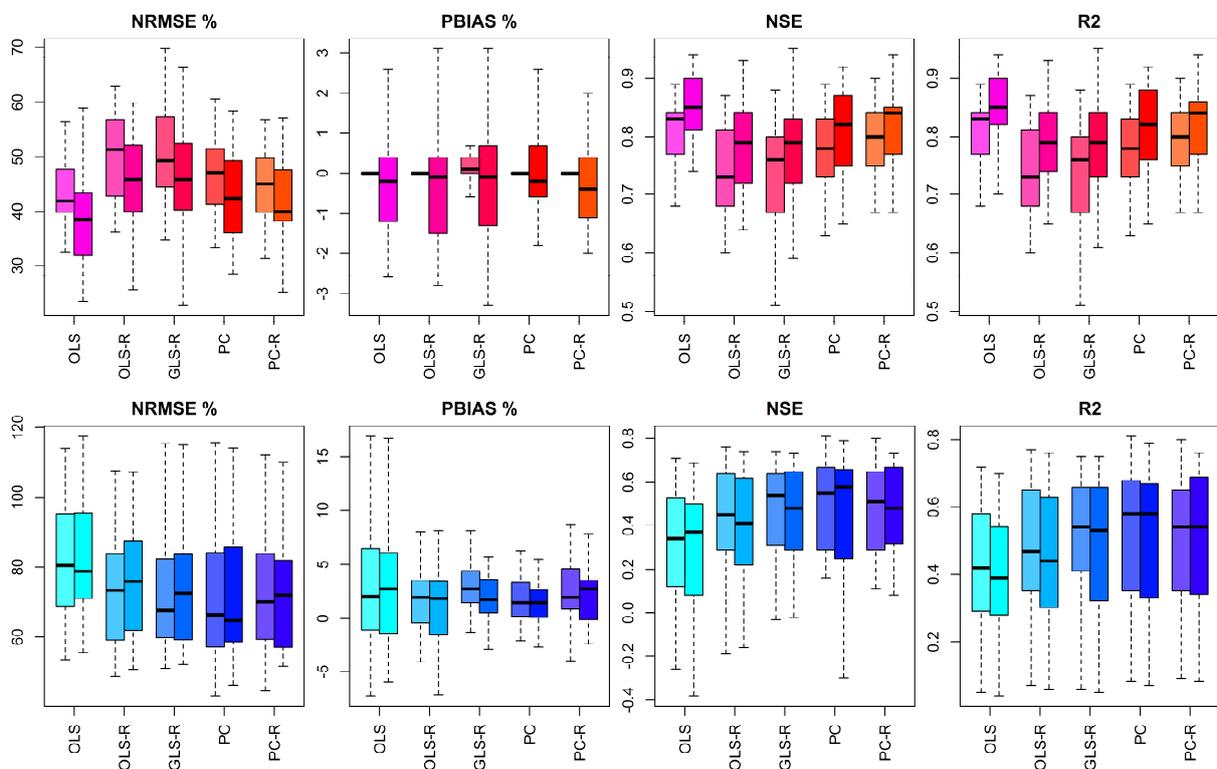
Figure 4.12: Criteria for goodness of fit of different model approaches in the calibration period (top) and validation period (bottom) over 28 stations. The left bars show MLR, the right bars SVR model performance.

Highest goodness of fit has been achieved with the principal component model without restriction during variable selection (PC). The NSE for the MLR model over all stations has a median value of 0.55, the coefficient of determination one of 0.58. The PC-model also shows smallest percent bias with a median of 1.4 % and smallest NRMSE with 66.4 %. The PC-R

model, where potential non-stationary variables have been removed, showed slightly lower performance.

Second best performance shows the GLS-fitted model with an AR-correlation structure and restriction during variable selection via exclusion of variables with non-stationary relationship to the target variable (GLS-R). The SVR models do not show a constant improvement of the OLS-fitted models and the results are overall comparable. It is thus assumed that SVR-fitting is not beneficial in terms of generality or slight non-linearity.

Exemplary, the individual models for all stations are listed in Appendix B3 for NM7Q estimation via GLS-R. When looking at the regressors that have been selected for the individual models it can be seen that many of them repeat for most of the stations. The most common predictor of the NM7Q appears to be the SPEI with different base periods and lead times. Most common are the 3-months minimum SPEI and the mean 6-months SPEI with 0 to 2 months lead. Average precipitation and upper precipitation quantiles follow in frequency. Some models also contain the ratio of average DSD to WSD.
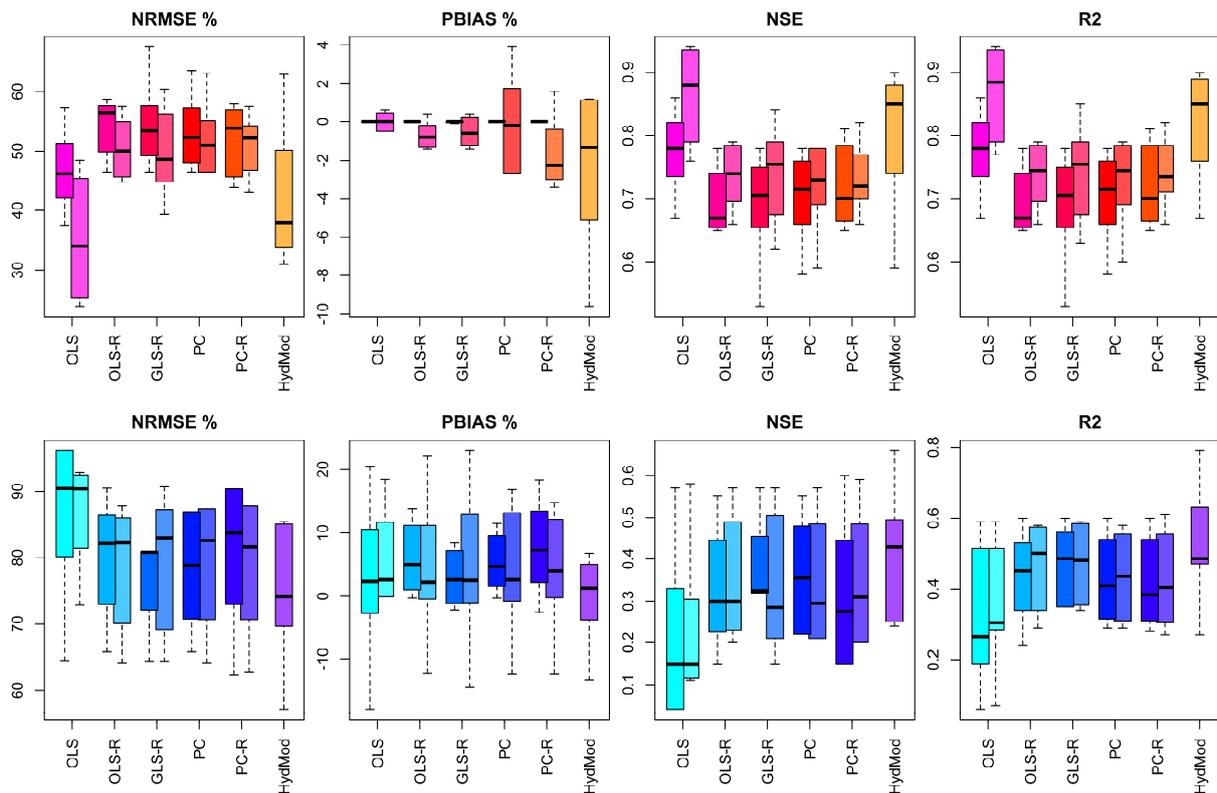
Figure 4.13: Criteria for goodness of fit of different model approaches in the calibration period (top) and validation period (bottom) over 7 stations. The left bars show MLR, the right bars SVR model performance.

Figure 4.13 shows the same criteria as Figure 4.12 with inclusion of the hydrological model, and thus for a total of only 7 stations. The performance of the HBV model in both calibration and validation period is substantially higher in all aspects with median validation NRMSE of 74.2 %, a percent bias of 1.3 %, an NSE of 0.43 and an R² of 0.49, compared to the PC-model with median values of 78.8 %, 4.6 %, 0.36 and 0.41, respectively. This performance, however, could only be achieved using the calibration strategy of matching exclusively the annual NM7Q values at each station. Calibration using the entire hydrograph or parts of the flow duration curve did not explicitly outperform the statistical approaches. Nevertheless, HBV-IWW appears to be suitable for assessment of future low flow development when calibrated specifically on the desired low flow index.

The statistical models appear to be positively biased. At almost all investigated stations a positive mean error could be observed. As models are fitted in both directions to the time series, this error appears to originate in the model itself, rather than in underlying processes of the time series that cannot be captured.

When testing the residuals of the fitted models, the preconditions seem to be satisfied by almost all model types. Table 4.2 summarizes the number of successful tests (α = 5 %) for non-normality, dependence and heteroscedasticity applied to the residuals of the individual model types, as well as to the original time series. The numbers refer to the total set of 28 stations.

The stations that show non-compliance with the preconditions of OLS-fitting are similar for all criteria and methods. These are primarily two gauges situated on the Vechte in the very West of Lower Saxony and three gauges on the Oker in the East. No method appears to be superior in terms of error distribution. GLS-fitting using an AR process as correlation structure appears to add significantly to the dependence of the residuals. This effect originates in the selection process of the correlation structure, which is done explicitly on basis of maximized log-likelihood.

Table 4.2: Number of successful tests out of 28 stations for different criteria applied to the original time series and several regression approaches.

	Time series	OLS	OLS-R	GLS	PC	PC-R
Non-normality	9	4	2	2	1	2
Dependence	2	0	2	7	4	3
Heteroscedasticity	3	3	4	5	5	5

Apart from the questionable performance of the above mentioned stations, differences in model performance could not be correlated to any specific regions or catchment features. The quality of the simulation did not depend on catchment size, position in the study area or any other distinguishable factor. A major challenge to the approach, however, poses anthropogenic interference. Even if indirectly considered in the statistical models, changes in the management

patterns would significantly alter the prognosis; a factor that needs to be considered when applying the models.

Table 4.3: Average absolute difference in quality criteria between calibration and validation period for 28 stations.

	OLS	OLS-R	GLS-R	PC	PC-R
NRMSE	32.7 %	20.5 %	20.0 %	21.8 %	25.0 %
PBIAS	3.7 %	2.9 %	2.4 %	2.1 %	2.4 %
NSE	0.37	0.25	0.24	0.23	0.26
R²	0.32	0.22	0.24	0.20	0.26

An obvious problematic is the significant difference in model performance between calibration and validation period. All quality criteria certify a much better performance during calibration than during validation. Table 4.3 and Table 4.4 show the median differences in goodness-of-fit measures for all model variants over all stations and over the stations at which hydrological modeling was possible. The differences are severe, even though quite a number of precautions have been taken during calibration. The NSE for unrestricted PC, for example, differs by 0.23 between calibration and validation. For the 7 stations this difference increases to 0.28. Overfitting appears to be an issue, even though the number of regressors has been restricted. Despite of choosing only one or two external variables for each model, the issue of lack of generality of the selected models prevails. Since only lumped meteorological statistics are used as input for the regression models, leaving out a range of potential information, a perfect fit to the observed low flows is not expected. However, automatic model fitting often leads to an unexpectedly good fit in the calibration period. The differences are also observable for the hydrological model, in some cases even more drastic than for the statistical models. The median difference in NSE over the 7 stations is 0.42 between calibration and validation period.

Table 4.4: Average absolute difference in quality criteria between calibration and validation period for 7 stations.

	OLS	OLS-R	GLS-R	PC	PC-R	HydMod
NRMSE	38.7 %	28.5 %	23.3 %	20.4 %	30.9 %	35.6 %
PBIAS	6.6 %	8.1 %	7.1 %	6.7 %	8.1 %	3.0 %
NSE	0.58	0.40	0.33	0.28	0.46	0.42
R²	0.45	0.23	0.19	0.23	0.33	0.31

Another possible reason for these discrepancies could be a non-stationary behavior in the low flow – meteorological index relationship, which could not be captured during calibration and which may cause the significant deviations between the former and the later period. Fitting of

models to moving windows along the time series shows that the selection of regressors varies from part to part of the series. One obvious reason for this behavior is the great similarity of the explanatory variables, resulting in each segment of the time series favoring another variable that is slightly different from the previous one. Nonetheless, this may indicate that the low flow – meteorological index relationship is not stationary or linear after all.

By separating the validation period into three parts it becomes obvious that performance decreases steadily with distance from the validation period, as shown in Figure 4.14. Compared are the estimated and the observed means for all methods. The error increases also for methods that have been supposedly restricted to stationary relationships between target variable and regressors. The effect is not observable for the hydrological model, which emphasizes once more its superiority over the statistical approaches and shows that non-stationary relationships between low flow and meteorological indices are of major relevance and need to be considered in modeling. Maximum likelihood fitting of models with linear time dependence of the coefficients (validated via likelihood ratio tests) was tried during this study to encompass the problem, but could not yield the expected results for the small calibration period.
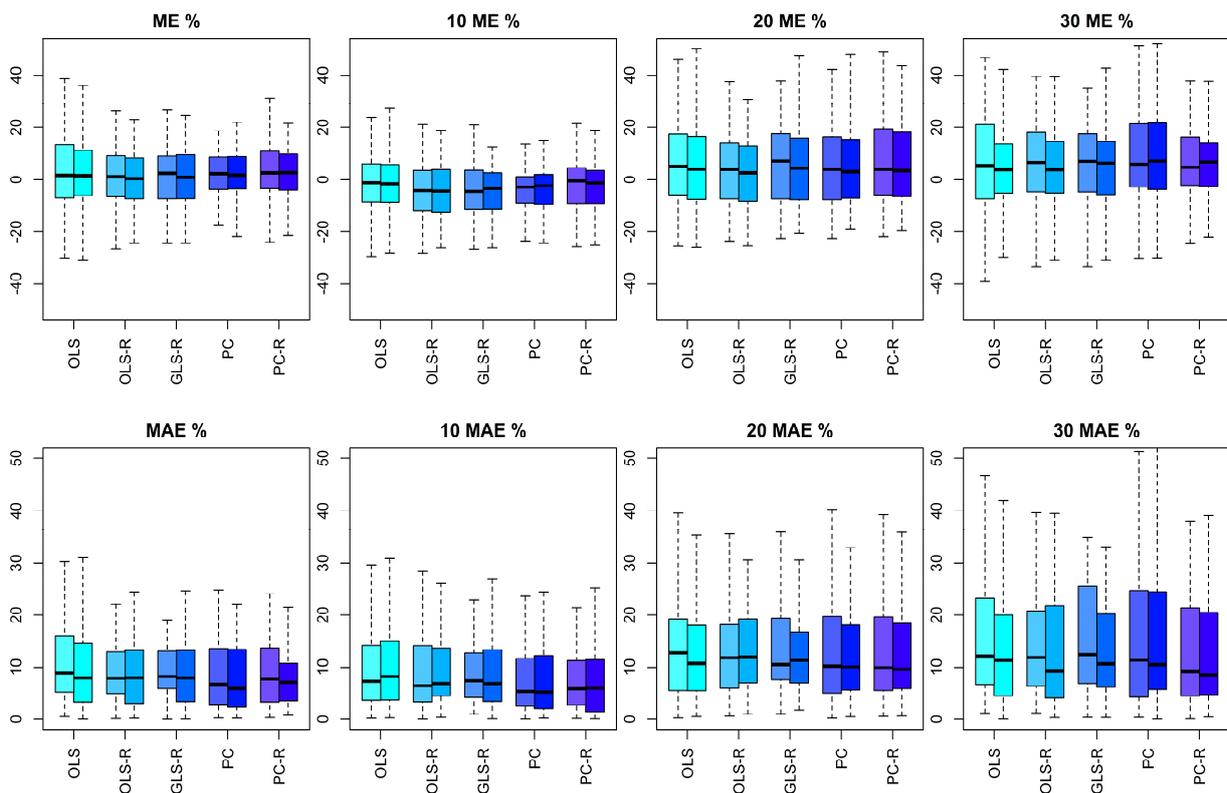
Figure 4.14: Mean (top) and mean absolute (bottom) deviation of estimated from observed means for the whole validation period (left), as well as the first, second and final 10 years of validation. The left bars show MLR, the right bars SVR model performance.

Modeling of the other tested indices showed quite a differentiated picture. As shown exemplary for the restricted GLS approach in Figure 4.15, some indices were reproduced more (top), others less successfully (bottom) than the NM7Q. Estimation of the Q_{95} appeared slightly better in terms of all quality criteria (median NSE of 0.54 compared to 0.50). The D_{max} was modeled effectively via GLS in terms of NRMSE, NSE and R² (median values of 47.9 %, 0.57 and 0.59) but showed quite a significant negative bias (median -2.1 %). V_{max} and especially the annual low flow timing could not be modeled successfully by any of the statistical approaches (median NSE of 0.22 and -0.5, respectively).

It was noted that the overall model performance was slightly higher for the more average values, i.e. NM30Q and Q_{80} than for the more extreme indices NM7Q and Q_{95}. The values for the larger quantile exceed the lower ones on average, stations with extremely poor fit are no longer present. The median NSE values of the NM7Q and the NM30Q compare as 0.5 and 0.58, the ones of the Q_{95} and Q_{80} as 0.54 and 0.62. Compared to the D_{max} and V_{max}, the annual average D_{mean} and V_{mean} cannot be predicted by the fitted models. The validation yields negative NSEs for both cases.
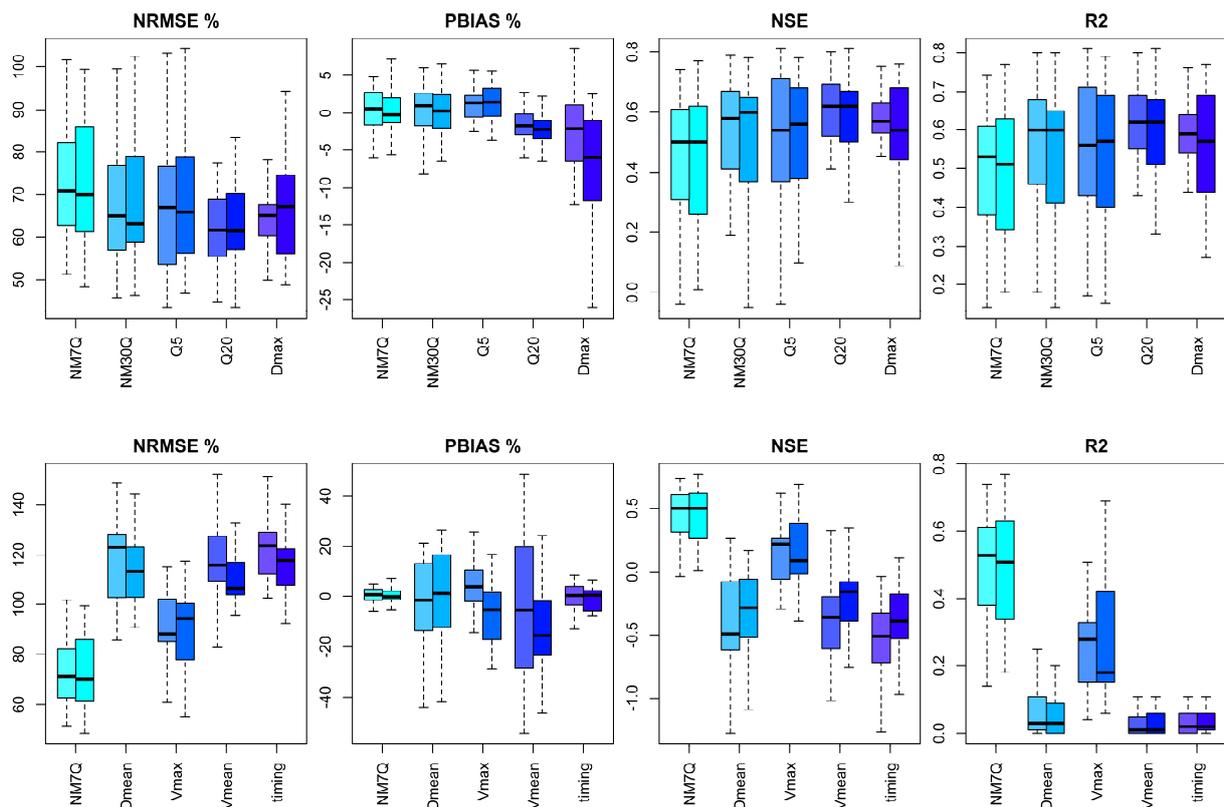
Figure 4.15: Validation results for the restricted GLS model for various low flow indices. The left bars show MLR, the right bars SVR model performance.

The problems that arise for modeling of deficit volume, timing and average event duration lie within the nature of these indices and the structure of the statistical model. The regression model uses aggregated meteorological features over a specific period of time for prediction of the desired low flow variable. The shorter the required period, the lower the degree of averaging and the greater the possibility of capturing extremes that cause a subsequent low flow event. Therefore, indices related to a specific event, like the NM7Q or the D_{max} can be modeled quite effectively based on previous meteorological states. The V_{max}, however, though equally related to a specific event is barely reproduced by the models at hand. This may be due to the dimensionality and definition of this very index. The V_{max} depends on a defined threshold and is characterized by duration of the event and its magnitude, thus poses some kind of combination of D_{max} and NM7Q. Even though both indices can be successfully estimated individually, prediction in combination appears not feasible. Based on the findings of Van Loon and Laaha (2015) one would expect the deficit volume to be better predicted by climatic indices than duration.

The fact that more average indices like NM30Q and Q_{80} are reproduced better than more extreme ones can be explained by the same principle. Meteorological indices computed for longer base periods are required to model more average index values. Errors that occur if extremes cannot be explained by external variables are lower. Dmean and Vmean, however, do not represent overall annual average values but averages of several events. These features cannot be captured by small sets of regressors as used in this study. The hydrological model has not been applied to other indices than the NM7Q but is expected to yield significantly better results, especially for deficit volume and low flow timing.

Considering the models of the other indices, it can be seen that the selected regressors differ between the different types. While Q95 and Q80 are primarily predicted by water balance parameters like the NM7Q and NM30Q, the models for Dmax, Vmax and timing count a high number of indices related to durations.

Despite all challenges, the approach of modeling specific low flow indices as a function of meteorological indices appears promising. Model set up and computation are straightforward and quick, even over large study areas with high numbers of catchments. The method appears to be able to simulate flow for any catchment from small to large and flat to steep without any consideration of physiographical characteristics. Being able to consider larger periods for calibration would assumingly lead to an increase in robustness of the models and extend the potential horizon for prognosis. Regarding the latter, it is difficult to say how far into the future prognoses can be trustable based on the limited time series lengths available for this study, but "near future" scenarios will definitely be easier to predict than periods more than 50 years after the calibration period.

The effect of non-stationarity adds to the error that will arise if the climate models that deliver the input for the statistic models are not capable of reproducing the required meteorological indices

adequately. Application of the regression models thus requires prior examination of the climate model data. However, the input variables represent indices that are lumped over a significant amount of time (3 – 12 months), averages that can potentially be better reproduced than the daily variability required for other impact model types.

All in all the statistical approaches can be of major assistance for decision making. Even though the prognosticated values should not be taken as basis for dimensioning they can offer a good approximation of the future development, especially when considered in a framework of model ensembles.

5   Spatial modeling of low flows

Chapter 4 has shown that temporal modeling of low flows appears feasible at gauges with sufficiently long records for proper model fitting. These gauges, however, represent a minor fraction of the river network. Prognoses for catchments with short or completely without data thus require regionalization of the available observations. In this chapter, a selection of methods will be presented and evaluated that are potentially suitable for regionalization of low flows. It should be noted that this chapter focuses on the regionalization of single point values in space, rather than the entire time series and hence allows a differentiated view on the spatial variability of low flows.

5.1   Methods

5.1.1   Multiple linear regression

Multiple linear regression, as described in section 4.1.2.1 is applied for regionalization. Apart from adapting it to the regionalization problem, i.e. using regional target variables observed over stations rather than local variables observed over time, the method remains unaltered.

In order to reduce the regional heterogeneity and improve the predictive power of the multiple linear regression model, the study area is divided into regions with similar low flow behavior. This is done by fitting a global multiple linear regression model to the entire set of low flow index values y in the area. Prior to model fitting the set of potential explanatory variables $X = (x_1, x_2, \ldots x_p)$ are scaled via

$$x_i{}^* = \frac{x_i - \bar{x}_i}{s(x_i)},$$
(5.1)

where $\bar{x}$ is the mean and $s(x_i)$ the standard deviation of the observations of the respective variable x_i. The global model becomes

$$y = X^{*^T} b .$$
(5.2)

The variables that are selected as regressors during the fitting are then multiplied with their respective regression coefficient estimates

$$X_W = X^{*^T} b .$$
(5.3)

In this way, they are weighted according to their influence on the target variable, as proposed by Nathan and McMahon (1990a). The weighted set of explanatory variables is then subjected to a

k-means clustering algorithm. This algorithm aims at clustering data points in an *m*-dimensional vector space by finding *k* centers with minimum Euclidean distance to the individuals of a cluster. Accordingly, for a fixed number of centers

$$S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_j - \mu_i \right\|^2 \tag{5.4}$$

is to be minimized, with n_i the number of data points and μ_i the means of the clusters. The algorithm applied for optimization is the one according to Hartigan and Wong (1979), which uses several random starting points of the cluster centers.

The *m* dimensions of the problem correspond in this case to the number of columns in X_W, i.e. the number of regressors in the global regression model. The clustering algorithm is run with different numbers for *k*. The optimal number of groups should yield low internal variance within the groups, while at the same time keeping the number of groups minimal.

The advantage of using k-means for clustering is that only explanatory variables are used for classification that are observable for any point in the study area. Thus, sites with unknown target variable but known set of explanatory variables can be easily assigned to a group by finding the cluster center with minimal Euclidian distance. Still, the selection and weighting of *X* through regression prior to classification allows to consider the influence of the target variable in the clustering indirectly, resulting in a more relevant grouping for subsequent model fitting within the groups.

Multiple linear regression models within the subregions are fitted in the same way as the global model but without scaling.

Equally to the multiple linear regression models for temporal low flow modeling, preconditions are tested accordingly with help of the residuals. Instead of testing for autocorrelation, cross correlation in the respective regions are considered. GLS-models are fitted in case correlation between the catchments is high. Instead of using AR correlation structures, as explained in section 4.1.2.1, spatial correlation structures are included. These are incorporated using theoretical variogram functions, which will be explained in section 5.1.3.1.

5.1.2 Index-flood method

The index-flood approach was, as indicated by its name, developed for regional frequency analysis of floods (Dalrymple, 1960). It is based on the assumption that for a hydrologically homogeneous region, the distribution of peak flows is equal for all belonging catchments, when their flow values are scaled by an index value. This index value usually represents the average peak flow for any observed catchment and needs to be estimated for unobserved basins based on catchment characteristics. Here, the procedure is adapted to low flows while still referring to it by its original name. The application involves three steps: a) Separation of the study area into

hydrologically homogeneous regions, b) fitting of a regional frequency distribution for each region, and c) regionalization of the index value.

Homogeneous regions

Regional homogeneity is realized by grouping the catchments in the study area according to similarities in physiographic catchment descriptors that are particularly related to the target low flow variable. In order to be able to consider this relationship during grouping, a regression tree approach has been applied, as it appeared to work more efficiently than the k-means grouping described in section 5.1.1.

Classification of data via regression trees is done via recursive partitioning. Starting point is a whole set of observations of some target variable y. This set is split into two subsets, which each may again be split into two subsets and so on. These subsets are called the nodes N of the tree. Basis of the splitting is the maximum reduction of the variance of y within a node. The variance reduction for a node N_T split into nodes N_L and N_R is defined as

$$\Delta s^2 (N_T) = s^2 (N_T) - s^2 (N_L) - s^2 (N_R) \qquad (5.5)$$

where

$$s^2 (N) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}(N))^2 \; . \qquad (5.6)$$

$\bar{y}$ *(N)* denotes the mean of the observations of y in node N, n is the number of observations in node N.

The splitting is done using the independent variables x rather than the target variable y. A candidate variable x is sorted according to size. Accordingly, the individuals of the target variable are arranged. Now the split is sought, which maximizes the variance reduction in y. This is repeated for all candidate variables and splits until maximum variance reduction is achieved. The algorithm continues until no further variance reduction can be obtained. The number of nodes required for this analysis is determined using heterogeneity measures, as will be explained in the following paragraph.

Homogeneity of a region refers to the similarity of its belonging catchments in their low flow behavior. Catchments are thus considered similar if the distribution of their low flow time series is similar, which constitutes the idea of the index flood method, i.e. a single distribution function can be fitted to an entire region of catchments. In terms of populations, this signifies that all at-site moments are identical throughout the region - with exception of the mean. Observed differences are hypothetically due to sampling variability given the short periods of observation. In order to assess this variability, Hosking and Wallis (1997) suggested a heterogeneity measure based on L-moments. L-moments will be described in detail in section 6.1.3 including their

advantages over traditional moments and their utilization in distribution fitting. Here it should suffice to note the analogy of L-moments to conventional moments and sample L-moments to sample estimates of population moments.

For each site in a region, sample L-moments can be obtained, including the L-coefficient of variation (L-CV) t, the L-skewness t_3, and the L-kurtosis t_4. These moments are, like conventional moments, related to the scale and shape of a probability distribution. From these at-site L-moments, a regional average over N sites can be obtained, weighted according to the record length n_i at site i:

$$t^R = \frac{\sum\limits_{i=1}^{N} n_i t_i}{\sum\limits_{i=1}^{N} n_i} \, .$$

(5.7)

With the regional L-CV, the standard deviation of the sites' L-CVs in a region can be obtained as

$$V = \left(\frac{\sum\limits_{i=1}^{N} n_i \left(t_i - t^R \right)^2}{\sum\limits_{i=1}^{N} n_i} \right)^{\frac{1}{2}} \, .$$

(5.8)

The regional L-CV, L-skewness and L-kurtosis, as well as a value of 1 for the regional L-location are used to determine the 4 parameters of a regional kappa distribution

$$F(x) = \left(1 - h \left(1 - \frac{\kappa (x - \xi)}{\alpha} \right)^{\frac{1}{\kappa}} \right)^{\frac{1}{h}} \, .$$

(5.9)

with ξ as the location, α as the scale and κ and h as the shape parameters.

The kappa distribution is selected due to its high flexibility resulting from its high number of parameters. The distribution is used to simulate a large number of realizations for N sites with record lengths n_i. V is obtained for each simulated region, which are all homogeneous, as they come from the same distribution. The set of simulated V values will be denoted V_{sim}, the observed V value V_{obs}. From the set of calculated V_{sim}, mean μ_{Vsim} and standard deviation σ_{Vsim} are computed. A region's heterogeneity measure is than computed as

$$H = \frac{\left(V_{obs} - \mu_{Vsim} \right)}{\sigma_{Vsim}} \, .$$

(5.10)

According to Hosking and Wallis, regions with H < 1 can be considered as acceptably homogeneous, regions with 1 < H < 2 are possibly heterogeneous and values of H > 2 indicate definite heterogeneity.

Heterogeneity is tested in this way for all nodes of the regression tree. The ideal number of nodes is determined as the minimum number of nodes required to reduce H in all subgroups to below 1, or the minimum number of nodes at which no further improvement of homogeneity can be obtained.

Fitting of regional probability distributions

Once the homogeneous regions are determined, probability distributions can be fit for each region individually. Distribution fitting is done using weighted regional L-moments, computed according to equation (5.7), to estimate the respective distribution parameters. The goodness of fit of the fitted distribution is evaluated using another measure suggested by Hosking and Wallis. This measure is calculated by firstly fitting a set of candidate distributions to the regional L-moments, along with a kappa distribution, as described in equation (5.9). From this kappa distribution, again a large number of realizations are simulated, as explained above. For all simulated regions, the regional L-kurtosis $t_{4,m}$ is computed and the bias

$$B_4 = \frac{1}{N_{sim}} \sum_{,=1}^{N_{sim}} \left(t_{4,m} - t_4^R \right)$$

(5.11)

and standard deviation

$$\sigma_4 = \left(\frac{1}{N_{sim} - 1} \left(\sum_{,m=1}^{N_{sim}} \left(t_{4,m} - t_4^R \right) - N_{sim} B_4^2 \right) \right)^{\frac{1}{2}}$$

(5.12)

are calculated. For each of the fitted candidate distributions, the L-kurtosis τ_4 is obtained and the goodness of fit is evaluated via

$$Z^{DIST} = \frac{\tau_4 - t_4^R + B_4}{\sigma_4}.$$

(5.13)

For values of $|Z^{DIST}| \leq 1.64$ the fit of a candidate distribution is considered acceptable.

Index value regionalization

The last step of the index-flood method is estimating the index value that adjusts the regional frequency distribution to the level of the low flow indices at the gauges. The index corresponds to the mean of the respective low flow series at each site, and needs to be estimated for ungauged catchments. The regionalization of this index value is done using the multiple linear

regression approach described in section 5.1.1. The index method thus allows through mere regionalization of the mean an estimation of the entire low flow behavior at ungauged sites.

5.1.3 Geostatistical approaches

In addition to the statistical approaches named above, a number of geostatistical methods, called kriging (Krige, 1951; Matheron, 1963), are tested for regionalization of low flow indices. Rather than making predictions based on the relationship between catchment characteristics and low flow indices, these methods make use of the spatial correlation structure between gauges. This structure is represented by the so-called semivariogram, which poses the basis for all the methods described below. The empirical semivariogram is obtained by computing the semivariances for several distance classes available in the data. For any such distance class h, the empirical semivariance $\hat{\gamma}(h)$ is computed as

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} \left(z(x_i) - z(x_i + h) \right)^2 \, ,$$
(5.14)

where $n(h)$ is the number of point pairs in that distance class. $z(x_i)$ and $z(x_i + h)$ denote the observations at points x_i and $x_i + h$. This type of representation indicates that in kriging the observations are treated as realizations of a random variable Z. This variable has a continuous spatial variation, which is modeled as a function of distance with help of the semivariogram.

For the semivariogram to be appropriate to model the spatial variance structure, the random variable Z is assumed to be intrinsically stationary. Intrinsic stationarity implies constant expected values throughout the domain and expected values of the variance to be invariant with respect to location.

The empirical semivariogram can be used to fit a theoretical variogram model, yielding semivariance values also for unobserved lags. There exist a variety of models, able to describe the semivariance structure, some of which will be described later on. All models use three important parameters to describe the variogram, i.e. the nugget C_0, which indicates the semivariance between points at zero distance (discontinuity at the origin) that results from several effects like measurement errors, the sill s, which represents the maximum semivariance between point pairs at large distances, and the range r, the distance for which the semivariogram values approach or equal the sill.

With help of the theoretical variogram model, prediction for any unobserved point x_0 can be made. The methods described below all make use of the semivariogram, but differ partly in their assumptions and implementation.

5.1.3.1 Ordinary kriging

The first geostatistical method tested is Ordinary Kriging (OK). OK bases on the intrinsic stationarity assumption made for the variogram. Predictions for any value $\hat{Z}(x_0)$ at an unobserved point x_0 are made according to OK via

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i), \qquad (5.15)$$

where λ_i represents a vector of weights. The weights are determined via optimization. The estimator is BLUE, like the OLS estimates of the regression coefficients described in section 4.1.2.1 and thus subject to two constraints: it needs to minimize the prediction variance and be unbiased. Thus

$$Minimize \; Var\left(\hat{Z}(x_0) - z(x_0)\right),$$

$$(5.16)$$

$$while \qquad E\left(\hat{Z}(x_0) - z(x_0)\right) = 0 \, .$$

In terms of the weights, unbiasdness is guaranteed by setting their sum to 1:

$$\sum_{i=1}^{N} \lambda_i = 1 \, . \qquad (5.17)$$

The prediction variance can be expressed in terms of semivariogram values obtained from the theoretical variogram model

$$Var\left(\hat{Z}(x_0) - Z(x_0)\right) = 2\sum_{i=1}^{N} \lambda_i \gamma(x_i - x_0) - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \gamma(x_i - x_j), \qquad (5.18)$$

where $\gamma(x_i - x_0)$ represents the semivariogram value for the distance between the observed points and the unknown point, and $\gamma(x_i - x_j)$ those for distances between the observed points. Minimizing equation (5.18) subject to equation (5.17) requires inclusion of a Lagrange multiplier $\psi(x_0)$ in order to achieve minimization. The resulting kriging system for solution of the optimization problem in terms of semivariogram values is

$$\begin{bmatrix} \lambda_i \\ \psi(x_0) \end{bmatrix} = \begin{bmatrix} \gamma(x_1, x_1) & \cdots & \gamma(x_1, x_N) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_N, x_1) & \cdots & \gamma(x_N, x_N) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(x_1, x_0) \\ \vdots \\ \gamma(x_N, x_0) \\ 1 \end{bmatrix} . \qquad (5.19)$$

The weights thus not only depend on the variance between unobserved and observed points but on relationships between all observable point pairs. Similar data points will be given lower

weight, enabling kriging to account for spatial clusters in the observed data. With the estimated weights, the prediction for point x_0 may be calculated using equation (5.15). The prediction variance can be obtained via

$$\sigma^2(x_0) = \sum_{i=1}^{N} \lambda_i \gamma(x_i - x_0) + \psi(x_0). \tag{5.20}$$

5.1.3.2 Topological kriging

Topological Kriging (TK), developed by Skøien et al. (2006) adapts the OK approach to the stream network problem. Accordingly, nested catchments are hypothetically more similar than independent catchments. Furthermore, measured discharge is assumed to represent an aerial integral over the belonging catchment up to the point of measurement. The latter point is taken into consideration in TK through the application of block kriging. The realization of a variable Z is thereby regarded as an integral over the entire spatial support, i.e. the catchment area, made up by position vectors x_i

$$z(A_i) = \frac{1}{A_i} \int_{A_i} z(x_i) dx_i . \tag{5.21}$$

Consequently, the kriging estimator according to equation (5.15) becomes

$$\hat{Z}(A_0) = \sum_{i=1}^{N} \lambda_i z(A_i). \tag{5.22}$$

Semivariance calculation needs to be extended to handle areas rather than points. This is achieved by regularization of a point variogram γ_p. For any pair of observed values, the expected semivariance $\bar{\gamma}(x_i, x_j)$ is thereby calculated via

$$\bar{\gamma}(x_i, x_j) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} \gamma_p(x_i - x_j) dx_i dx_j$$

$$- 0.5 \left(\frac{1}{|A_i|^2} \int_{A_i} \int_{A_i} \gamma_p(x_i - x_j) ds du + \frac{1}{|A_j|^2} \int_{A_j} \int_{A_j} \gamma_p(x_i - x_j) dx_i dx_j \right). \tag{5.23}$$

In the first part of the equation, the total variance between the catchments is integrated. In the second part, the averaged variance within the catchments is substracted. The larger the overlapping area between the two catchments, the smaller becomes the expected semivariance. Integration over the catchment area cannot be achieved analytically and is thus done via discretization of the catchments using regular grids. By using the pairwise computed expected

semivariance values in the kriging system given in equation (5.19), the weights for prediction can be determined. Likewise, the prediction variance can be obtained using equation (5.20) with $\overline{\gamma}(x_i, x_0)$.

According to Skøien et al. (2006) regularization of point variograms result in disappearance of the nugget, even if discontinuity at the origin is prominent. Therefore, they suggest separate regularization of the nugget effect. The nugget may be regularized via

$$C_0\left(A_i, A_j\right) = 0.5\left(\frac{C_{0p}}{A_i} + \frac{C_{0p}}{A_j} - \frac{2C_{0p} \cdot Meas\left(A_i \cap A_j\right)}{A_i A_j}\right),$$

(5.24)

where $Meas(A_i \cap A_j)$ denotes the overlapping area of two catchments with areas A_i and A_j. The regularized nugget is simply added to the regularized variogram.

TK is carried out using the R package "rtop" for interpolation of data with variable spatial support (Skøien, 2016).

5.1.3.3 Kriging with external drift

Kriging with external drift (EDK) extends OK through the inclusion of external variables that are linearly related to the target variable Z. The kriging estimator for an unobserved location x_0 is knowingly given by equation (5.15). Additionally, Z at any observed and unobserved location can be predicted by a set of K variables $y_1(x)$, …, $y_K(x)$ with unknown coefficients $\beta_{1, …, } \beta_K$

$$Z(x) = \sum_{k=1}^{K} \beta_k y_k(x) + \varepsilon(x).$$

(5.25)

The expected value of $\hat{Z}(x_0)$ becomes

$$E(Z(x)) = \sum_{k=1}^{K} \sum_{i=1}^{N} \beta_k \lambda_i y_k(x_i)$$

(5.26)

and a second unbiasedness criterion is included, namely

$$\sum_{i=1}^{N} \lambda_i y_k(x_i) = y_k(x_0) \quad for \ k = 1, 2, …, K.$$

(5.27)

Accordingly, the kriging system in equation (5.19) is extended to

$$
\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \psi_0 \\ \psi_1 \\ \vdots \\ \psi_K \end{bmatrix} = \begin{bmatrix} \gamma(x_1,x_1) & \cdots & \gamma(x_1,x_N) & 1 & y_1(x_1) & \cdots & y_K(x_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(x_N,x_1) & \cdots & \gamma(x_N,x_N) & 1 & y_1(x_N) & \cdots & y_K(x_N) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ y_1(x_1) & \cdots & y_1(x_N) & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_K(x_1) & \cdots & y_K(x_N) & 0 & 0 & \cdots & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(x_1,x_0) \\ \vdots \\ \gamma(x_N,x_0) \\ 1 \\ y_1(x_0) \\ \vdots \\ y_K(x_N) \end{bmatrix} . \qquad (5.28)
$$

With the found kriging weights, the prediction variance can be determined according to equation (5.18) (Webster and Oliver, 2007).

5.1.3.4 Physiographical space based interpolation

A different approach for optimization of the spatial interpolation is physiographical space based interpolation (PSBI). PSBI, developed by Chokmani and Ouarda (2004), aims at improving interpolation, which commonly is done in 2-dimensional geographic space, by transferring it to a so-called physiographic space. Rather than of x- and y-coordinates, the new dimensions consist of combinations of physiographical and climatic catchment descriptors. These are created using multivariate techniques. Here, principal component analysis has been applied. The procedure is equal to the one described in section 4.1.2.1 with the set of catchment descriptors as input. Once the physiographical space has been established, interpolation can be conducted with the desired technique. Close attention has to be paid regarding the stationarity prerequisites of the kriging methods, as in physiographic space trends will become prominent through ordering of the spatial values according to external factors. Therefore, universal kriging (UK), which is EDK with the physiographic x- and y-coordinates as drifts, and deterministic inverse distance weighting (IDW) are applied here. In IDW the value at an unobserved location is determined as the weighted sum of observed values of surrounding stations

$$
\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i), \qquad (5.29)
$$

as for the geostatistical approaches shown before. The individual weights are in this case determined according to the inverse distance of an observed location to the unobserved point

$$
\lambda_i = \frac{1}{d^x(x_0,x_i)} . \qquad (5.30)
$$

The power x can be adjusted to represent the loss of similarity with distance. Since the sum of weights is to add up to 1, the following estimator is obtained

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \frac{\dfrac{z(x_i)}{d^x(x_0, x_i)}}{\dfrac{1}{d^x(x_0, x_i)}}. \tag{5.31}$$

5.2 Data preparation

For regionalization record length is of much lesser importance than station density. Thus, a minimum of 20 years has been chosen, in order to facilitate proper calculation of sample statistics at each gauge. The period from 1988 to 2009 emerged with the highest number of available stations, namely 137. This period is used to calculate average annual low flow indices at each gauge. A common period for the analysis is important under the presumption of non-stationarity, i.e. that long-term statistics vary significantly in time. In order to include other stations, whose records lie outside of this period or just overlap partially with it, a transfer method is applied. The objective is to identify the relationship of each stations' record to the period of 1988 to 2009 and use it to transfer the low flow indices from their observed to the main period.

The principle of this approach is depicted in Figure 5.1. Gauge number 3 does not have observations for the entire main period, while all surrounding gauges have. Out of these gauges, those are selected, that have a common period with gauge 3, which are gauges 1, 2 and 4. Based on these gauges, a simple linear regression model is fitted. Target variable is the average low flow index value for the main period, the regressor is the low flow index for station 3's period. The identified model is then used to transfer the low flow index at gauge 3 to the main period. This transfer is carried out for all available gauges, so that in the end, 221 stations are available for analysis.

Primary target variable for all regionalization approaches is the mean annual low flow index value for the period 1988 – 2009. The MLR, as well as the index method are applied to additionally regionalize low flow quantiles, i.e. quantiles for 10-, 20- and 50-year return periods. Since the index method requires the full set of annual index values for L-moment calculation, the original, non-extended data set of 137 stations is used for quantile prediction. This allows the comparison of regionally estimated quantiles with local ones at every station without validation of the stationarity criterion.

Where appropriate, the target variables are normalized through division of the values by the respective catchment areas. This applies to the NM7Q, Q_{95}, as well as V_{max}. Normalized variables are denoted NM7q, q_{95}, and V_{max}. For all analyses entire catchments up to the point of measurement are considered, i.e. the individual low flow values represent integrals over the belonging catchment area. Overlapping areas are tolerated and upstream measurements are not substracted from downstream ones. Potential dependencies that consequently remain in the

data sets are handled from case to case. For geostatistical approaches that require point data as input in order to compute Euclidean distances, the center of gravity of the catchment polygons is considered.
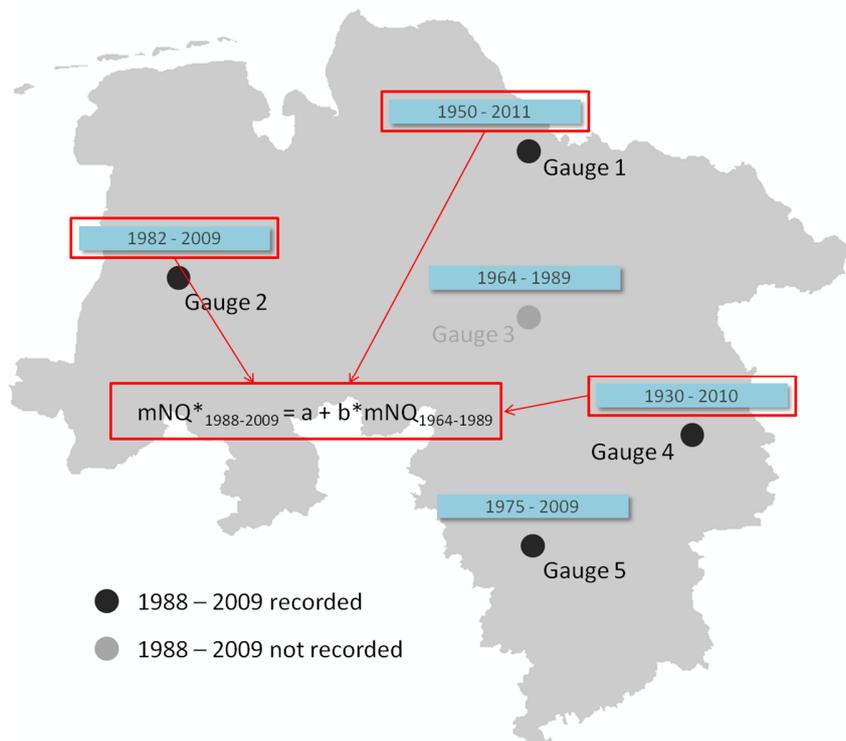
Figure 5.1: Data set expansion approach.

The spatial distribution of the individual indices can be seen in Figure 5.2. The NM7q appears quite randomly distributed but catchments in the north and the south appear to show higher values than catchments in the center of Lower Saxony. Almost an identical pattern can be seen for q95. The average maximum V appears to be related to catchment size, as highest values are observed in the smallest catchments, which goes along with their highly variable flow. The distribution of the average maximum D is quite heterogeneous without distinguishable spatial patterns. The average low flow timing, on the other hand, exhibits a clear regional pattern. In the southern parts, low flows occur much later than in the northern parts. Individual headwater catchments show earlier or later low flow occurrence independent from the main river or surrounding gauges.

The MLR, EDK, and PSBI require input of external variables. Initially, the set of physiographic catchment descriptors described in section 3.2.3 was designated for this task. However, the set was extended through inclusion of aggregated meteorological indices for the period of 1988 - 2009. Aggregation was simply achieved by computing mean, minimum and maximum of the catchments' annual meteorological index values for all base periods and lead times.
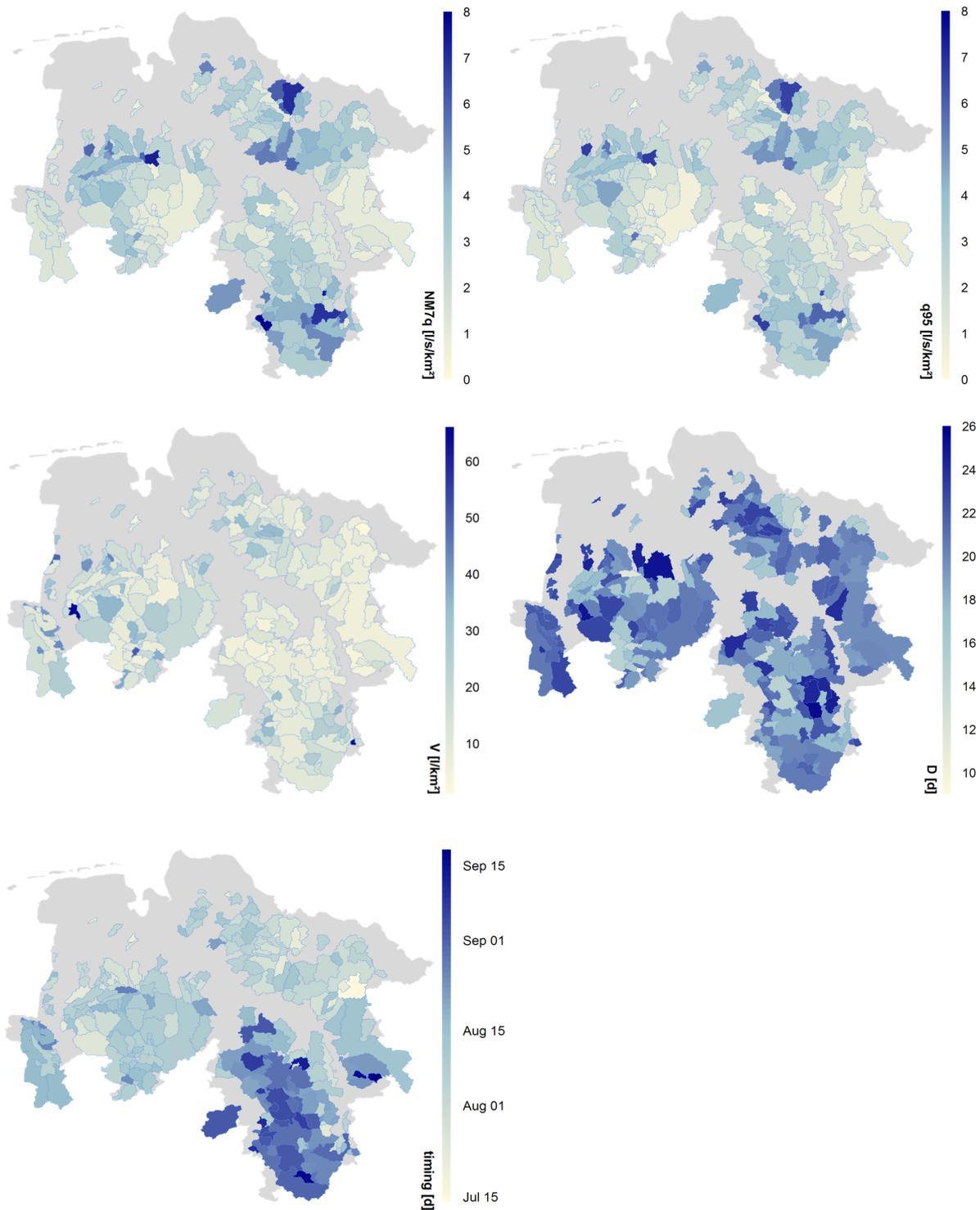
Figure 5.2: Observed Regional distribution of the 5 low flow variables NM7q, q95, V_{max}, D_{max} and timing.

The catchments included in the spatial analysis exhibit quite a range of sizes. The mean catchment area is 394 km², the median 110 km² and maximum and minimum areas are 6443 km

² and 11 km², respectively. The distribution of catchment areas is depicted in Figure 5.3. Despite the high variance in areal support, no catchments were excluded from the analysis based on their size. Possible resulting effects are to be analyzed during application of the individual methods.
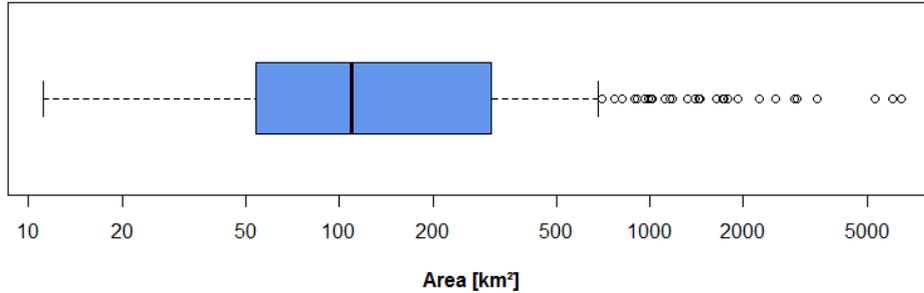
Figure 5.3: Distribution of areal extent of the catchments used in the study.

Regional low flow data appeared to be non-normally distributed. Since normality is a precondition for the majority of the applied methods, the target variables have been transformed using Box-Cox transformations (Box and Cox, 1964). This is simply done according to

$$y_\lambda = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ ln(y) & \text{if } \lambda = 0 \end{cases}, \tag{5.32}$$

where λ is a power parameter, which is estimated as the value that maximizes the normal log-likelihood function.

5.3 Model fitting and evaluation of model performance

For all considered regionalization approaches, model fitting is carried out on the entire data set. Leave-one-out cross validation (LOOCV) is then applied to assess the performance of the individual models.

Variable selection for the regional MLR is done using the same two-way BIC minimization algorithm, as described in section 4.3. The same restrictions for normality, independence and homoscedasticity of the residuals apply. Multicollinearity is penalized accordingly. The number of variables selected is additionally restricted by the number of available stations, i.e. 221 for fitting of the global model and respectively fewer for the model fitting inside the k-means clusters. As a rule of thumb, the maximum number of variables is set to 1 per 10 observations. The k-means clusters are reordered for each target variable. Spatial correlation is tested for every target variable and GLS is applied rather than MLR if necessary.

The mean low flow indices estimated during LOOCV pose the scaling parameter value input for the index flood method. Before being able to fit regional distribution functions and derive local quantiles, a clustering into homogeneous regions is necessary. As target variable for the regression tree approach the sample L-CV is selected, as estimate of a stations' shape of probability distribution. Since annual observations of low flow indices are required to fit regional frequency distributions, the non-extended data set of 137 stations is used for analyses with the index method. Again, under the premise of temporal non-stationarity in the data, only the record from 1988 - 2009 is used for model fitting, even if longer records are available. Weighting due to different observation periods thus does not need to be considered.

In order to evaluate the suitability of the regional frequency distribution in a homogeneous region, it is fitted using all but one station in a region. The left-out stations' quantiles are then estimated using the regional frequency distribution and the estimated index value. Performance is assessed through comparison of estimated quantiles with quantiles derived from distributions fitted directly to the observed record. Both k-means clusters and homogeneous regions are set as fixed and not re-estimated during cross validation for every left-out station.

Before application of OK, the data is screened for potential regional trends. If any such trend is observable, an appropriate function is fitted and subtracted from the data.

EDK and PSBI both require variable selection. For both methods the complete set of regressors, found during global MLR-fitting serves as set of candidate variables, for which all possible combinations are tested.

5.4 Results and discussion

In the following section, the results of the analyzed approaches will be shown and discussed in detail. Before carrying out any of the described methods, the data was normalized using a Box-Cox transformation with a λ of 0.4 for all approaches. The goodness-of-fit criteria shown for cross validation are computed for back-transformed values.

Multiple linear regression

Global MLR was carried out on the scaled data set, in order to pose the basis for weighted k-means clustering. The algorithm for variable selection was applied and 14 regressors were identified. The variables and their respective coefficient estimates are shown in Table 5.1. Assessment of the model performance for the entire study area using cross validation yielded an NRMSE of 55.4 %, a percent bias of -2.0 %, and an NSE and R^2 of 0.69.

Table 5.1: Estimated coefficients and k-means class centers for the individual variables for attribution of unobserved catchments.

Variable	Estimate (scaled)	1	2	3	4	5
Intercept	*-0.0006*					
Maximum SPI $_{3\text{-}1}$ [-]	0.72	1.59	1.36	1.48	1.55	1.28
Range $T_{min\ 3\text{-}2}$ [°C]	-0.57	5.73	4.29	5.73	4.97	4.50
Minimum $WSD_{mean\ 3\text{-}11}$ [d]	0.43	1.89	1.99	1.62	1.69	1.74
Minimum $G_{max\ 3\text{-}6}$ [W/m²]	-0.41	12.83	13.99	13.06	13.12	12.95
Minimum $P_{90\ 6\text{-}9}$ [mm/d]	0.38	7.38	5.14	4.74	5.29	5.08
Range $DSD_{mean\ 3\text{-}8}$ [d]	0.37	5.92	7.57	5.28	9.42	7.77
Maximum $DSD_{mean\ 12\text{-}6}$ [d]	0.27	4.79	5.67	5.27	5.68	5.55
Minimum $G_{mean\ 6\text{-}0}$ [W/m²]	-0.20	14.33	14.39	14.72	14.43	14.44
Range $T_{min\ 3\text{-}10}$ [°C]	-0.20	10.96	11.12	11.20	10.41	10.77
Minimum $DSD/WSD_{max\ 3\text{-}0}$ [-]	0.20	0.48	0.40	0.62	0.46	0.47
Range $WSD_{mean\ 12\text{-}11}$ [d]	-0.18	0.99	1.13	0.90	0.98	0.88
Maximum $P_{max\ 3\text{-}8}$ [mm/d]	-0.17	34.80	21.19	23.24	23.95	24.61
Percentage of urban area [%]	0.15	6.58	4.24	8.59	3.87	5.57
Air capacity [mm]	0.14	31.55	90.72	63.59	104.53	68.06

The global model comprises of quite a large set of explanatory variables. The model is consequently unhandy and potentially overfitted. However, it poses the ideal basis for selection and weighting of variables for grouping of the catchments. Based on the standardized set of catchment descriptors, weighted by their respective coefficient estimates, k-means clustering has been carried out for different numbers of groups, i.e. for 1 to 15. Figure 5.4 shows a very moderate turning point in the average within-group variance at a number of 5 clusters. This configuration has been selected for clustering as it suggests maximum variance reduction for a small number of groups.
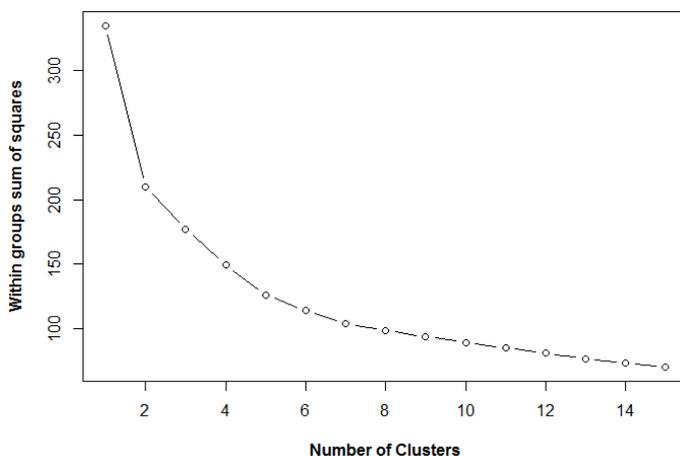
Figure 5.4: Number of clusters vs. average within groups sum of squares for weighted k-means clustering.

The resulting grouping can be seen in Figure 5.5. The cluster that is most prominent and usually identified first for all low flow indices is the Harz area, shown in yellow. Logically, being the only mountainous catchments in the study area subject to substantially different flow governing processes, the region sticks out. Unfortunately, the number of gauges within is small, which is problematic for proper model fitting. Otherwise, a west-east to northwest-southeast gradient becomes apparent in the positioning of the groups. The groups are more or less geographically continuous. The clusters for the other low flow indices can be found in Appendix C1.

Next to the coefficient estimates for the scaled data set, Table 5.1 summarizes the rescaled group centers of each variable for the five groups. Assignment of unobserved catchments can easily be conducted in terms of minimal Euclidean distance to the class centers. A cross-validation procedure has been carried out, forming k-means clusters with one station left out and subsequent assignment of the left out station to one of the groups. The procedure lead to misclassification of a single station only, indicating that the grouping is robust and that reclassification does not need to be considered during cross validation of the regionally fitted models.
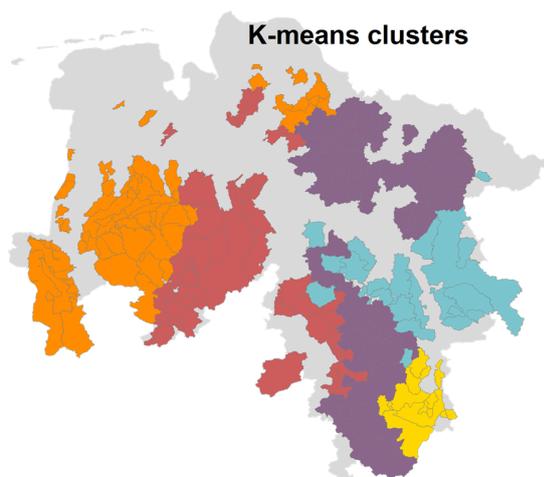
Figure 5.5: Clusters found with the weighted k-means clustering algorithm for interpolation of the mNM7q.

The models fitted for the individual regions can be found in Table 5.2 alongside the respective cross-validation results. It should be noted that clustering of the data set already contributes to explanation of some portion of the variance in the mNM7q (R^2 of 0.23). Beyond that, model fitting appears differently successful in the individual groups. Best model fit is achieved for the Harz region (region 5). However, against the set rule of choosing one regressor per 10 observations, 2 external variables are needed to achieve this result. Overfitting might be an issue. For the northwestern region (region 4), model fitting turned out almost equally successful.

Regionalization in the very East of Lower Saxony (region 1) is not quite as effective and regions 2 and 3 show the worst cross-validation results. Nevertheless, regional model fitting appears generally favorable, given the small numbers of regressors required and the increase in model performance. The overall quality criteria are significantly better than for the global model in all aspects, e.g. a regional NSE of 0.79 in comparison to a global NSE of 0.69.

The MLR results for the other indices in the form of scatterplots can be found in Appendix C2. As expected, the performance for the regionalization of the q95 is comparable to the NM7q but slightly lower. The regional NSE is 0.72. Prediction of V_{max} and D_{max} in space appears to be difficult. NSE values of 0.60 and 0.47 could be obtained. The index that was most readily regionalized is the low flow timing with a cross-validation NSE of 0.81. These results are in accordance with the findings of Sanborn and Bledsoe (2006) who found that flow magnitude and timing could be regionalized better than flow variability. The regional models of all indices are shown in Appendix C3.

Table 5.2: Regional regression models for the four groups and NSE calculated for cross and split validation

Group	No. of stations	Coefficients		NRMSE	pbias	NSE	R²
1	21	*Intercept*	*-1.05*	54.4 %	-0.3 %	0.69	0.69
		Minimum $WSD_{max\ 6-8}$ [d]	0.23				
		Range $P_{70\ 3-1}$ [mm/d]	0.30				
2	61	*Intercept*	*-3.55*	60.9 %	-0.6 %	0.62	0.63
		Minimum $WSD_{mean\ 12-10}$ [d]	1.64				
		Maximum $HWD_{mean\ 3-2}$ [d]	-0.28				
		Minimum $DSD_{mean\ 6-3}$ [d]	1.40				
		Minimum $P_{70\ 12-9}$ [mm/d]	2.55				
		Maximum $P_{60\ 3-4}$ [mm/d]	-0.93				
		Minimum $WSD_{mean\ 3-9}$ [d]	-1.20				
3	59	*Intercept*	*-0.07*	59.9 %	-1.5 %	0.64	0.64
		Maximum $P\text{-}ETP_{mean\ 3-2}$ [mm/d]	0.86				
		Minimum $DSD_{max}/WSD_{max\ 6-2}$ [-]	0.46				
		Highly conductive material [%]	-0.005				
		Minimum $HWD_{mean\ 12-7}$ [d]	2.82				
		Range $T_{min\ 3-10}$ [°C]	-0.51				
		Minimum $P_{50\ 6-8}$ [mm/d]	4.25				
4	66	*Intercept*	*8.97*	43.0 %	-0.9 %	0.81	0.81
		Maximum $SPI_{\ 3-1}$ [-]	0.78				
		Range $P_{80\ 6-10}$ [mm/d]	0.47				
		Maximum $DSD_{mean}/WSD_{mean\ 12-11}$ [-]	-1.08				
		Minimum $DSD_{max\ 6-2}$ [d]	0.14				
		Minimum $Tmax\ 3-2$ [°C]	-4.23				
		Circumference/area [-]	-2.87				
5	14	*Intercept*	*4.62*	39.9 %	0.1 %	0.83	0.84
		Maximum $P_{max\ 6-1}$ [mm/d]	0.27				
		Mean $HWD_{max\ 6-11}$ [d]	-2.07				
			Total	45.6 %	-0.8 %	0.79	0.79

Figure 5.6 and Figure 5.7 compare the global and the regional models directly. The regional distribution of the simulated values clearly shows a slightly smoother picture for the global than for the regional models, even though both methods reproduce the observed pattern in the mNM7q quite well. The scatterplot of the estimated vs. the observed mNM7q does not only emphasize the better fit of the regional models in comparison to the global one but also shows the partitioning of the regional mNM7q values into different magnitudes through grouping. Region 1, represented by the blue dots contains generally the lowest mNM7qs, while region 5 contains the highest.
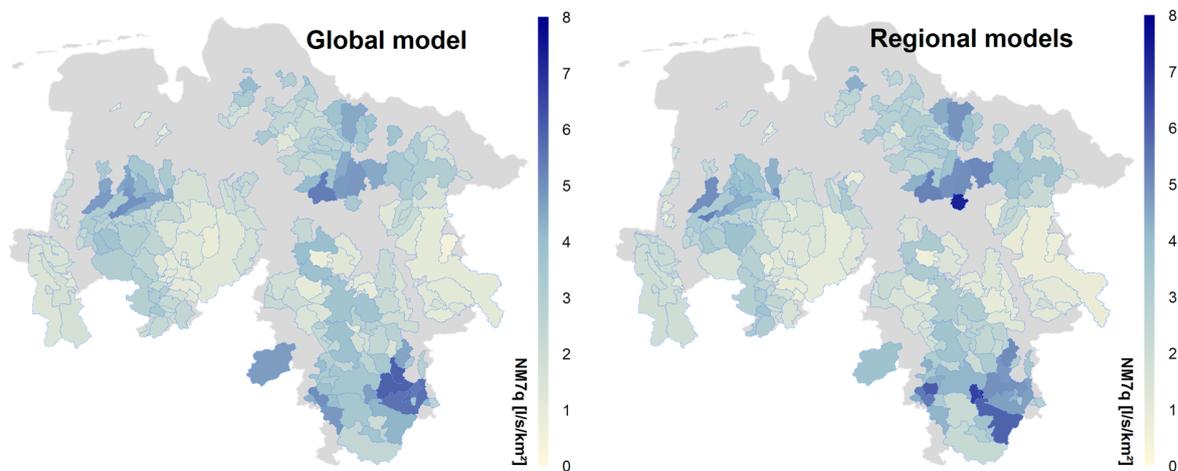
Figure 5.6: LOOCV estimated mNM7q for the global MLR model (left) and regional MLR models for the 5 k-means clusters (right).

For all regions, as well as the global model, the residuals have been examined after model fitting. Empirical variograms have been calculated and a range of theoretical models were tested as possible correlation structures in a GLS model fitting attempt. As shown in Figure 5.8 for the residuals of the global model no obvious spatial pattern was observable and GLS fitting was considered redundant.

It is striking that the majority of regressors in the global and regional models is made up of aggregated meteorological indices, rather than physiographic catchment descriptors. Inclusion of the indices thus appears favorable. Two main reasons come into consideration. The first one is the peculiar homogeneity of the landscape of the study area. The flatness, similarity of soils etc. makes it hard to identify relevant physiographic descriptors that explain the differences in low flow magnitude. The choice of physiographic variables appears poor and not readily applicable for low flow estimation. Also, it seems possible that differences in magnitude arise primarily through scaling effects. The advantage of using meteorological indices may be the reproduction of these effects in the flow variable. Scaling effects arise due to variable support of the

catchments. As described in section 5.1.3.2, the flow measured at a gauge can be considered as the integral over the entire catchment's area. The larger the area, the more moderated the flow, due to long flow times and counterbalancing of extreme events. Small headwater catchments, on the other hand, react fast to meteorological events and show much more erratic flow. Integrating the flow causing mechanisms, i.e. the meteorological variables over the catchment area just like the flow variable itself accounts for these very effects. This is of special relevance for low flows, which are caused by long phases of low precipitation and high evapotranspiration. Computing indices for meteorological time series averaged over the respective catchment area will result in more extreme events for smaller and more moderate events for large catchments, in accordance with the low flow index values at the belonging gauges. Additionally, the averaged meteorological indices are naturally capable of reflecting climatic gradients in the study area and are suitable to pattern topographic effects. All in all, apart from representing climatic differences in the study area, the aggregated meteorological indices are capable of accounting for effects, like the degree of alleviation due to catchment extent, which the physiographic descriptors, including actual catchment size, cannot.
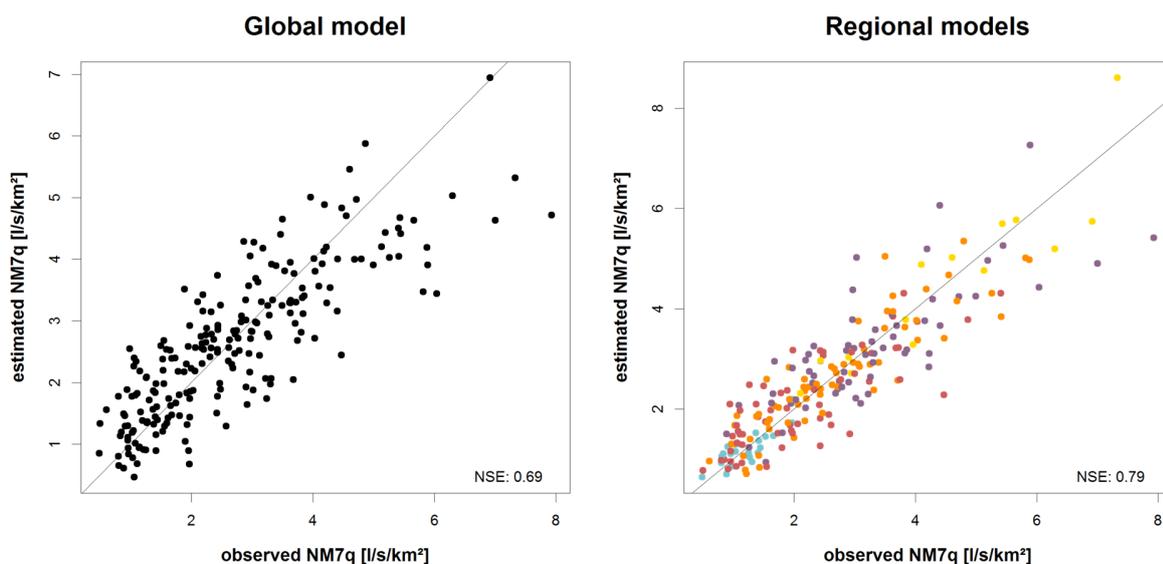
Figure 5.7: LOOCV estimated vs. observed NM7q for the global MLR model (left) and regional MLR models for the 5 k-means clusters (right).

Most of the regional models contain regressors related to wet or dry spell or heat wave duration. As seen in Figure 5.9, those indices exhibit a small-scale regional pattern, capable of explaining significant differences in low flows at catchments with direct spatial proximity. Out of the entire set of meteorological indices, these are the ones related the most to variability, and, since they are computed for precipitation and temperature time series that are integrated over the entire catchment area, may account for scaling effects more than any of their counterparts.

Figure 5.9 shows the spatial distribution of the first five external variables that have been selected during fitting of the global MLR model. P_{90} shows a pattern that highlights the Harz region with highest precipitation in the study area but also shows lower values in the central parts. Average global radiation portrays an inverse picture with highest values in the center without any anomaly in the Harz. The SPI is characterized by high values in the very Northeast and Southeast of Lower Saxony with driest areas again in the center. In the range in minimum temperature a prominent gradient from smaller variations in the West to larger differences in the East can be observed. Ranges in average wet spell duration, on the other hand, show quite a small-scale variability between the individual catchments, something, which is not observable in any of the physiographic catchment descriptors. Still, the fact that climatic indicators give better approximation of low flow indices than do physiographic variables contradicts various findings, like the one of Yaeger et al. (2012) and Gudmundsson et al. (2011), who found that low flows rely more on catchment characteristics than on climate variations.
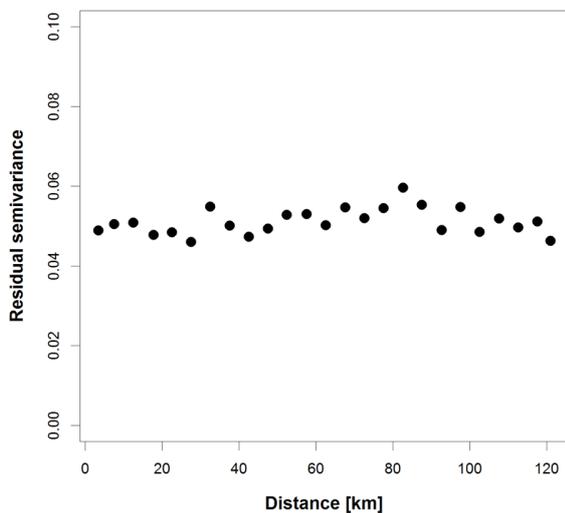
Figure 5.8: Empirical semivariogram of the residuals of the global MLR model.

Geostatistical methods

When looking at the spatial distribution of the observed mNM7q, a certain pattern can be detected, as mentioned before, namely that values turn out to be smaller in the center of the study area and larger in both North and South. This trend logically becomes most prominent in North-South direction, as shown in Figure 5.10. In an attempt to make the spatial data stationary for OK, the trend has been estimated as a simple polynomial model

$$trend = 1.58 - 0.61y + 2.45y^2 \,, \tag{5.33}$$

where *y* denotes the northing of the gauges, and removed. It is added back after kriging and included in the cross-validation results.
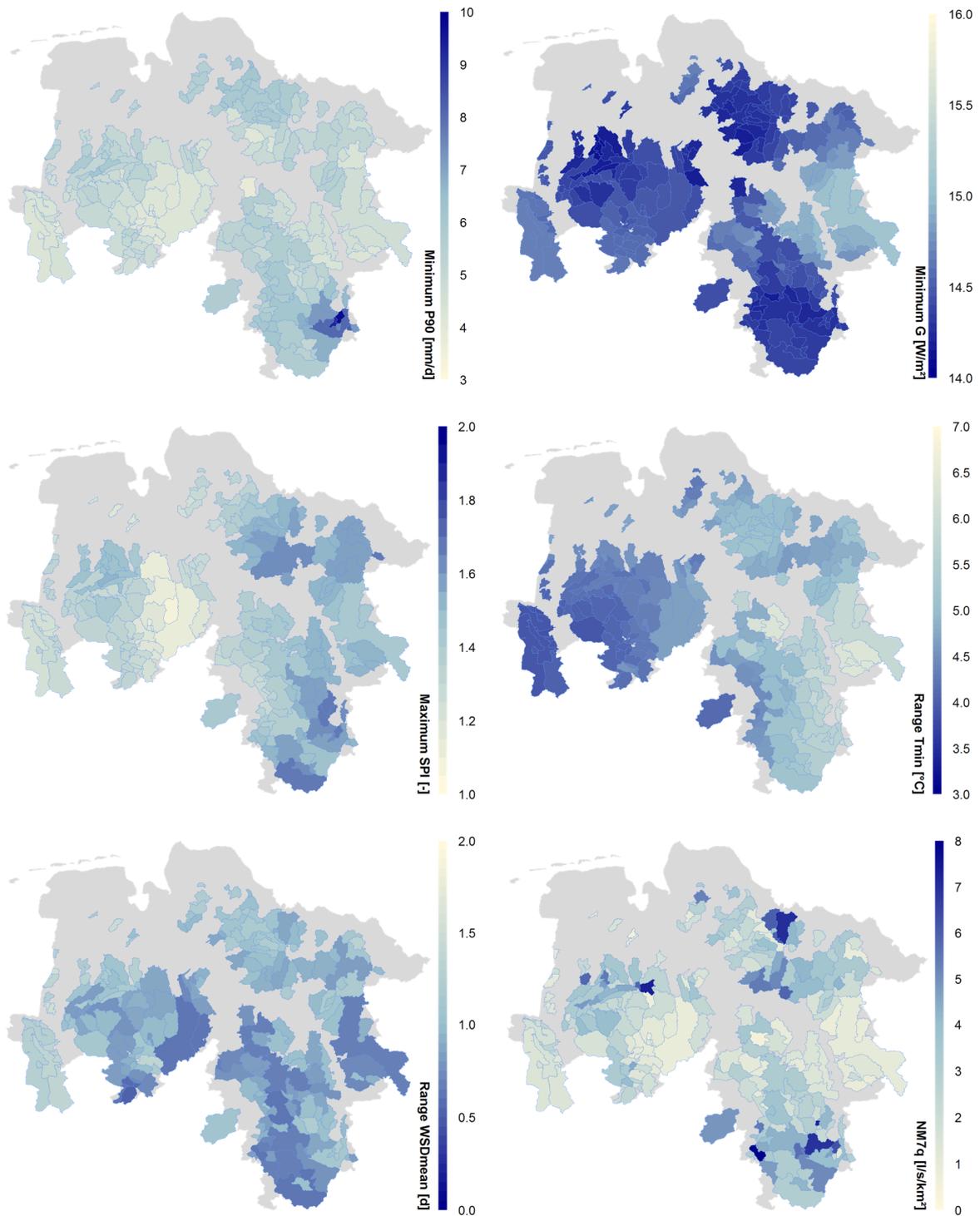
Figure 5.9: Spatial distribution of the first five external variables selected for the global MLR model in comparison to the observed spatial distribution of the mNM7q.
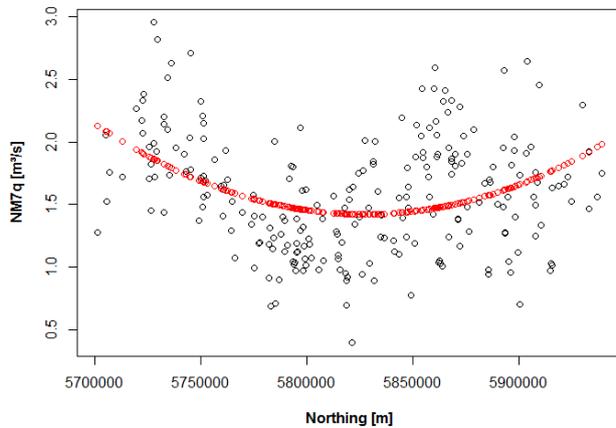
Figure 5.10: Observed mNM7q in North-South direction and estimated trend.

The residual data set has been used for semivariogram fitting. The empirical variogram shown by the dots in Figure 5.11 indicates that a spatial correlation structure exists between catchments. A rise in semivariance can be observed for a distance below roughly 60 km before it stagnates. The nugget is quite high, which is expectable for the type of problem at hand; nearby gauges that belong to catchments of different rivers naturally show different values.
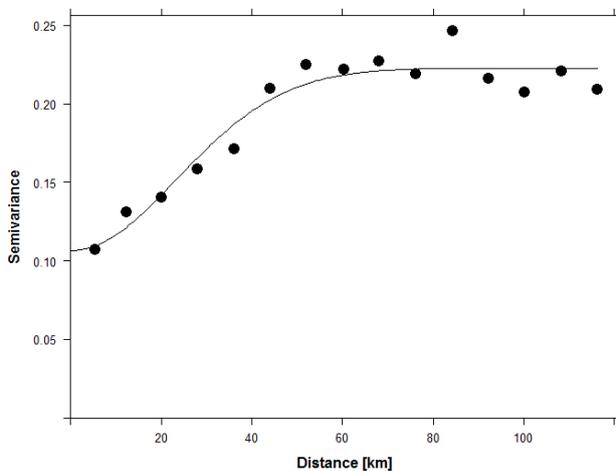
Figure 5.11: Empirical and fitted theoretical semivariogram for OK.

Despite averaging over generously wide distance bins, the empirical variogram did not appear very smooth, indicating a large variability within the bins and therefore quite significant error of the fitted theoretical variogram. A theoretical variogram was fitted, nonetheless. A good approximate fit was achieved using a Gaussian model in the following form:

$$\gamma(h) = 0.12 \left(1 - exp\left(-\frac{h^2}{33.26} \right) \right) + 0.11 \; , \tag{5.34}$$

displayed by the line in Figure 5.11.

Local Kriging has been applied using 20 neighbors and a maximum radius of 60 km for each station. The LOOCV yielded an NRMSE of 71.9 %, a percent bias of -1.6 %, and an NSE and R² of 0.39. In order to increase the performance, additional kriging neighborhoods were tested. Best results were obtained when kriging was carried out inside the homogeneous regions with 10 neighbors, but only in terms of NSE and R², which both increased to 0.44. NRMSE and percent bias, however, also increased to 75.4 % and -2.7 % respectively. As shown in Figure 5.12, the regional prediction appears severely smoothed but the general pattern of the observed NM7q seems to be met. The prediction variance is generally higher for smaller than for larger catchments. Compared to the MLR approach shown above, OK consequently performs significantly worse.
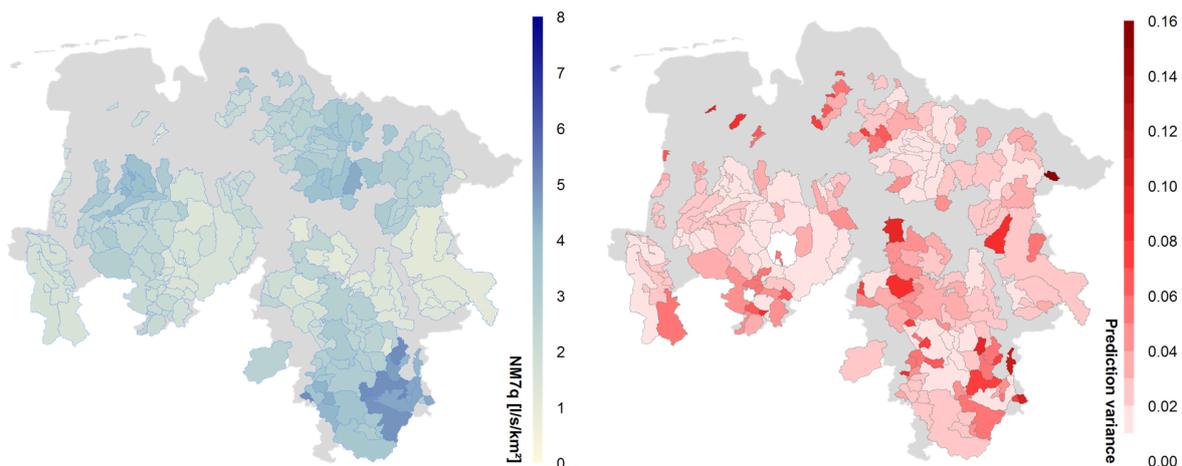
Figure 5.12: Predicted mNM7q and prediction variance obtained for OK during cross-validation.

In order to investigate possible effects of area and nestedness of catchments on the spatial correlation structure, separate empirical semivariograms have been computed for point pairs that are directly connected through a stream, i.e. whose catchments are nested, and for the remaining unconnected point pairs. Figure 5.13 shows the difference between the two variants. The left graph shows the semivariogram for the nested catchments. One cannot see the typical initial increase in semivariance and a subsequent flattening, but the opposite. Semivariance is generally low for all lags until it finally increases significantly. This behavior most probably indicates a spatial trend in the in-stream mNM7q. The unconnected point pairs, on the other hand, exhibit the typical correlation structure and appear to be unaffected by any trend. Since

the number of point pairs directly related through the same river is small in comparison to the unrelated pairs, the observed effects disappear within the combined semivariogram.
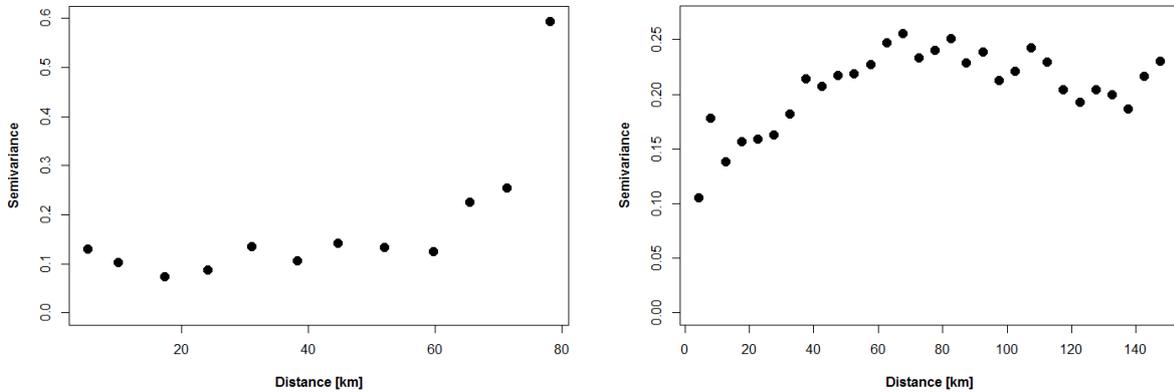
Figure 5.13: Empirical variograms for connected (left) and unconnected point pairs (right).

For further investigation of the influence of catchment size on the similarity of values, variograms have been calculated as a function of difference in area rather than of distance between point pairs. The result can be found in Figure 5.14. For nested catchments it becomes obvious, that the variance increases with increasing difference in area. This is expected, as for connected catchments, areal difference is equivalent to distance. For unconnected point pairs, however, the effect is reversed. The larger the difference in catchment size, the more similar the values. This is contrary to any expectation and eludes any logical reasoning.
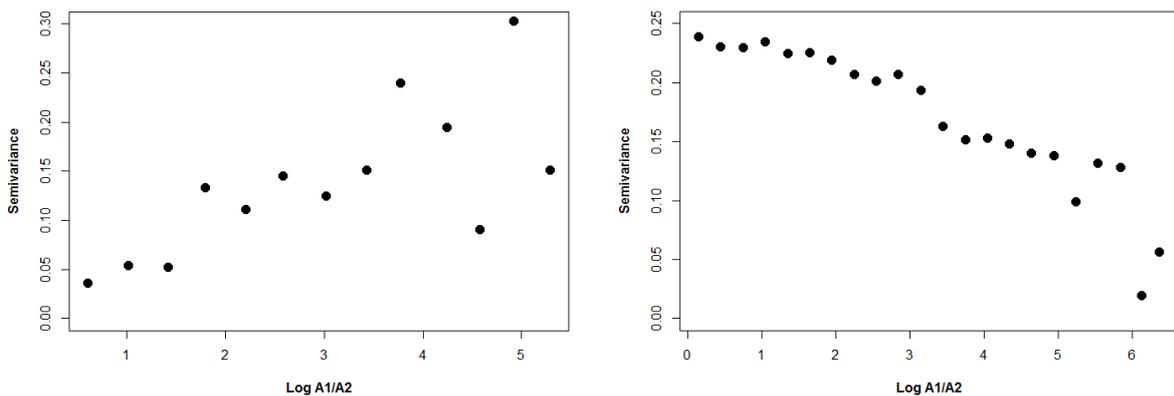
Figure 5.14: Empirical variance for areal differences between connected (left) and unconnected point pairs (right).

In summary, negligence of the spatial support of the catchments, as done in OK is consequently not meaningful for geostatistical approaches. Therefore TK has been applied as a method capable of incorporating areal effects. As described before, TK is able to simultaneously capture

the alternating variance over distance and catchment size between observations by considering separate variograms for various areal classes.
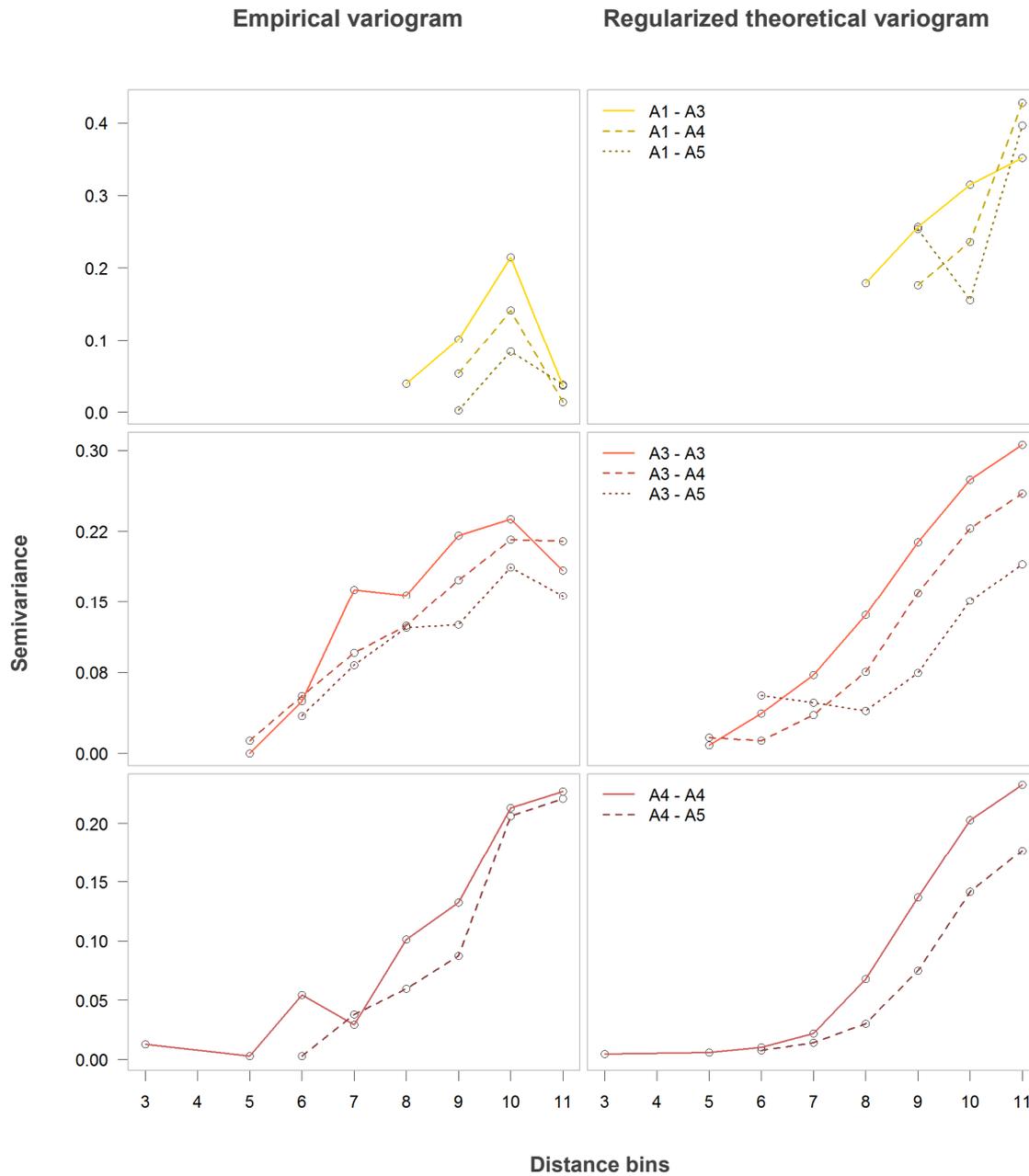
Figure 5.15: Empirical (left) and regularized theoretical variograms (right) for various areal and distance bins.

In order to achieve these different variograms, a theoretical point variogram needs to be selected, which can then be regularized for each areal setting. The selected variogram for this study is the multiplication of a modified exponential with a fractal model according to Skøien et

al. (2006), which has proven suitable in various previous studies (e.g. Laaha et al., 2013) and which proved superior over other tested variants. The model takes the form

$$\gamma_p(h) = 0.29\,h^{0.10}\left(1 - exp\left(-\frac{h}{27.44}\right)^{1.07}\right) + 1.05 \ .$$

(5.35)

11 distance and 5 areal bins have been selected. Regularization occurs for each combination of areal bins. Figure 5.15 shows a comparison of empirical (left) and regularized variogram values (right). In general, the fit of the regularized variogram is acceptable. The range of the empirical values is met best for the combinations of large areal bins, as seen in the lower panels. For medium sized bins, the regularized semivariance exceeds the observed one for greater distances. For combinations including small bins, as shown in the upper panels, overestimation by the regularized variogram is profound for all distance bins. Several scaling approaches on the input data have been tested to resolve this issue, but no satisfactory solution could be found. The issue is also well observable in Figure 5.16, where the theoretical variogram values are plotted against their empirical counterparts. Here, the respective numbers of point pairs per bin are indicated, showing a possible reason for the error: the population of the smallest areal bins is minor, which may cause distortion by measurement errors or similar effects.
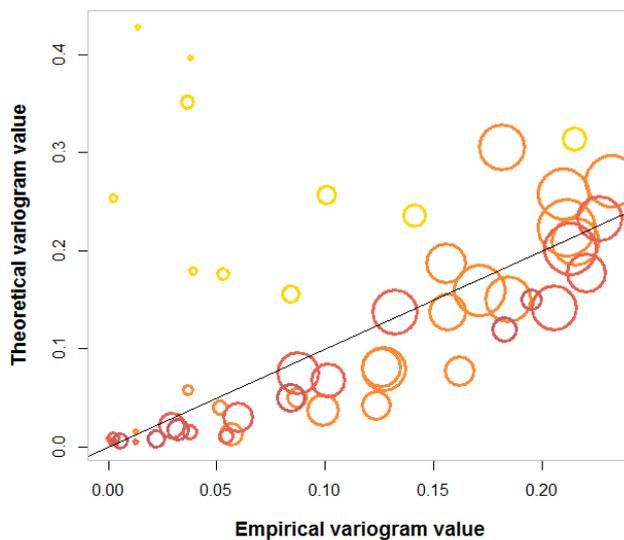
Figure 5.16: Scatterplot of regularized theoretical vs. empirical variogram values for areal and distance bins. Circle sizes indicate the number of point pairs per bin; the colors are analogous to Figure 5.15 and represent the size of the areal bins.

Despite the mentioned shortcoming the theoretical point variogram and its regularizations were used for kriging. Cross validation resulted in an NRMSE of 71.1 %, a percent bias of 2 %, an NSE of 0.49 and an R² of 0.50, and shows that TK outperforms OK. The values predicted during

LOOCV, as seen in Figure 5.17, show much higher spatial variability than the smoothed picture produced by OK. Also, the prediction variance naturally appears much more variable in space. As expected, variance is highest within the smaller catchments. For comparison: when all catchments smaller than 100 km² are removed before kriging, the cross validation NSE and R² increase to 0.64, while NRMSE and percent bias decrease to 60.1 % and 0.7 %, respectively. These obeservations are in accordance with those of e.g. Laaha et al. (2012) and Castiglioni et al. (2011), who showed that TK appeared to perform much better for larger rivers than for headwater catchments. The overall performance, however, is much higher in both studies.
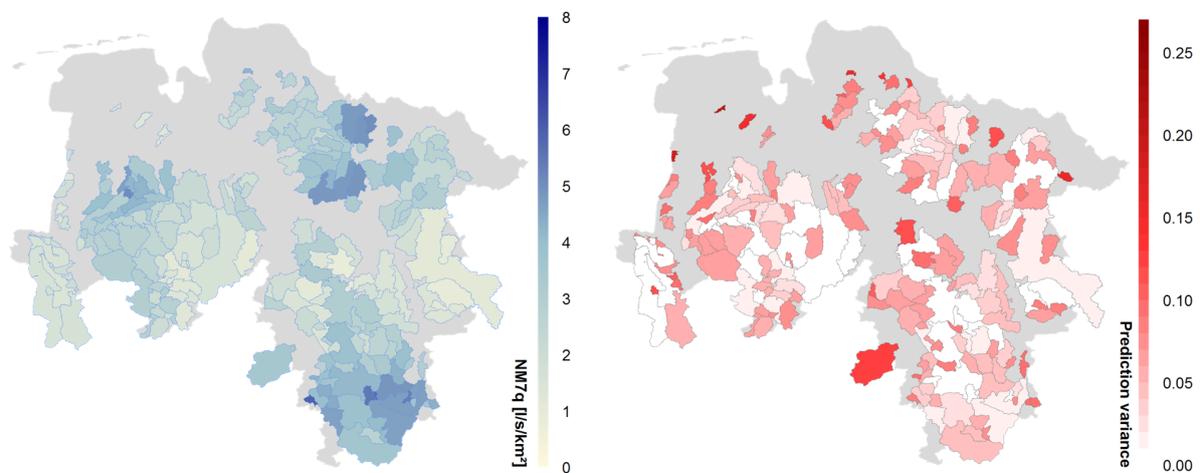
Figure 5.17: Regional mNM7q estimated during cross-validation using TK (left) and prediction variances (right).

The next step of the analysis aims at assessing the benefit of combining statistical and geostatistical approaches, i.e. the inclusion of external variables for regionalization. The first method tested is the PSBI. In order to build the first two principal components that serve as the physiographical coordinates, a range of combinations of different numbers of external variables have been tried. It turned out that a set of 6 variables proves ideal. A higher number of variables is naturally capable of explaining more of the variance in the target variable, but the portion of variance that is explained by the first two principal components only, does not appear to increase with increasing number of explanatory variables. The selected variables and their respective loadings and contribution to the individual components can be found in Table 5.3. For this setting the first two principle components make up 69.6 % of the entire variance. The resulting distribution of the mNM7q in the obtained physiographical space is shown in Figure 5.18.

The mapping in physiographic space yields a clear separation of mNM7q values. A gradient in the direction of the second principal component is observable. This trend needs to be considered before regionalization, as some kriging methods are prone to non-stationarity. Out of

the methods tested, the deterministic inverse distance weighting (IDW) with a power of 2 and consideration of all neighbors appeared to be most effective. The cross-validation yielded an NRMSE of 65.5 %, a percent bias of 0.0 %, an NSE of 0.56 and an R^2 of 0.56.

Table 5.3: List of selected variables and their respective loadings and contribution to the first two principal components.

Variable	PC 1		PC 2	
	Loading	Contribution	Loading	Contribution
Minimum $P_{90\ 6-9}$	0.07	0.19	0.85	43.16
Minimum $G_{mean\ 6-0}$	-0.33	4.3	-0.63	23.81
Maximum SPI $_{3-1}$	-0.59	13.93	0.61	21.81
Minimum $WSD_{mean\ 3-11}$	0.8	25.55	0.3	5.46
Minimum $P_{50\ 6-10}$	0.81	26.17	-0.29	5.16
Minimum $P_{60\ 6-10}$	0.86	29.85	0.1	0.6

Application of external drift Kriging did not appear promising, based on the previous findings. Having no observable correlation structure left after MLR fitting indicates that inclusion of information from neighboring stations will not improve the modeling results. As expected, EDK is capable of significantly improving the OK regionalization when meaningful external variables are added. However, even if all regressors identified using global MLR are included, the EDK model does not outperform MLR but rather shows poorer predictive power.
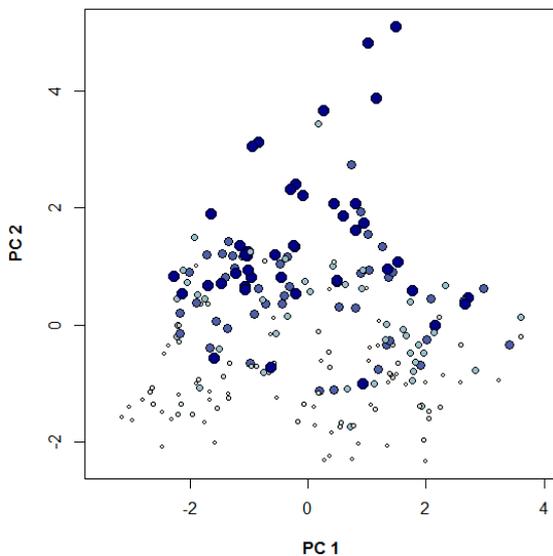
Figure 5.18: Distribution of the observed mNM7q in principal component space.

The shape of the fitted variogram is irrelevant and merely determines the range of the prediction variance, no matter if the variogram in (5.34) or a residual variogram is considered. The underlying variogram behind the results shown here is therefore the same as for OK. EDK results are only shown for a global kriging approach without any restrictions. Improvements are found for group wise application but still remain in the range of MLR.
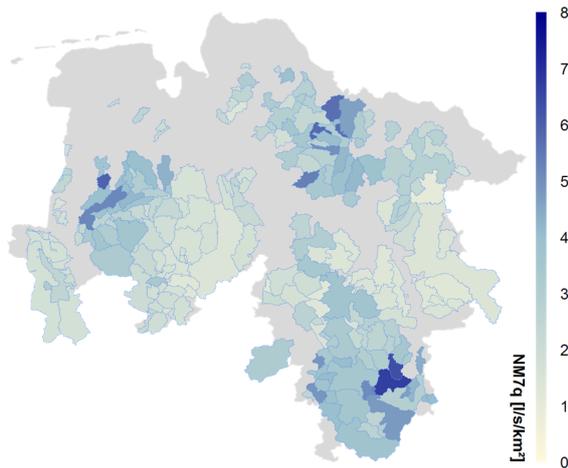
Figure 5.19: Regional mNM7q estimated during cross-validation using PSBI with inverse distance weighting.

Figure 5.20 shows the mNM7q estimated during LOOCV. The picture is almost identical to the global MLR model. Also the performance measures show similar values, i.e. an NRMSE of 62.2 %, a percent bias of -2.0 %, and an NSE and R² of 0.61. The prediction variance is smoothly distributed in space.
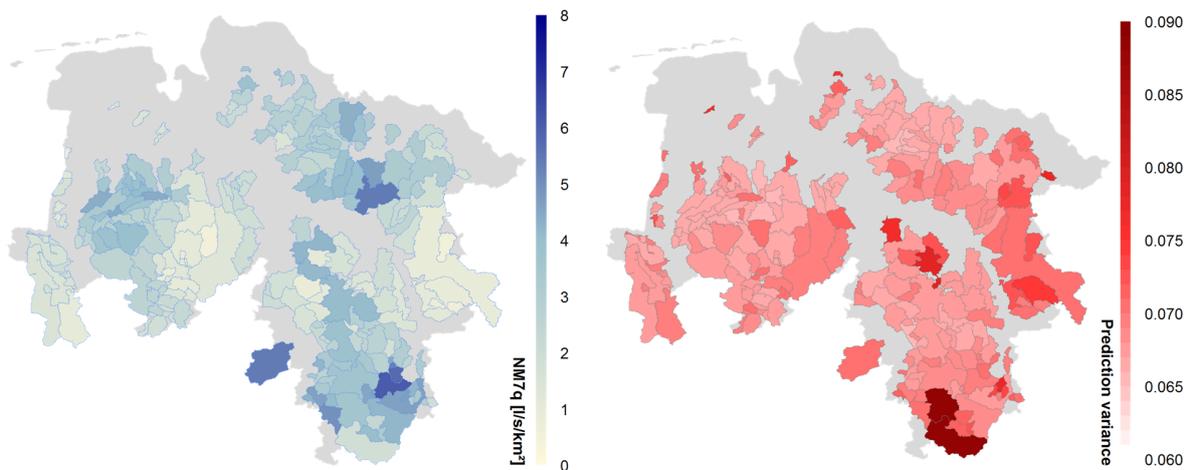
Figure 5.20: Regional mNM7q estimated during cross-validation using EDK (left) and prediction variances (right).

For the tested configuration, there is obviously no benefit in combining MLR with geostatistical methods, as the major portion of explainable variability in the target variable appears to be reproduced by MLR itself. In different configurations, the combination could be beneficial indeed. As shown before, there exists a trend for in-stream point pairs, i.e. nested catchments in the study area. Introducing external variables into TK or applying EDK exclusively for nested catchment pairs could potentially yield much better results than the simple MLR approach.

For comparison, Table 5.4 summarizes the cross-validation results for all analyzed methods for regionalization of the mNM7q.

Table 5.4: Comparison of cross-validation results for all five tested methods.

	NRMSE	pbias	NSE	R²
MLR global (grouped)	55.4 % (45.6 %)	-2.0 % (-0.8 %)	0.69 (0.79)	0.69 (0.79)
EDK	62.2 %	-2.0 %	0.61	0.61
PSBI	65.5 %	0.0 %	0.56	0.56
TK	71.1 %	2.0 %	0.49	0.50
OK	75.4 %	-2.7 %	0.44	0.44

Index-flood

So far, only the average NM7q has been considered for regionalization. Of greater interest, however, is the complete distribution of the NM7q at any point in space. Information about specific quantiles can be obtained by regionalizing respective values directly using one of the above methods, or, more conveniently, by application of the index flood method under input of regionalized means.

The index flood method requires homogeneity of flow in a region in order to be applicable. For formation of homogeneous regions in Lower Saxony, the set of meteorological indices in combination with the physiographic catchment descriptors have been used to build a regression tree for the classification of catchments. Figure 5.21 depicts the final tree, which assigns the available catchments into 15 groups. It can be seen that variables related to the maximum heat wave duration and dry and wet spell duration appear most relevant for classification.

The spatial distribution of the groups is shown in Figure 5.22. The heterogeneity of each region has been assessed using Hosking and Wallis' heterogeneity measure. The results are listed in Table 5.5. Blue values indicate definite homogeneity, purple values probable heterogeneity, and red values definite heterogeneity. The latter is the case for 1 of the 15 groups. Further separation of this group could not be achieved using the regression tree or any other approach. Discordant stations were neither removed nor resorted into different regions in order to leave classification to an objective algorithm applicable to unobserved locations.
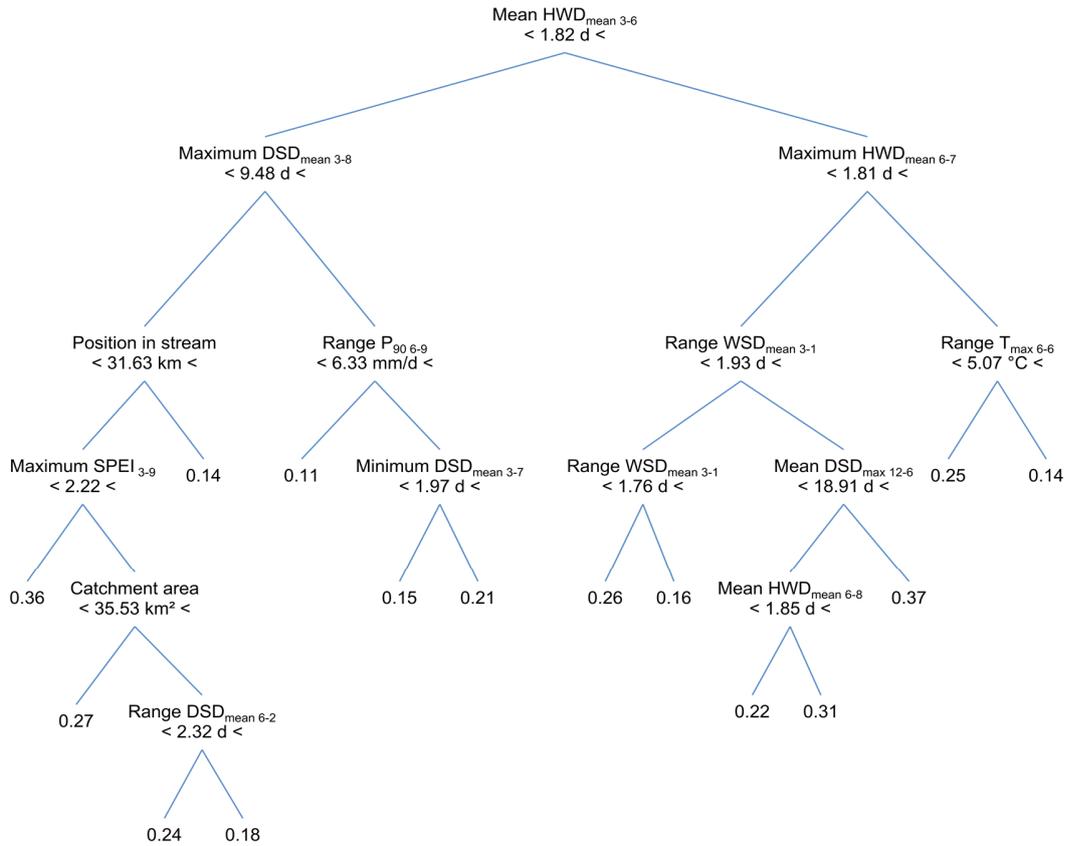
Figure 5.21: Regression tree for classification of catchments into homogeneous regions.

Table 5.5 additionally lists the regional distribution fitted for each region and the respective goodness of fit measure. In most cases, a GEV distribution was fitted; in 4 cases a general logistic distribution (GLO) showed best performance. Again, for the goodness-of-fit measure, blue entries indicate acceptable fit, red values poor fit. It needs to be mentioned that the GEV distribution did not always yield the best goodness-of-fit statistics but was fitted anyway out of mere convenience. Analyses showed that the effect of the type of distribution fitted was minor, if recognizable at all. Only if the goodness-of-fit indicated poor fit of the GEV, a different distribution was sought. For regions 13 and 14, the GLO showed lowest values in terms of goodness of fit, still, the fit was not acceptable. Region 14 is the largest of all regions and could potentially be separated further in order to improve the fit. Since for low flow only moderately low quantiles are of interest, the fit of the distribution in the extreme tails is not as important as for floods, for example. Despite poor goodness-of-fit measures, the fitted distributions still showed good fit for the more average values and lower tails. Thus, the regions and distribution functions were used as shown.
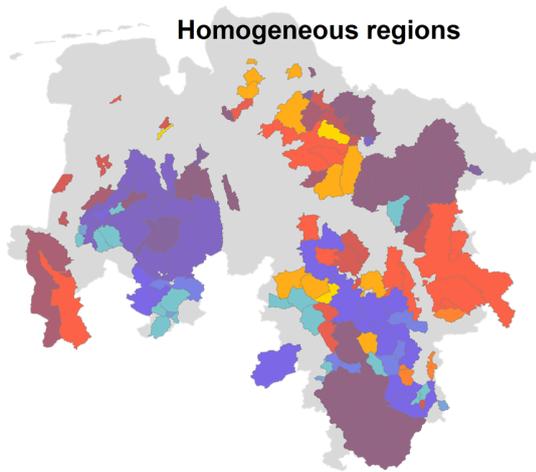
Figure 5.22: **H**omogeneous regions found via regression tree analysis for the index flood method.

As index values for local quantile estimation from regional frequency distributions the cross validation mNM7q predictions of the regional MLR have been selected, as the best estimates found in this study. The LOOCV results of the index flood method using these indices are summarized in Table 5.6 and compare the stations' locally observed with the regionally estimated quantiles. Shown are 3 quantiles with the return periods of 10, 20 and 50 years. In general, the performance of the index flood method is quite good, but appears to deteriorate with increasing return period. The only quality criterion that seems to improve is the bias, i.e. underestimation is higher for the NM7q10 than for the NM7q50. Since those values are overall negligibly small, this effect seems not important.

Table 5.5: Homogeneous regions with heterogeneity measure and goodness-of-fit statistic for the fitted distribution according Hosking and Wallis (1997).

Region	No. of members	H_1	Distribution	$|Z^{DIST}|$	Region	No. of members	H_1	Distribution	$|Z^{DIST}|$
1	8	0.35	GEV	0.51	9	2	1.78	GEV	0.42
2	14	-0.24	GEV	1.30	10	12	-1.44	GLO	0.62
3	5	-0.51	GEV	0.29	11	6	0.32	GEV	0.01
4	12	2.04	GEV	0.31	12	10	-0.38	GLO	0.95
5	10	0.15	GEV	0.68	13	5	0.02	GLO	3.69
6	4	-0.98	GEV	1.07	14	17	-0.23	GLO	2.27
7	5	-1.41	GEV	0.75	15	9	0.35	GEV	0.66
8	9	-0.24	GEV	0.63					

Figure 5.23 shows the regional deviation of the estimated quantiles during LOOCV from the locally observed quantiles and the corresponding error in the cross-validated index value. The

effect of error enhancement with increasing return period becomes visible for the individual catchments.

Table 5.6: Cross-validation results for three quantiles of the NM7q estimated using the index-flood method.

	NRMSE	pbias	NSE	R²
NM7q10	46.1 %	-3.8 %	0.79	0.79
NM7q20	48.4 %	-3.7 %	0.76	0.77
NM7q50	53.1 %	-3.1 %	0.72	0.72

In order to be able to assess the benefit of the index flood method over direct quantile regionalization via the above methods, the LOOCV of the estimation of the NM7q10 is compared to the best performing method, the regional MLR model. The model has been directly applied to the NM7q10 for the same catchments available for the index flood method. The cross-validation showed poorer performance with an NRMSE of 56.3 %, a percent bias of -4.9 %, an NSE of 0.68 and an R² of 0.69. It is thus assumed that the index flood method is superior to direct quantile regionalization using any of the analyzed approaches in this study.

Overall, low flow regionalization in Lower Saxony appears challenging, much more than in other regions of the world. Potential reasons behind the problematic are manifold. In comparison to other study areas, the northern German physiography is rather homogeneous. Apart from the small areal portion of the Harz mountains the region is consistently flat. Variation in soil and aquifer type, as well as climate is similarly low in comparison. Capturing differences in (low) flow governing processes, which is especially relevant for MLR approaches, is therefore challenging, especially in the light of anthropogenic interference, whose effects potentially exceed these factors. Human intervention is a major issue for the study area at hand. Pristine catchments are rare. The streams of the Harz mountains are dammed up, groundwater extraction for the intensive agriculture in the area is high, especially in the low-flow relevant drier periods, channeling and diversion of river water for economic reasons is affecting various rivers, etc. Even though enormous effort has been made to identify heavily and moderately influenced catchments, anthropogenic effects may still cover natural processes, distorting the natural spatial distribution of low flows, which heavily restricts modeling without special consideration of these very effects. Schnier and Cai (2014) found for catchments in the USA that inclusion of human factors, like population and groundwater use, are especially relevant for the regionalization of low flows, while they do not need to be considered for mean or flood flows.

Another problem may be related to the accuracy of the estimated means of the relatively short period of only 22 years. As shown, e.g. by Rodriguez-Iturbe (1969), 40 to 60 years of annual flow values should be available for robust estimation due to sampling errors. Using a data-set extension approach as done here may introduce even larger errors, when observed means are

biased. As shown by Westerberg et al. (2016), uncertainties due to gauging errors need to be evaluated before regionalization in order to not falsely attribute spatial differences in flow that result from measuring to catchment characteristics.
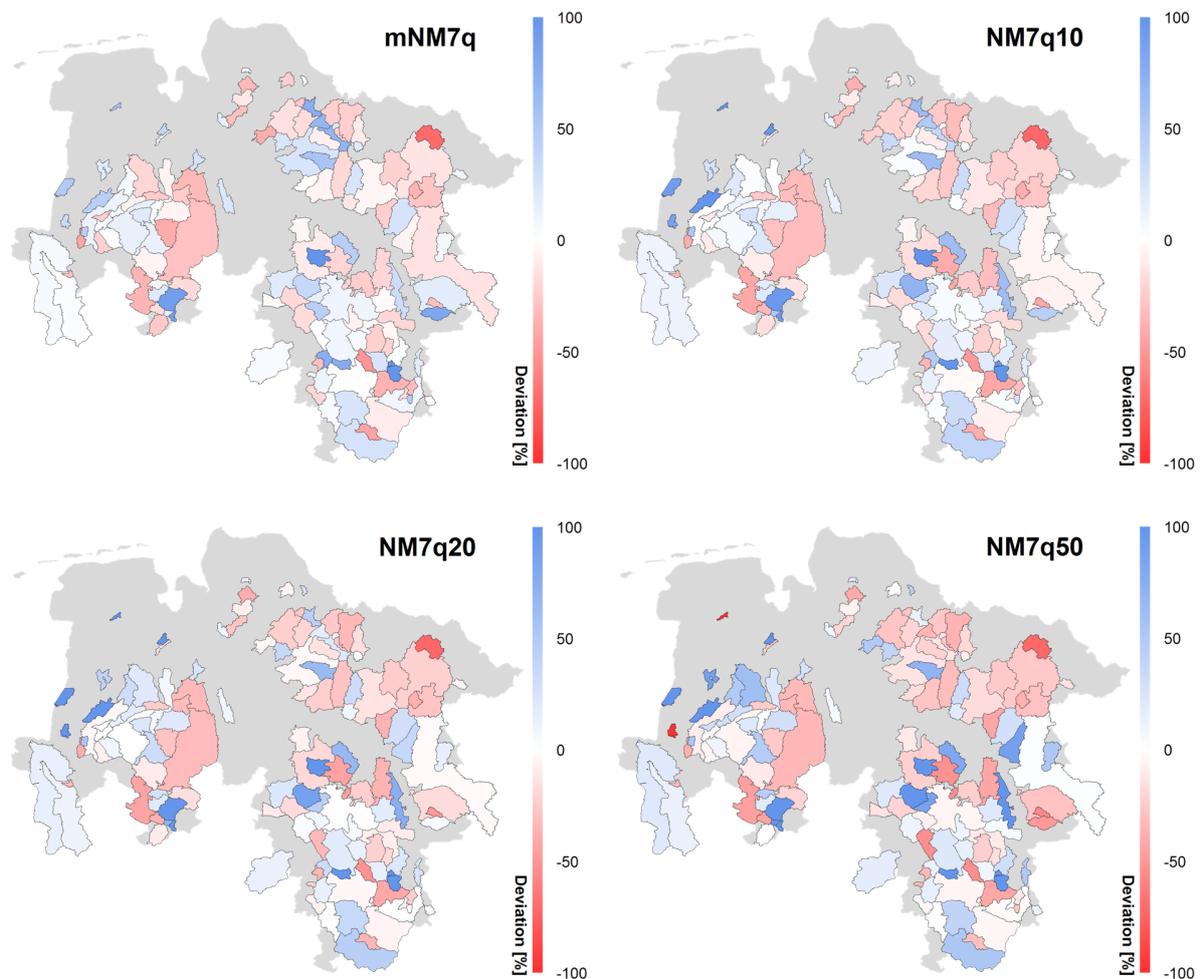
Figure 5.23: Percentage deviation of the regionalized from the observed mNM7q, as well as quantiles estimated with the index method from locally derived ones with 10-, 20- and 50-year return period.

Nevertheless, despite their comparably mediocre performance, MLR with meteorological indices as spatial explanatory variables and the ensuing index flood method appear to grant a method capable of reproduction of the low flow variance in space.

6   Spatiotemporal modeling of low flows

The third type of analysis deals with the combination of the temporal and spatial aspects of low flows within a single model, allowing a direct prognosis of future low flows in ungauged catchments. Three approaches will be evaluated: a) fitting a model to all available time series simultaneously for all stations (ST-1), b) fitting a model to flood index values obtained from the time series simultaneously for all stations (ST-2), and c) fitting a model to L-moments obtained from the time series simultaneously for all stations (ST-3).

6.1   Methods

6.1.1   Panel data regression

In order to incorporate both the temporal structure of the low flow indices at the stations, i.e. the dynamic component, and the heterogeneity between the stations into a single model, panel data regression has been selected as the underlying method for all three approaches named above. Panel data is a term originating from econometrics and usually describes data obtained from socio-economic surveys. It is structured in a specific way, i.e. data is simultaneously observed over time and over certain units, which is analogous to the data used for spatiotemporal modeling. The units are given by the different gauges, which are heterogenous in space, and whose values are observed over time. Thus, applying panel data regression appears meaningful for this type of analysis. The advantage of panel data regression over a repeated application of simple linear regression models over all stations is the significant increase in sample size and thus of degrees of freedom. Overfitting becomes less of an issue and coefficient estimates will be more robust.

Linear panel data regression extends simple regression by introducing a second "dimension" for which observations exist. Here the two dimensions would be space and time. The panel data regression model becomes

$$y_{it} = X_{it}\beta_{it}^{T}{}_{it} + u_{it} .$$

(6.1)

y represents an observation for individual i at time t, X is an observed vector of explanatory variables, observed for the same i and t, and u is the error component. Panel data regression offers a variety of ways to address so-called effects, i.e. differences in parameters over time and/or across units with the option to handle the error term as random or predictable component. Those alternatives will not be presented here, but can be found e.g. in Baltagi (2013). The simplest approach is to pool the data, assuming that parameters do not vary over time and across units.

Pooling has a major advantage for this type of analysis: prediction becomes straightforward, since modeling effects for unobserved locations is complex. Assuming constant parameters over the entire region facilitates prognosis at any point in the study area at any time. However, pooling restricts the model enormously.

Whether data is poolable and to which degree is tested using Chow's test (Chow, 1960). The test bases upon the assumption that a data set described via a multiple linear regression model can be split into several groups, which can each be described by an individual regression model with higher accuracy. For the test, the total regression model is compared with the individual regression models fitted to the individual groups. For n groups, k variables and a total of N observations the test statistic is

$$F_C = \frac{\dfrac{\left(SS - \sum_{i=1}^{n} SS_i\right)}{k(n-1)}}{\dfrac{\sum_{i=1}^{n} SS_i}{n(N-k)}} \, , \tag{6.2}$$

where SS and SS_i denote the total and the within-group sum of squared residuals, respectively. The test statistic is approximately distributed according to

$$F\left(k(n-1), nN - nk\right). \tag{6.3}$$

The Chow test requires normal error distribution.

Model fitting itself is done using feasible general least squares (FGLS) in order to achieve robustness against heteroscedasticity and serial correlation. Thereby, an OLS model is fitted first and the residuals are used to estimate the error covariance matrix. Panel data regression and belonging analyses are carried out using the "plm" package in R (Croissant et al., 2016).

6.1.2 Index flood method

As a second regionalization approach capable of considering temporal variability the index flood method described in section 5.1.2 is extended through addition of temporally variable predictors for index value estimation. A new set of homogeneous regions was formed for the new data set, which is substantially smaller, using once more a regression tree. Considering the same groups as before would result in several regions having only one observation, which would make LOOCV impossible.

The spatiotemporal index method approach is based on two major assumptions: a) low flow magnitude and hence low flow index values vary over time, resulting in changes of the flood index values at the stations and b) the shape of the frequency distribution remains completely unaffected by these changes.

6.1.3 L-moments

The third attempt at comprising spatial and temporal estimation of low flow indices within a single model is carried out via regionalization of temporally varying L-moments. This procedure is comparable to the index method, but instead of regionalizing a single index-value, several L-moments, calculated from the stations' samples are regionalized.

L-moments are used to describe the shape of a probability distribution. Any probability distribution can be described by moments of the underlying population, which in turn can be estimated from an available sample of this population. Of special relevance for inference of the distribution parameters are mean, standard deviation, skewness and kurtosis of the data. However, sample sizes are usually small, often too small to robustly estimate the population moments. Especially higher order moments like skewness and kurtosis may be heavily biased. Inference via L-moments instead of conventional moments provides a more robust way to estimate distribution parameters.

L-moments are derived from linear combinations of order statistics, explicitly as the expected value of these statistics multiplied by a scalar

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(X_{r-k:r}),$$

(6.4)

where r denotes the order of the L-moment, $X_{k:n}$ the kth smallest of n observations from the distribution of the random variable X, and E the expected value. L-moment ratios τ are obtained by dividing higher order L-moments by the second L-moment λ_2:

$$\tau_r = \frac{\lambda_r}{\lambda_2} \text{ for } r > 3.$$

(6.5)

The most relevant L-moments and L-moment ratios are the L-location λ_1, the L-scale λ_2, the L-coefficient of variation τ, the L-skewness τ_3 and the L-kurtosis τ_4.

Sample L-moments can be obtained using probability weighted moments. Estimation of a probability weighted moment of order r is achieved by summing over subsamples of length r from an ordered sample of length n

$$b_r = \frac{1}{n} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} x_{j:n}.$$

(6.6)

The sample L-moments can then be derived as linear combinations of different order probability weighted moments, namely

$$l_{r+1} = \sum_{k=0}^{r} p_{r,k}^* b_k \; , \tag{6.7}$$

where

$$p_{r,k}^* = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} . \tag{6.8}$$

In order to apply the sample L-moments for parameter estimation, relationships between L-moments and parameters of the desired distribution need to be established. The parameters of the GEV distribution, for example, whose function has been given in equation (4.30), are related to the sample L-moments as follows: the shape parameter κ cannot be solved explicitly but is approximated according to Hosking et al. (1985) via

$$\kappa \approx 7.8590 \, c + 2.9554 \, c^2$$
$$c = \frac{2}{3 + \tau_3} - \frac{\log 2}{\log 3} \qquad . \tag{6.9}$$

The other parameters can then be estimated using

$$\alpha = \frac{\lambda_2 \kappa}{\left(1 - 2^{-\kappa}\right) \Gamma\left(1 + \kappa\right)} \tag{6.10}$$

and

$$\xi = \lambda_1 - \frac{\alpha\left(1 - \Gamma\left(1 + \kappa\right)\right)}{\kappa} \tag{6.11}$$

with Γ as the gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \; . \tag{6.12}$$

The number of L-moments needed depends on the distribution function of choice. Regionalizing several L-moments instead of a single index value increases the flexibility and thus the adaptability of the selected regional frequency distribution to potential future changes beyond the mean, as it is the case for the index-flood approach. Nevertheless, the assumption that the type of frequency distribution does not change over time remains.

6.2 Data preparation

For the spatiotemporal analyses both record length and station density become relevant. The best possible compromise was found by selecting a record length of 40 years, allowing a

temporal separation into calibration and validation period. The period that is fully overlapping with the available climate record and shows the highest station density runs from 1966 to 2005. 51 gauges are available for this analysis.

Base data remains the set of annual low flow values, normalized by the belonging stations' catchment areas, while the annual meteorological indices serve again as potential explanatory variables. In order to be able to capture the spatial differences between the stations, the set of catchment characteristics, as well as the aggregated meteorological indices, used in section 4.2 are added to the data set.
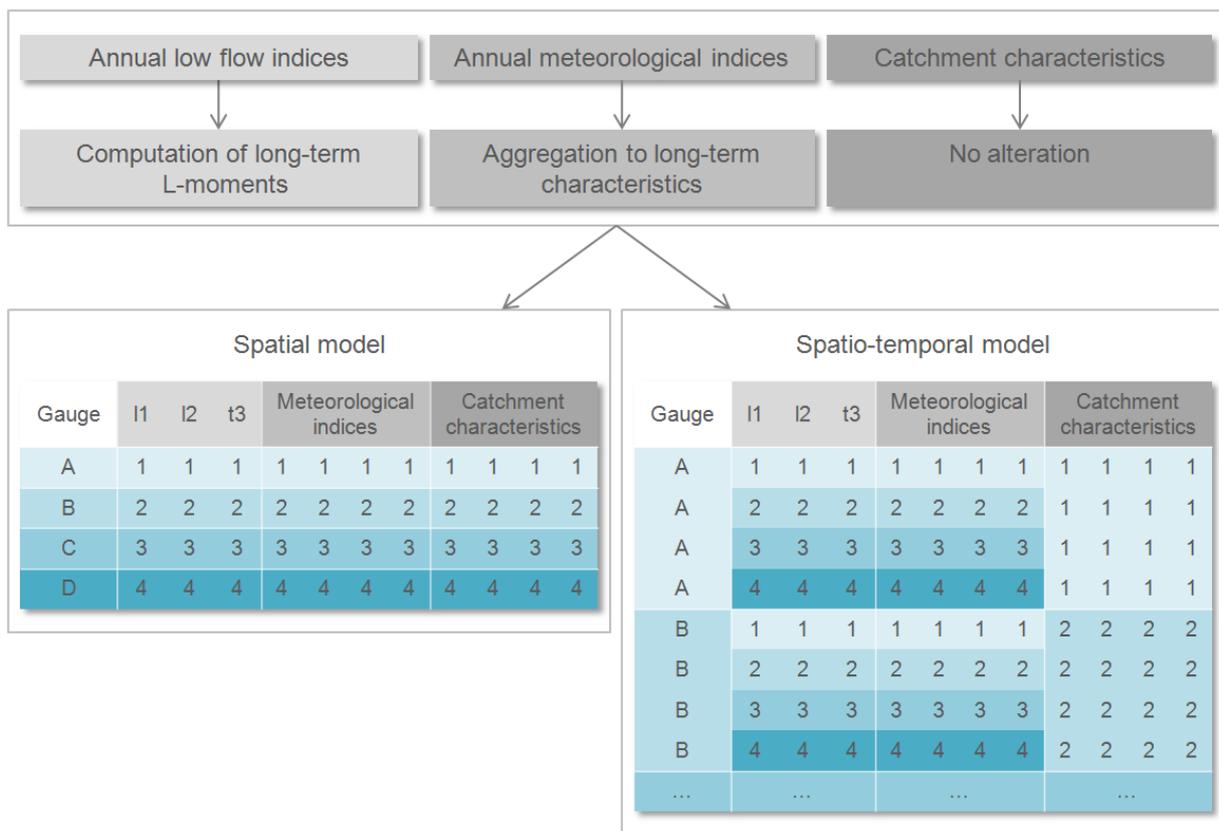
Figure 6.1: Data preparation for the fitting procedure of the index and the L-moment regional model approach

For version ST-1, i.e. simultaneous model fitting for the entire time series over all stations, the annual index values are subjected to model fitting without any alteration. Versions ST-2 and ST-3, i.e. spatiotemporal modeling of index values and L-moments, respectively, the data set needs to be modified. Instead of calculating the index values and L-moments for the entire available time series, they are repeatedly computed for bootstrap samples of 20 years. This enables the final model to not only account for the spatial variability between stations but also for the temporal variability at the individual stations, since more than one value per gauge will be

available for model fitting. Thus the setting still represents a panel data problem. For ST-3 three L-moments will be regionalized, i.e. l_1, l_2 and t_3. This allows later fitting of distributions with up to 3 parameters. L_1 simultaneously serves as the flood index value required for application of the ST-2 model.

The extension of the data set via bootstrapping is carried out as follows: Samples of individual years are drawn randomly from the low flow and the meteorological index time series. For each sample the low flow statistic of interest is calculated, be it the mean or L-moments and L-moment ratios, respectively. Accordingly, the meteorological indices are aggregated to form long-term characteristics for the exact same samples. Aggregation is accomplished by computing mean, minimum and maximum, and range of the annual meteorological index values. In order to capture the spatial structure, catchment characteristics are included that remain unaltered over the bootstrap samples. Figure 6.1 shows the procedure in detail and compares the mere regional approach with the spatiotemporal one.

The random sampling naturally results in equal or similar samples, which in turn yield equal or similar statistics. Including these multiples in the model fitting procedure will result in a preferable weighting towards these – most probably rather average – values. Thus, the individual bootstrap samples are tested against each other using non-parametric parameter tests. These are the Wilcoxon-Mann-Whitney test for homogeneity (Wilcoxon, 1945; Mann and Whitney, 1947) and the Levene test (Levene, 1960) for equality of variances between samples. The former test is carried out as follows: at first, ranks need to be assigned for all observations of both samples. The ranks thereby need to refer to the combined set of both samples. For both samples the ranks are summed up separately and a test statistic U can be computed for each sample according to

$$U_i = R_i - \frac{n_i(n_i + 1)}{2},$$
(6.13)

where R represents the sum of ranks, and n denotes the number of observations of sample i. The overall test statistic is obtained as the minimum value of U for both samples. The critical values are obtained from an approximated normal distribution of the test statistic, whose parameters are determined as

$$\mu_U = \frac{n_1 n_2}{2}$$
(6.14)

and

$$\sigma_U = \sqrt{\frac{n_1 n_2}{12}\left((n+1) - \frac{t_1^3 - t_1}{n(n-1)} - \frac{t_2^3 - t_2}{n(n-1)}\right)},$$
(6.15)

where t_i denotes the number of ties present within the respective sample.

Levene's test statistic for two samples 1 and 2 is computed as

$$L = \frac{n_1(\overline{X}_1 - \overline{X})^2 + n_2(\overline{X}_2 - \overline{X})^2}{\sum_{j=1}^{n_1}(\overline{Y}_{1j} - \overline{X}_1)^2 + \sum_{i=1}^{n_2}(\overline{Y}_{2j} - \overline{X}_2)^2} \qquad (6.16)$$

with the sample means $\overline{X}_i$, the mean over both samples $\overline{X}$ and

$$Y_{ij} = \left| X_{ij} - \overline{X}_i \right|, \qquad (6.17)$$

where X_{ij} denotes the j observations of sample i. The test statistic is approximately F distributed according to

$$F(1, n-2), \qquad (6.18)$$

with n as the total number of observations. Both the Wilcoxon-Mann-Whitney and the Levene test require independence of data.

Only those statistics that are significantly different from each other are subjected to the respective model fitting procedure. The number of accepted bootstrap samples per station is thus subject to the number of successful tests. In order to minimize the difference in numbers of observations between the stations, the initial bootstrap samples are drawn equally over all stations.

The testing nevertheless results in different bootstrap samples for the different L-moments that are used for model calibration. Since comparison of skewness between samples was not feasible, t_3 model fitting was tested using the two other samples consisting of significantly different means and variances. This procedure is not conform to the specified conditions but will prove irrelevant in any case.

6.3 Model fitting and evaluation of model performance

Variable selection will be carried out as done in the previous chapters. Since the sample size is large, especially for the first spatiotemporal modeling approach involving all observed time steps, penalty due to lacking degrees of freedom will not be given during the fitting process, depriving the variable selection algorithm of a distinct stopping point. Also, all selected regressors will most likely be highly significant. The appropriate number of variables is hence determined subjectively via the rate of convergence in the fitting criterion.

To facilitate an evaluation of both temporal and spatial quality of the model, validation is carried out in a combined split- and cross-validation procedure. Therefore, the selected 40 years of

record length for the analysis (1966 – 2005) are split into 30 years of calibration (1966 – 1995) and 10 years of validation (1996 – 2005). Variable selection occurs within the calibration period, as explained in section 4.1.2. Leave-one out cross-validation is then applied, not leaving out a single data point at a time, but an entire station's record. The whole time series or set of index values or L-moments for the left-out station will thus be estimated by a model fitted solely to the remaining stations within the calibration period, as depicted in Figure 6.2. Evaluation of the overall quality of fit, as well as the difference in performance between calibration and validation period give hence insight into regional and temporal model quality, respectively.
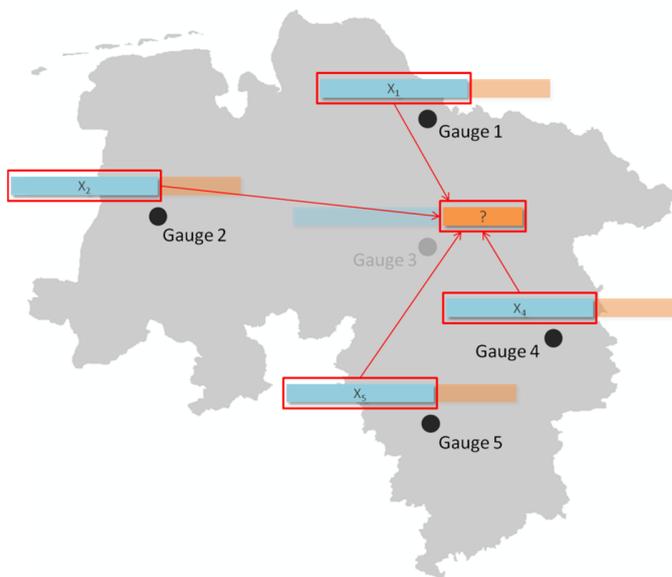
Figure 6.2: Validation strategy for performance assessment of the spatiotemporal models.

Again, the validation period is not considered for model selection. The best model is to be found merely on grounds of restrictions within the calibration process. In order to analyze whether pooling is a viable option for the spatiotemporal modeling via panel data regression, Chow's test is carried out in two ways: at first, the total regression model with equal intercept and slope parameters is tested against individual regression models for each station with completely variant regression coefficients. In a second step, a panel data regression model with spatially variable intercept but equal slope coefficients is tested against the independent models. In this way it can be determined if a model with fixed parameters over all stations, a model with variant intercepts or completely independent models yield the best performance. It should be noted that for the task at hand, i.e. simultaneous spatial and temporal prediction, not pooling the data is not an option, since parameters need to be valid for the entire region, including unobserved locations. Nevertheless, the outcome of the tests could suggest that group sizes could be reduced to achieve greater similarity between stations or that combined prediction in some areas may not pose the optimal option.

In order to assess whether spatiotemporal models are capable of reproducing future changes that are not observable in the past, the ST-3 L-moment model is exemplary subjected to further testing. In these tests, the entire 40-year period is considered for bootstrap sampling, individually for each L-moment. Case wise the maximal or minimal estimated value is left out and the model is fitted to the remaining values. Validation is then done via the left-out values. In order to assess whether variable changes in space can be captured by the model, the decision if a station's maximum or minimum is left out for validation is decided randomly. Selection of the distribution is carried out stationwise based on the observed annual low flow index values within the calibration period

6.4 Results and discussion

For the ST-1 pooled panel data regression model, a total of 10 variables have been selected, as listed in Table 6.1. The quality criteria for the calibration and validation period can be found in Table 6.2. Given are the total goodness-of-fit measures, obtained by simply comparing all estimated and observed values directly, as well as the regional and temporal quality criteria. These are obtained as the mean over all time steps' quality criteria and the mean over all stations' quality criteria, respectively. It can be seen that the overall goodness of fit is quite good with a calibration NSE of 0.71 and a validation NSE of 0.62.

Table 6.1: Estimated global pooled panel data regression model.

Coefficient	Estimate	Coefficient	Estimate
Intercept	*-6.15*	Range DSDWSD$_{max\ 3-1}$	0.27
P-ETP$_{mean\ 6-1}$	0.78	Minimum HWD$_{mean\ 6-8}$	4.42
Minimum SPI $_{3-12}$	-3.80	Minimum SPI $_{3-7}$	0.90
Maximum DSD$_{mean\ 3-4}$	-0.50	Range P$_{60\ 3-2}$	-0.65
Range WSD$_{mean\ 6-11}$	-0.99	Range P$_{50\ 3-7}$	0.56
Maximum WSD$_{max\ 6-1}$	0.16		

When distinguishing between temporal and spatial model quality, major discrepancies arise. The regional NSE shows values of 0.75 and 0.68 for calibration and validation period, while the temporal NSE yields values far below 0. When considering the other goodness-of-fit measures for the temporal case, it becomes obvious that the relative course of the annual NM7Qs at the individual stations appears to be reproduced quite successfully, as indicated by R² values of 0.57 and 0.54. The main problem is that the actual range of the values is not met sufficiently, which is also reflected by the high values of the pbias, which certify a significant overestimation of most stations' time series. This problem actually arises from poor regional estimation of the stations' means. The regional goodness-of-fit measures still indicate good performance due to the much larger spatial variability in contrast to the low temporal variability at the stations. Also,

as for temporal modeling in chapter 4, it is observable that the validation results are poorer than the ones for calibration.

Table 6.2: Cross validation results for the global ST-1 model measuring total, as well as regional and temporal performance for the calibration and validation period.

		NRMSE	pbias	NSE	R²
calibration	total	56.4 %	-1.4 %	0.71	0.71
	regional	49.4 %	-1.3 %	0.75	0.76
	temporal	98.6 %	5.0 %	-0.37	0.57
validation	total	62.1 %	6.3 %	0.62	0.65
	regional	55.7 %	6.3 %	0.68	0.72
	temporal	113.65 %	13.32 %	-0.80	0.54

Important for application of the pooled panel data regression model is the actual poolability of the time series at all stations, i.e. that the same regression parameters can be estimated for all stations in the data set. Table 6.3 gives an overview of the results of the Chow test for the tested model variants. Blue color indicates poolability with regard to both intercept and slope parameters, purple indicates poolability with regard to the slope coefficients only and red indicates no poolability, i.e. different intercepts and slopes should be considered for the individual stations. When looking at the global ST-1 model, poolability cannot be achieved by the selected model on any level, which is expected due to the heterogeneity in the study area. Thus, for all model variants, a regional approach is advised and the respective global models will not be discussed here. For regional analysis, the previously found k-means clusters are utilized.

Table 6.3: Poolability of data according to Chow's test ($\alpha = 5$ %).

	ST-1	ST-3 l_1	ST-3 l_2	ST-3 t_3
1				
2				
3				
4				
5				
global				

Poolable with same intercept and slope coefficients
Poolable with same slope coefficients
Not poolable

Poolability for the regional ST-1 models is achieved for the slope coefficients and three of the 5 regions. Regions 2 and 4 cannot be effectively pooled. Further separation of these groups is thus recommended but will not be carried out in this study. The results for the regional models are shown in Table 6.4. The final models are listed in Appendix D1, including all models described below. As expected, the goodness of fit for the individual models within the k-means

clusters is higher than for the overall global model. The NSE increases to 0.85 for calibration and 0.75 for validation. Figure 6.3 compares the cross-validation results for the global and regional models graphically via scatterplots. The right panel shows the regional model approach, which obviously performs significantly better than the global one. It can however be seen that some regions are modeled better than others. Especially the yellow dots, which represent group number 5, i.e. the mountainous Harz area, show extensive overestimation by the regression model.

Table 6.4: Cross validation results for the regional ST-1 models measuring total, as well as regional and temporal performance for the calibration and validation period.

		NRMSE	pbias	NSE	R²
calibration	total	39.0 %	-0.6 %	0.85	0.85
	regional	57.1 %	8.4 %	0.63	0.81
	temporal	100.4 %	6.5 %	-0.28	0.50
validation	total	50.0 %	10.2 %	0.75	0.83
	regional	51.5 %	10.1 %	0.69	0.84
	temporal	95.5 %	8.7 %	-0.18	0.61

The effect regarding the reproduction of temporal variability observed in the global model cannot be tackled by the regional models either. Temporal NSE values are still below 0 for both calibration and validation. The estimated index time series for the validation periods at the individual stations are shown in Appendix D2. It can be seen that the general course of the NM7q over the years is reproduced quite well by the regional models. The average coefficient of determination over all stations has a value of 0.61, indicating an acceptable reproduction of the internal temporal structure at the stations. R² performance ranges from 0.09 to up to 0.92.
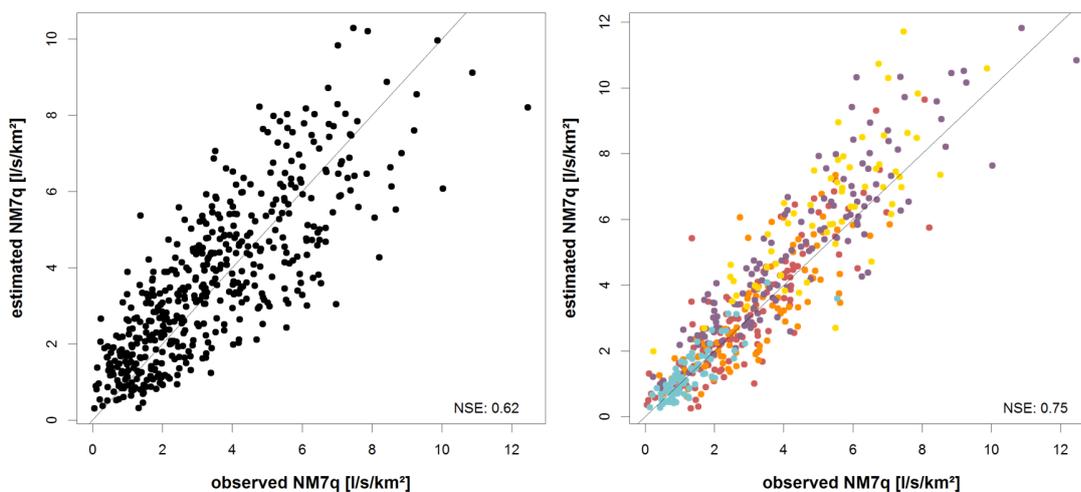
Figure 6.3: Scatterplots of observed vs. estimated annual NM7q values for the global ST-1 model (left) and the regional ST-1 models within the k-means clusters (right).

For the majority of stations, the actual range of the observed values is also met, for some stations however, significant under- or overestimations can be observed, resulting from the model not being capable of reproducing the variation in space. The average NSE of -0.18 for at-station temporal model performance indicates that the distance between observed and estimated values is the major issue for simultaneous spatial and temporal model fitting. Out of the 51 stations 17 show an NSE below 0. The spatiotemporal models hence appear to be suitable to predict the relative course of the low flow indices over time but not always the actual absolute values, as indicated before by the mere poolability of slope coefficients. The model therefore appears rather suitable for assessment of regional climate change signals than for exact prognosis of future low flow values at individual points in space and time.
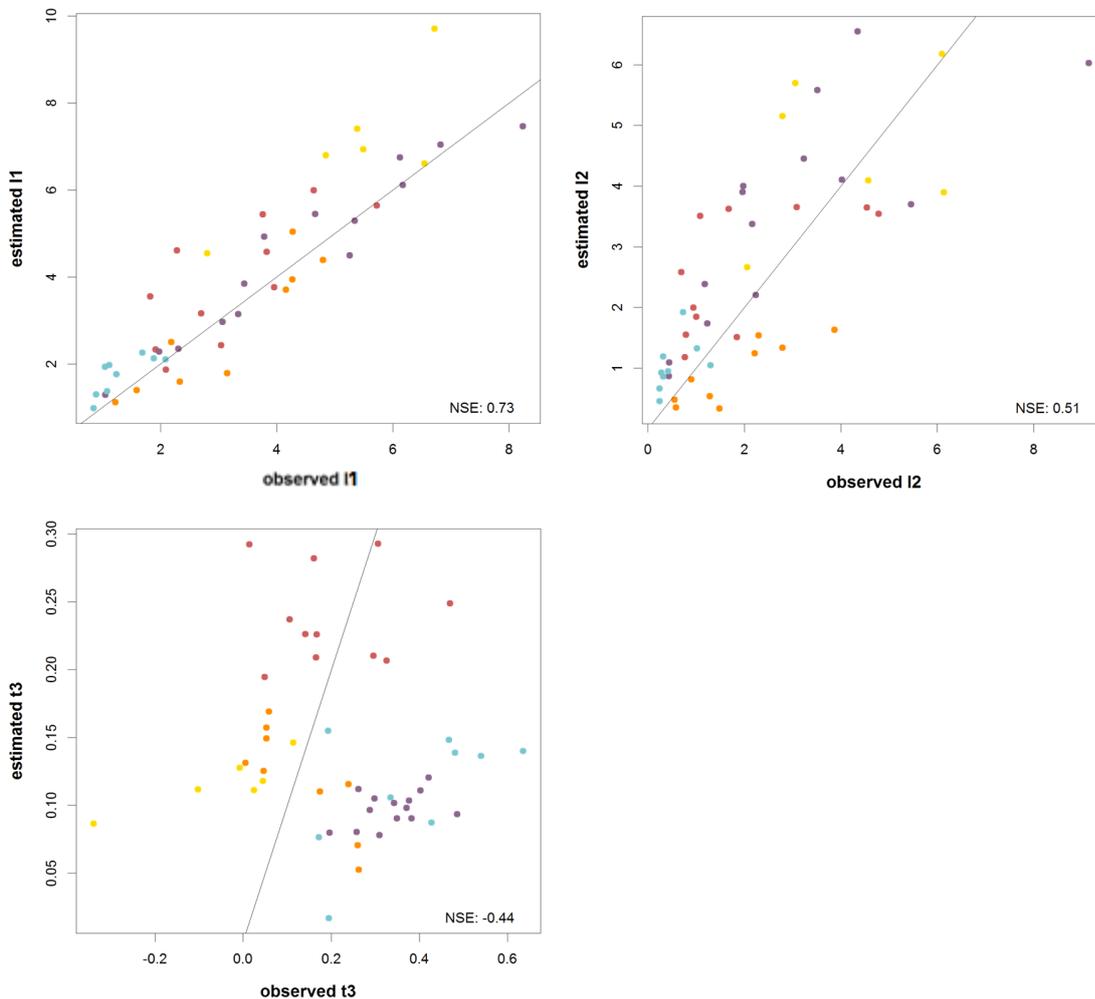
Figure 6.4: Scatterplots of observed vs. estimated values for the three L-moments modeled with the regional ST-3 model.

The ST-2 and ST-3 models require regionalization of L-moments. The first L-moment, which is utilized by both methods, appears to be readily modeled in space and time, as seen in Figure 6.4 (topleft). The NSE for the combined cross and split validation is 0.73. The other L-moments cannot be modeled as effectively. For l_2 an NSE of merely 0.51 can be obtained with severe overestimation for most of the regions. T_3 cannot be reproduced at all by the panel data regression model. Even though the Chow test certifies perfect poolability for two of the regions, as seen in Table 6.3, the modeling attempt is discarded. The enormous error that arises during regionalization (overall NSE of -0.44) could induce major distortion of the fitted distributions. Instead, for regional distribution fitting, the respective average t_3 of a homogeneous region is used.

Table 6.5: Homogeneous regions with heterogeneity measure and goodness-of-fit statistic for the fitted distribution after Hosking and Wallis (1997).

Region	No. of members	H_1	Distribution	$|Z^{DIST}|$
1	8	-1.51	GEV	0.41
2	8	0.03	GEV	0.64
3	6	-1.04	GEV	1.02
4	6	0.55	GEV	1.43
5	7	2.70	GEV	0.34
6	4	2.56	GEV	3.97
7	6	-0.81	GEV	0.96
8	6	1.11	GEV	0.72

For application of the ST-2 method, homogeneous regions, analogous to the ones in chapter 5.4 need to be found for the reduced data set. Again, a regression tree approach is applied, whose final tree can be found in Appendix D3. 8 homogeneous regions have been identified, as listed in Table 6.5. Two regions emerged as heterogeneous but were not treated further, due to the same reasons as in chapter 5. The GEV distribution appeared generally suitable for all regions, except for region 6. However, no distribution could be identified that would yield a better fit. Again, the fit for the lower tails was generally met, even if the overall goodness-of-fit suggested otherwise.

Table 6.6: Goodness-of-fit criteria for the NM7q10 estimated via regional ST-1, ST-2 and ST-3 models for the calibration and validation period.

	ST-1			ST-2			ST-3		
	pbias	NSE	R²	pbias	NSE	R²	pbias	NSE	R²
Calibration	7.6 %	0.88	0.89	-0.5 %	0.92	0.92	0.6 %	0.90	0.91
Validation	9.7 %	0.84	0.90	5.4 %	0.78	0.84	14.6 %	0.63	0.8

Accordingly, the GEV distribution was selected for all stations in the ST-3 modeling approach, fitted using the regionalized L-moments. In order to assess the quality of the ST-2 and ST-3 approaches, individual quantiles derived from the fitted distributions are compared to locally derived quantiles from the observation. The validation results are additionally compared to the ST-1 approach for which quantiles are extracted from the predicted time series. Table 6.6 shows the results for all three model types by means of the NM7Q10 for both calibration and validation, while Figure 6.5 shows the belonging scatterplots for the validation period. Highest performance during validation has been achieved by the ST-1 model, which performes slightly better than the ST-2 approach. The scatterplots show that the ST-1 error is quite uniform for the entire study area, while the ST-2 error is highest for the Harz region and larger values in general. Smaller values are reproduced much better than with method ST-1. ST-3 shows the overall weakest performance. The validation NSE of 0.63 is substantially lower than for the other methods. The percent bias of 14.6 % and the scatterplot indicate a major overestimation, especially for the Harz region. Also, for ST-3 the difference in performance between calibration and validation is largest, which indicates poor generality of the fitted models. The difference can also be noted for the ST-2 model, though weaker, but is almost unnoticeable for the ST-1 model. In contrast to the mere temporal models in chapter 4, the ST-1 model appears to reduce issues due to potential non-stationary relationships between target and explanatory variables and overfitting or omittance of relevant variables.

Figure 6.6 depicts the error distribution over all stations made with the individual models. The upper left panel show the errors for the spatiotemporal modeling of the L-moments. It is clearly observable that the error is smallest for the l-location and increases significantly toward the l-skewness. The remaining panels show the ME and MAE for the three ST models over three quantiles, i.e. the NM7Q10, the NM7Q20 and the NM7Q50. For all models it becomes obvious, that overall performance decreases with increasing return period. This effect is comparably small for ST-1 and ST-2 but quite significant for ST-3. While ST-1 shows highest performance for prediction of the NM7Q10, ST-2 appears slightly superior for prediction of the lower quantiles. The median MAE for the NM7Q50 compare as 27.46 % to 28.8 % for ST-2 and ST-1, respectively.

Both ST-1 and ST-2 are thus considered to have comparable predictive power for the observation period at hand. ST-1 could however prove advantageous for prediction of the far future. By modeling the entire set of annual low flow index values, any distribution could be fitted to the predicted data, depending on maximum goodness-of-fit. The ST-2 model relies on a stationary distribution within a homogeneous region over time, even for the far future. ST-1 thus is capable of transferring complete information about the low flow distribution, while ST-2 is only capable of estimating changes in the location. Furthermore, the stationary distribution of the homogeneous regions that are required for ST-2 is questionable. A regression tree analysis carried out for the validation period gave a slightly different configuration than for the calibration, even though the periods are directly adjacent. ST-3 would pose an intermediate way between

the two methods by still assuming the same type of distribution for the future but by simultaneously making it more flexible to adapt to future conditions via estimation of the complete parameter set rather than just the location. However, the error that arises from modeling of the parameters other than the mean is too high to make precise predictions. Assuming those parameters constant would yield better prognoses. There is consequently no benefit of an L-moment regionalization over the index method and especially over direct modeling of annual index values in space and time.
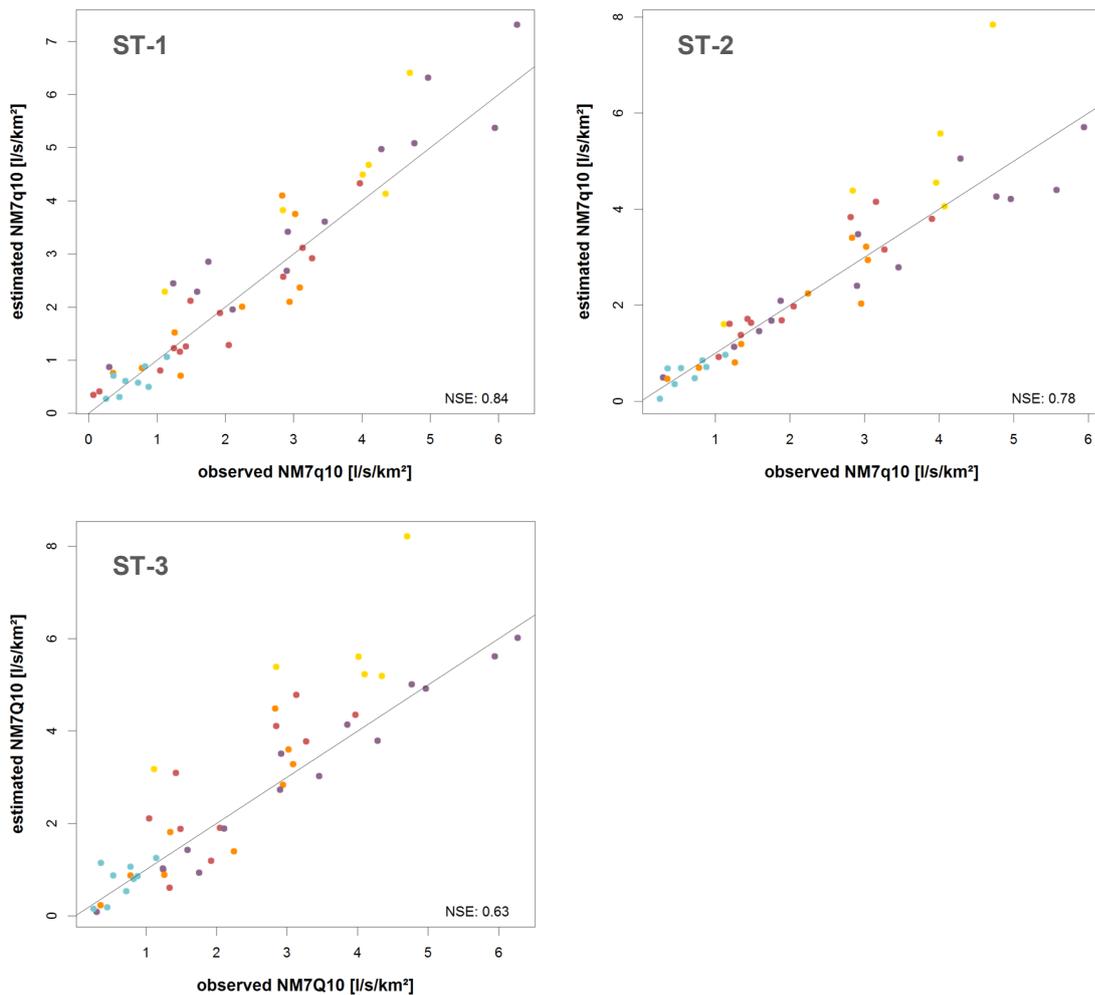
Figure 6.5: Scatterplots of the observed vs. predicted NM7Q10 for the three spatiotemporal methods in the validation period.

An issue that arises from the small record lengths is that the validation period used here is not significantly different from the calibration period, due to its direct adjacency. All models are definitely capable of reproducing temporal changes in the variables but the cross validation profits from the similarity of both periods. How well the models perform for states outside of this

range definitely needs to be tested using climate model data, as future values may be significantly different from the observed ones. A test that could be made given the 40 years of observation is the creation of artificial validation sets that include values, which are not included in the calibration data set. This analysis has been carried out on the spatiotemporal modeling of the first L-moment. For validation, all maxima, all minima, and randomly selected minima or maxima have been stepwise removed from the calibration set.
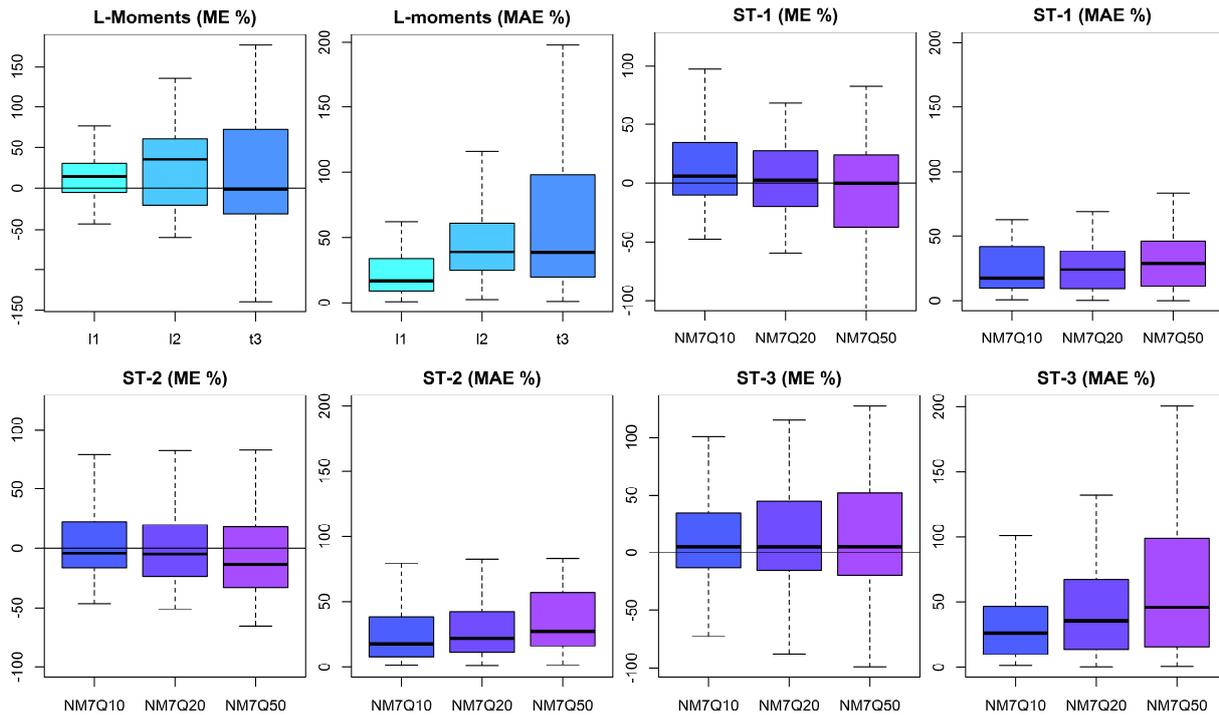
Figure 6.6: Mean and absolute mean error over all 51 stations for the validation period for the individual L-moments (top left) and specific quantiles of the NM7q estimated with the three spatiotemporal model variants.

Table 6.7 gives an overview of the validation results for the three validation types, while Figure 6.7 shows the respective scatterplots between observation and estimation and compares each scenario to the stationary case, i.e. using the mean of the calibration set as predictor. It can be seen that the quality criteria for all three scenarios certify a remarkable performance. The pbias indicates that the minima are slightly overestimated, while the maxima are rather underestimated. Since both extremes represent values that are significantly lower or respectively higher that the values in the calibration set, as substantiated by the Wilcoxon test, this outcome is more moderate than expected. The fact that the random mixed removal of maxima and minima over the study area yields an even higher performance proves that the spatiotemporal models are capable of reproducing the individual stations' development based on its meteorological indices independent from the development of the remaining stations in the area.

When looking at the scatterplots of the stationary scenarios, it becomes obvious that the changes in the validation period cannot be captured, as expected. One should note that the scatterplots shown for the ST-models represent the predictions made during combined split and cross validation, i.e. each station's prognosis has been made without the actual station in the calibration set and thus represent a combination of temporal and spatial modelling. The stationary values on the other hand are simply temporally transferred from calibration to validation for each station, which is the reason for the dispersion being much lower than for the ST models. This transfer is of course not possible for unobserved gauges. Consequently, the assumption of stationarity is not an option. If climate change impacts are to be assessed, modeling is essential.
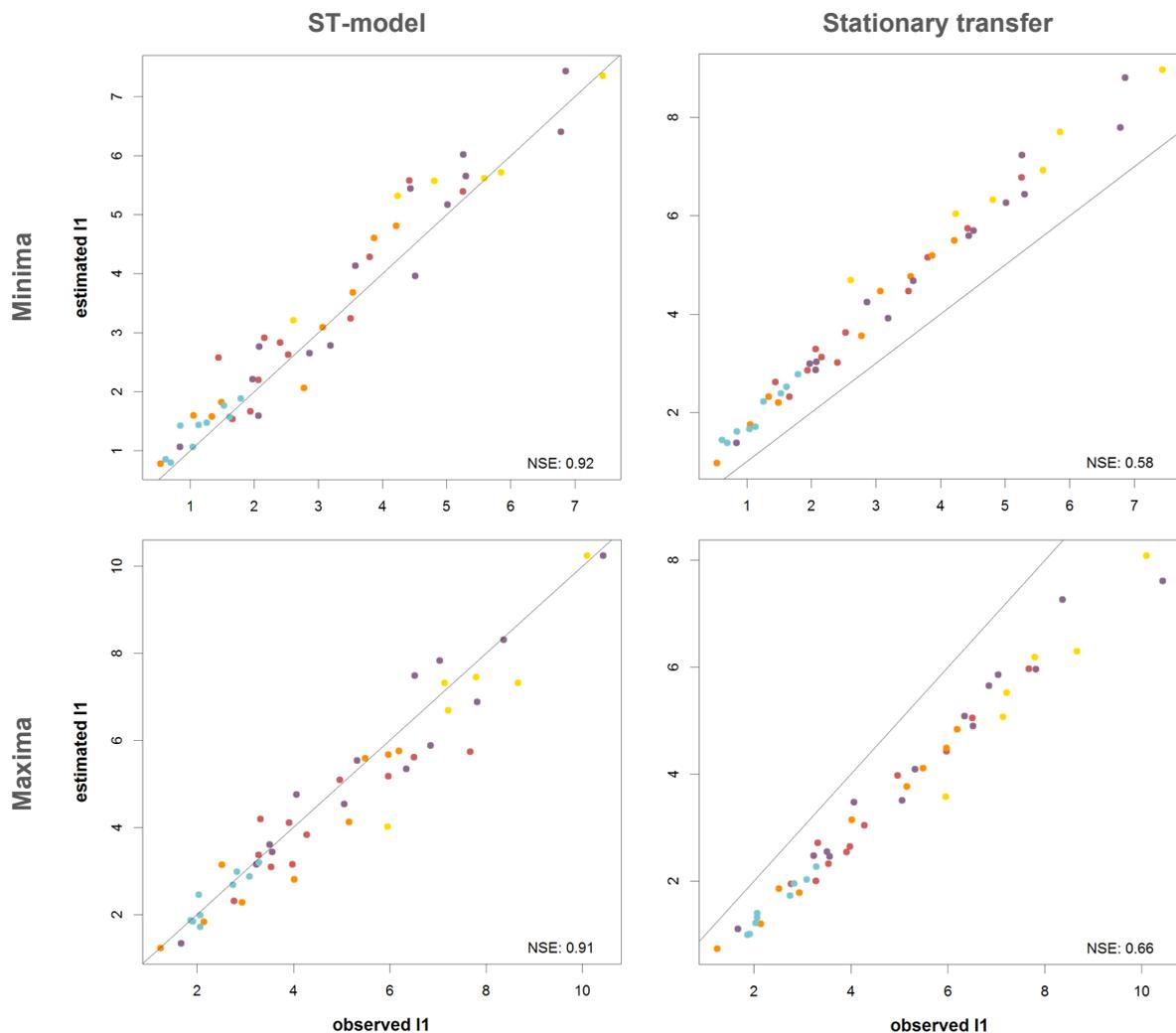
Figure 6.7: Scatterplots of observed vs. estimated l_1 in the validation period

Spatiotemporal modeling appears to be a straightforward approach for comprehensive regional climate change impact assessment. Simultaneous spatial and temporal prediction based on simple statistical assumptions is a major advantage over other procedures. If the combined approach is capable to even outperform the regionalization of prognoses made at individual stations will be tested in the following chapter.

Table 6.7: Cross-validation results for estimation of l_1 using three different validation scenarios.

	NRMSE	pbias	NSE	R²
Minimum	27.5 %	8.2 %	0.92	0.95
Maximum	29.3 %	-5.7 %	0.91	0.93
Mixed	19.6 %	0.20	0.96	0.96

The other indices have not been analyzed for this type of analysis. It is expected that the q95 performs similar to the NM7q. Since all other indices could only be modeled successfully in either space or time, the outcome of spatiotemporal modeling is expected to be significantly inferior to the results shown here.

7 Comparison of modeling approaches

In order to assess which approach is most suitable for regional estimation of future low flow indices, the approaches analyzed in chapters 4, 5 and 6 are combined and compared. The aim is to investigate whether it is more effective to first temporally estimate future low flows at stations with long record lengths according to chapter 4 and then interpolate the flow using the methods of chapter 5, or if simultaneous temporal and spatial modeling, as done in chapter 6 is advantageous.

Before the actual comparison there is a range of questions that need to be answered in this chapter, primarily regarding the regionalization of temporally estimated low flows. The major issue is the reduced station density. In chapter 5, regionalization has been applied to data sets of 221 stations. Since temporal model fitting can only be achieved at long record stations, the number of available stations will decrease drastically. The problem thereby lies within the validity of the regional models, especially the statistical ones, like MLR. The fitted model is explicitly valid for the data it has been fitted to. Consequently, the model found for the 1988-2009 period will not be applicable to the set of future data. A new model needs to be fitted, given the restriction of low station density. Pure distance based models, like OK and TK may be transferrable from past to future under the premise that the relative regional distribution of the low flow and thus the spatial correlation structure does not change over time. But even if this were the case, regional estimation would not be as successful with fewer direct neighbors in the data set.

In case of the regional MLR model it is to be tested whether it should be made use of the temporal variability of the meteorological catchment descriptors used for prediction or if it is more robust to assume stationary external variables, i.e. the data set of the past, and adapt the regression coefficient estimates only.

7.1 Data preparation and model fitting

The data set selected for the analyses is the one used for spatiotemporal modeling in chapter 6, posing the compromise between station density and record length.

The time series are divided into a pseudo past period and a pseudo future period. The past is used for model calibration, the future period for model validation, as before. As for the spatiotemporal model approaches, the period from 1966 to 1995 is selected as the "observed past" and the period from 1996 to 2005 serves as the "observed future".

A "simulated future" is obtained by application of local MLR models to the available stations, as described in chapter 4, calibrated on the "observed past".

On this data the following steps will be carried out for analysis:

a) Interpolation of the "observed past" data set in order to assess the effect of reduced station density in comparison to chapter 5.

b) Interpolation of the "observed future" data set in order to assess changes in the spatial structure between the periods.

c) Interpolation of the "simulated future" data set in order to assess the effects of uncertainty.

d) Comparison of the best regionalization technique for the "simulated future" with the spatiotemporal models ST-1 and ST-2.

In the first part of the analysis, mean low flow values are computed for the calibration period at each station. These values are regionalized using regional MLR (section 5.1.1) in order to evaluate the explicit performance of the regional model via cross-validation.

The same procedure is carried out for the "observed future". Regionalization of the station values is once more achieved using regional MLR. Several variants for model fitting are thereby tested. At first, the regional model is fitted to the observed future mNM7q using future CDs. The second model is fitted to the observed future mNM7q but with past CDs as regressors. The direct comparison between the two approaches may give an indication whether non-stationary CDs are required for regionalization of future data. It is also tested if the regional model found for the past may be directly applicable to the future data set by applying the past model to the set of future CDs. Finally, the error of a simple transfer of the past model to the future data set is evaluated.

In the next step, temporal MLR is applied using principal components (section 4.1.2.2), which appeared to be the best temporal modeling approach. For the "simulated future" mean low flow values are computed and a regional MLR model is fitted once more. Comparison of observed and temporally estimated means gives insight into the temporal model quality, while LOOCV comparison of regionalized with temporally estimated means shows the error of the spatial model in the future period. The overall error is given by comparison of future regionalized means with future observed means. This last error can be directly compared to the cross-validation results of the spatiotemporal models.

The steps a) to c) will be carried out using different model configurations. These come about through modification of two parameters: the set of catchment descriptors and the target variable used for model fitting. The need to analyze the effect of the utilized set of catchment descriptors arises once more from the underlying assumptions of this study. Considering non-stationarity in the low flow variable would simultaneously call for non-stationary CDs for regionalization. It may, however, be sufficient to use a static set of CDs and simply change model settings to match the future low flow. Hence, CDs and target variable are alternatively obtained from either the past or the future period. In total, 9 experiments arise from the different settings, as listed in Table 7.1.

The various comparisons between observed and simulated, local and regionalized data allow for an additional assessment of model quality of the individual modeling steps.

Data set extension according to section 5.2. is not carried out. Due to the large number of experiments only global models without grouping are shown for steps a) to c) for convenience. The final comparison between model approaches in step d) will be carried out for the k-means clusters found in chapter 5.

7.2 Results and discussion

Table 7.1 gives an overview over all tested modeling variants using different target variables and sets of CDs. Given are the period of the regionalized mNM7q, the type of variables that are compared during cross validation, i.e. station data (stat) and regionalized data (reg), and observed (obs) and simulated data (sim), respectively, the set of CDs that served as regressors for the regional MLR model, as well as the period of the target variable it has been fitted to, and finally the cross validation results.

Table 7.1: Comparison of variants for regionalization of observed and simulated mNM7q.

Case	Period	Comparison	Catchment descriptors	Target variable	Validation			
					NRMSE	pbias	NSE	R²
a.1	1966-1995	obs_{stat} vs. obs_{reg}	past	past	53.1 %	-1.5 %	0.71	0.72
b.1	1996-2005	obs_{stat} vs. obs_{reg}	future	future	53.0 %	-1.5 %	0.72	0.72
b.2		obs_{stat} vs. obs_{reg}	past	future	54.7 %	-1.9 %	0.70	0.70
b.3		obs_{stat} vs. obs_{reg}	future	past	102.1 %	-6.1 %	-0.06	0.02
b.4		obs_{stat} vs. obs_{reg}	past	past	56.4 %	-1.7 %	0.68	0.68
c.1		obs_{stat} vs. sim_{stat}	-	-	27.8 %	8.2 %	0.92	0.96
c.2		sim_{stat} vs. sim_{reg}	past	future	56.2 %	-2.2 %	0.68	0.68
c.3		sim_{stat} vs. sim_{reg}	future	future	57.2 %	-1.7 %	0.67	0.67
c.4		obs_{stat} vs. sim_{reg}	past	future	68.4 %	4.5 %	0.57	0.57

obs = observed mNM7q

sim = simulated mNM7q (via temporal station models)

stat = station data

reg = regionalized data

Before carrying out any regionalization the "observed past" and the "observed future" mNM7q values are compared. Expectedly, as seen in Figure 7.1, the difference between the observations in the "past" and "future" are very similar, due to the direct chronological adjacency and the small length of the validation period. Nevertheless, the overall NM7q is slightly lower for the "future" than for the "past".
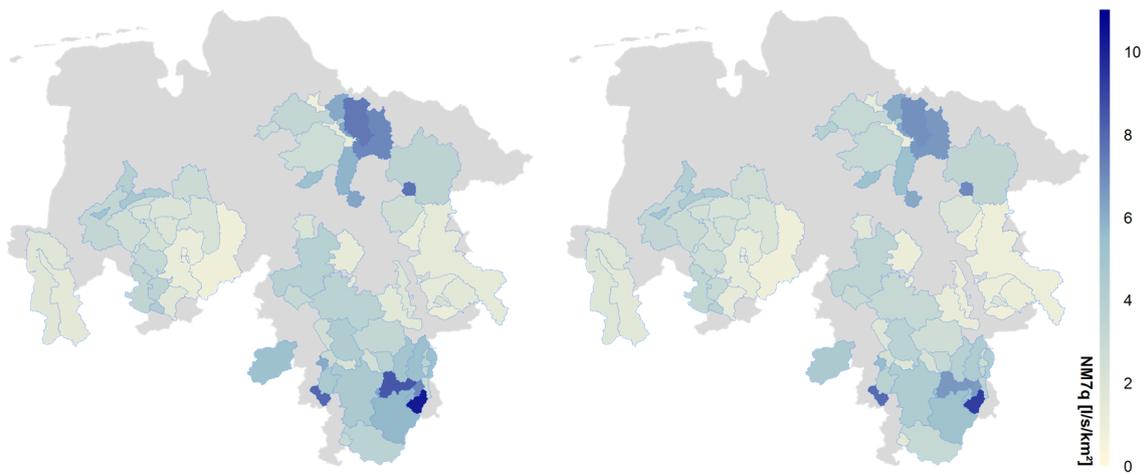
Figure 7.1: Observed regional mNM7q for the periods 1966 – 1995 (left) and 1996 – 2005 (right).

The first comparisons are made between local observations and regionalization by using individual models and respective climatic catchment descriptors for the two periods. The goodness-of-fit criteria, summarized in Table 7.1 (cases a.1 and b.1) are almost equal for cross-validation and indicate successful regionalization for both the past and the future observed values using past and future observed CDs, respectively. Also the spatial distribution of the predicted values, as depicted in Figure 7.2, is in both cases similar to the respective observation. The scatterplots in Figure 7.3, showing the comparison between observed and cross-validation estimations, indicate that in the future period several values are strongly under- or overestimated, which are reproduced better in the past period.
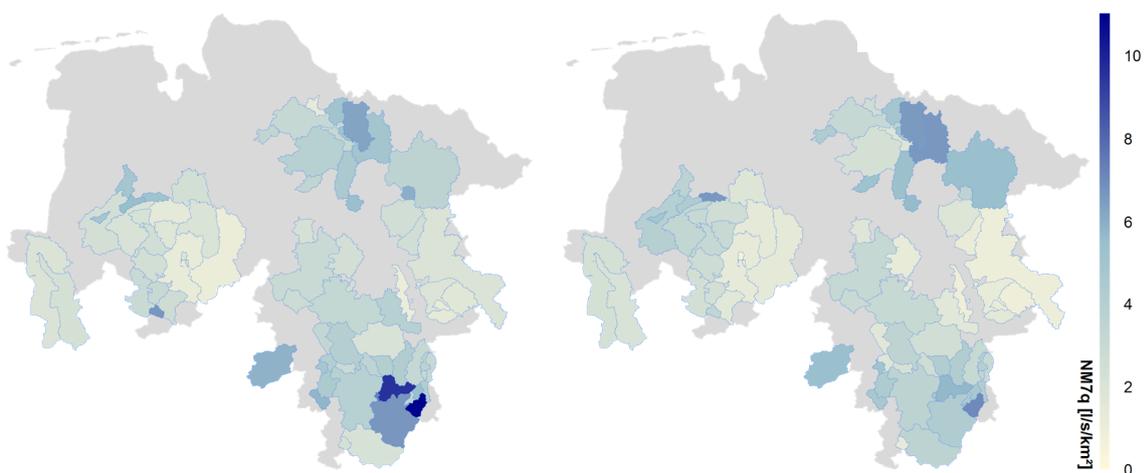
Figure 7.2: Regionalized mNM7q for the periods 1966 – 1995 (left) and 1996 – 2005 (right).
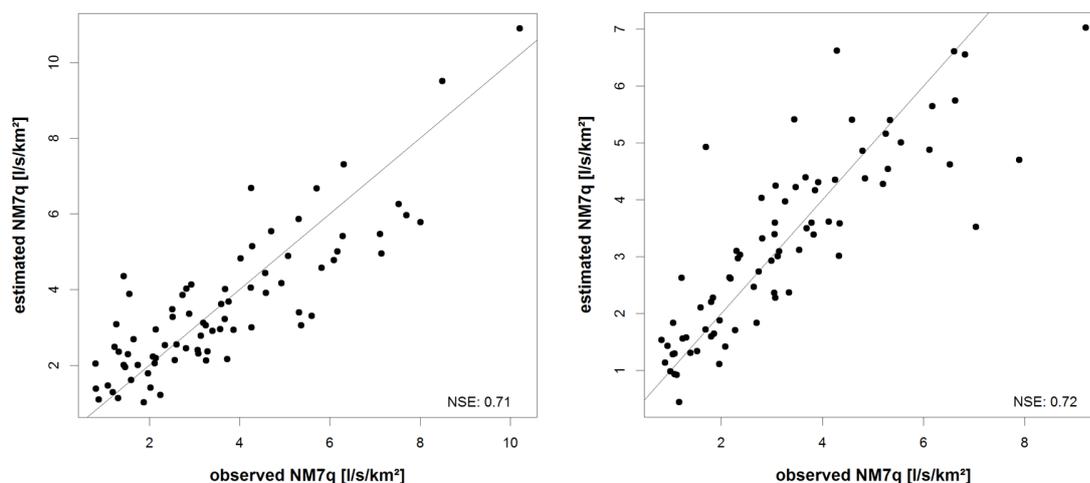
Figure 7.3: Cross validation result showing observed vs. regionalized mNM7q for the periods 1966 – 1995 (left) and 1996 – 2005 (right).

Even though the most accurate procedure, the use of climatic CDs for the future period is difficult in practice. When temporal and spatial models are applied to climate model data, the catchment descriptors would need to be extracted accordingly from this very data. The resulting effect could be two-sided. Either, the error introduced during temporal model fitting due to uncertainty in the input data could be further enhanced through uncertainty during spatial model fitting, or the fact that the same base data is used for temporal and spatial estimation leads to a better regional prognosis, when catchment descriptors from climate model data are used. The effect is to be analyzed through actual application of climate model data in a control period and cannot be exhaustively evaluated in this study. It is therefore assumed that the use of observed past catchment descriptors is preferable for regionalization of simulated future low flow indices. Case b.2 in Table 7.1 summarizes the scenario for observed data. The model is fitted to the observed future data but the catchment descriptors used are the ones from the past. The cross-validation results indicate slightly worse performance than for regionalization using future CDs (case b.1), but deviations are minor. Figure 7.4 (left) shows the regionalized mNM7q and Figure 7.5 (left) compares regionalized with observed values for case b.2.

To test whether the meteorological CDs are also capable of making temporal prognoses, as they differ in time, the model fitted to the past mNM7q using also CDs from the past is applied to the future set of CDs unaltered (case b.3). The results show poor performance, as seen in Figure 7.4 and Figure 7.5 (right) and point out that the CDs are merely capable of reproducing spatial patterns, not temporal ones. Thus, time definitely needs to be incorporated into the model fitting procedure, as done in chapter 6, to be able to model the temporal component. The regional model uses climatic catchment descriptors to model spatial dependencies between the low flow indicators but can by no means predict temporal changes in low flow through application of the same model to a different set of meteorological indices. Case b.4 shows a scenario where the

model fitted to the past period using past catchment descriptors is directly applied to regionalize future values. Due to the great similarity between the two periods, the performance is only slightly worse than for the models fitted to the future period. The comparison between case b.3 and b.4 shows that, even though the change in the target variable between two periods may be minor, changes in the catchment descriptors can be major, indicating that consideration of temporal effects in the relationship between CDs and target variable is essential.
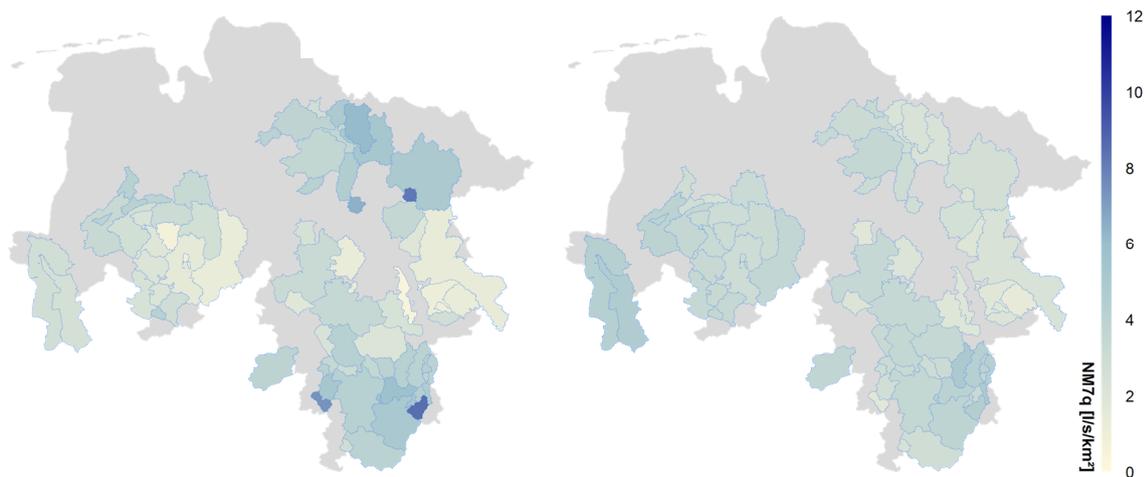
Figure 7.4: Regionalized mNM7q for the period 1996 – 2005 based on past CDs (left) and application of a past model on future CDs right).

Temporal prediction of the mNM7q (case c.1) is highly successful, as expected given the minor difference between past and future period. Figure 7.6 (left) shows the predicted regional mNM7q for the future period and Figure 7.7 (left) the respective scatterplot. 3 catchments are distinguishably overestimated but otherwise the fit is accurate, yielding an NSE of 0.92. The cases c.2 and c.3 pose a repetition of cases b.1 and b.2 and certify that in this setting the use of past catchment descriptors for regionalizing future predicted values is legitimate. The quality criteria hardly differ (NSEs of 0.68 and 0.67, respectively). In the right panels of Figures 7.6 and 7.7, the regionalized predicted mNM7q is depicted (case c.4). Regionalization has been done using past catchment descriptors. There is some additional smoothing in the regional distribution that adds to the error of temporal estimation. When comparing the cross-validated temporally and spatially estimated values with the observation, these errors add up, as can also be seen by the deterioration in goodness-of-fit measures.
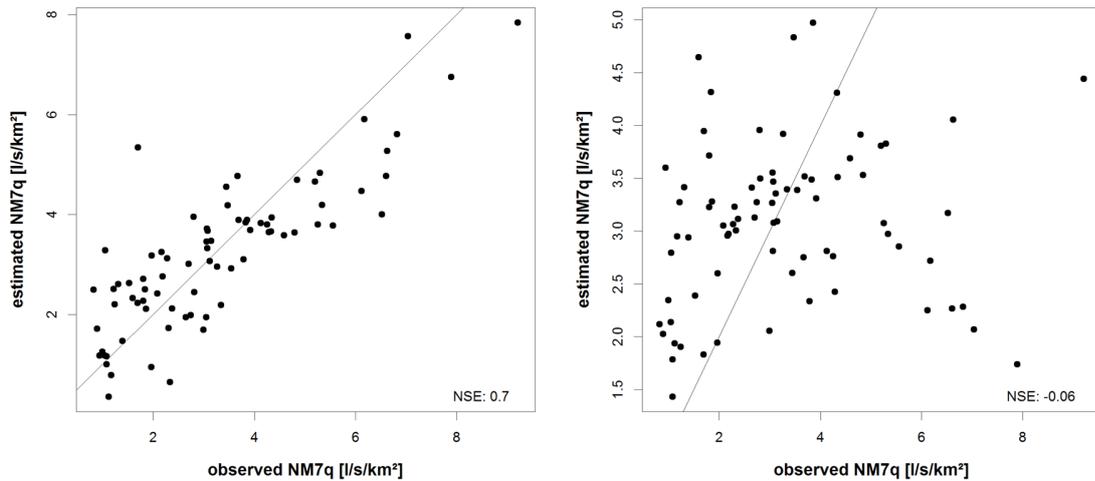
Figure 7.5: Cross-validation result showing observed vs. predicted mNM7q for the period 1996 – 2005 based on past CDs (left) and application of a past model on future CDs (right).

In order to determine how the regionalization of locally estimated index values compares to the direct spatiotemporal models, all methods are applied for the 5 k-means clusters found in chapter 5. The results can be found in Figure 7.8. Compared are the observed future values with the ones simulated from the observed past via subsequent or simultaneous temporal and spatial modeling. Shown are the results for the mNM7q, as well as for the NM7Q10, NM7Q20 and NM7Q50. The quantiles are obtained via regional distribution fitting within the homogeneous regions of chapter 6 for the regionalization of temporally estimated values and the ST-2 index method. For the ST-1 approach, they are directly derived from distributions fitted to each station's simulated annual index values. The subsequent temporal and spatial estimation will in the following be termed T+S for reasons of conciseness.
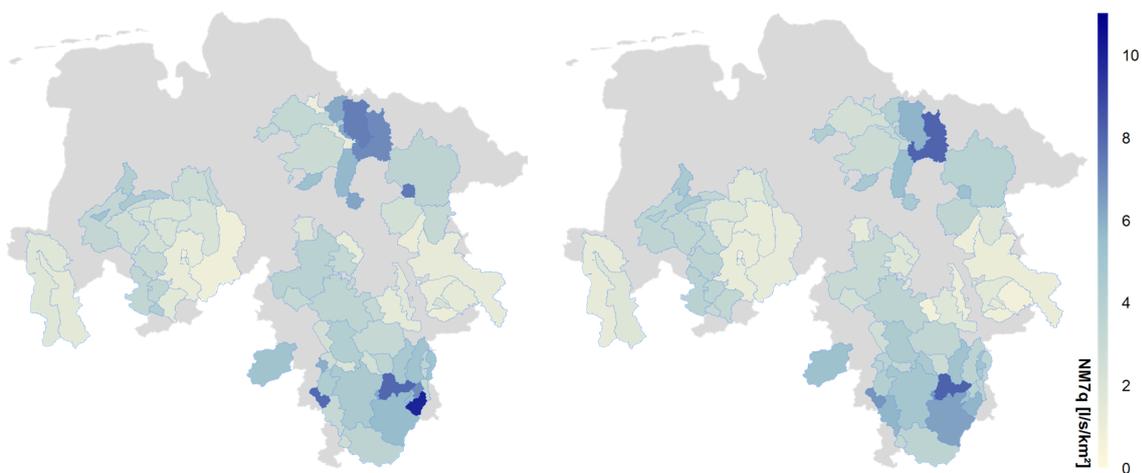
Figure 7.6: Predicted mNM7q for the period 1996 – 2005 simulated using temporal MLR models (left) and cross-validation results from regionalization (right).

It can be seen that the goodness-of-fit measures certify satisfactory performance for all three approaches. However, the ST-1 method appears to outperform the other two methods in almost all aspects. The T+S method surpasses the ST-2 slightly in terms of goodness of fit. The bias is highest for the T+S method, as it indicates significant underestimation of the mean and considered quantiles. The spatiotemporal methods, on the other hand, rather show an overestimation. There appears to be a tendency in bias from the mean to the lower quantiles: it decreases from the mean NM7q to the lower quantiles for all three model approaches. For the T+S model this means that the error increases further, while it becomes smaller for the ST models; the NM7Q50 even yields a negative bias for the ST-2 method. Apart from the bias, a continuous decrease in model performance from mean to lowest quantile can be observed for the ST-1 model, while T+S and ST-2 have lower goodness of fit for the mean than for the NM7q10 or even NM7q20. This effect has already been observed for the regional index method in chapter 5 and is reflected within the spatiotemporal modeling.
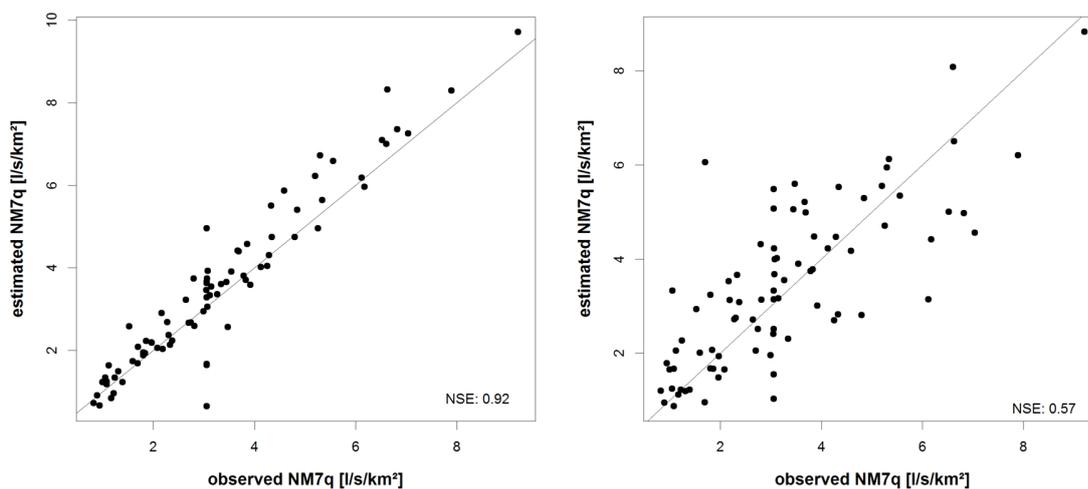
Figure 7.7: Comparison of mNM7q simulated with the temporal station models with the observation and cross-validation result of regionalized simulated mNM7q compared with locally observed mNM7q.

The superiority of the ST-1 method over the ST-2 approach has already been discussed in the previous chapter and is associated with the straightforwardness of modeling annual values rather than aggregated statistics and the estimation of complete future distributions rather than assuming second order stationarity. Also, the direct modeling of the time series in space appears advantageous over station wise modeling and subsequent regionalization. The benefit lies within the obvious problematic of regionalizing predicted values, as the errors made during temporal prediction aggravate the fitting of regional models, since the natural spatial structure is somewhat disrupted. Considering both spatial and temporal effects simultaneously can preclude this error amplification. Additionally, spatiotemporal modeling in the form of panel data regression allows inclusion of significantly larger numbers of explanatory variables for temporal and spatial estimation without the danger of overfitting, due to the increased degrees of freedom

resulting from the panel data set. Thus, leaving out of potentially relevant variables is not an issue compared to the individual station models that allow inclusion of only small numbers of regressors for temporal modeling due to limited calibration lengths and for spatial modeling based on small number of stations in the area.
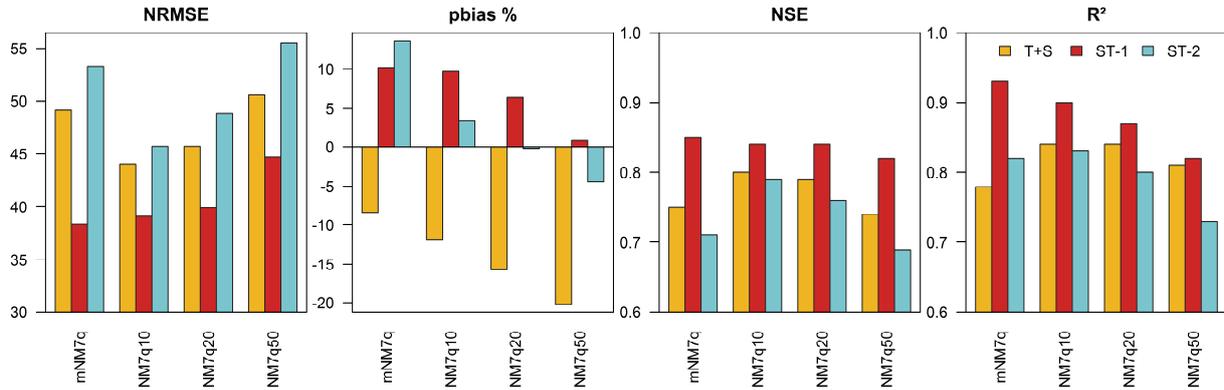
Figure 7.8: Cross validation results for the mean and various quantiles of the NM7q for the subsequent temporal and spatial model (T+S) and the spatiotemporal ST-1 and ST-2 model.

8 Conclusions and outlook

Low flow prediction in time and space is an important but challenging task, especially for the quite extensively managed area of Lower Saxony. However, based exclusively on simple statistical approaches, a framework could be established that may pose an effective means for regional climate change impact assessment in both its spatial and temporal dimension.

Identification of repetitive patterns and trends within low flow time series and elongation into the future for assessment of climate change impacts did not prove feasible. Even though the approach has the major advantage of not requiring any external climate model data input, its predictions are too imprecise to even determine the proper direction of expected developments. Predictions depend greatly on the selected period for model calibration and are significantly aggravated by inconsistencies in trend direction and non-linear changes in the underlying distribution. The combination of a separation of time series components and non-linear model approaches performed significantly better than mere extrapolation of changes in the mean but prediction errors were still higher than for the assumption of stationarity.

External climatic data was thus assumed to represents the only source of information about future low flows. Simple MLR models based on climate data input in the form of meteorological indices showed satisfying results for predicting an observed future. A simple OLS-fitted model, which uses input of principal components of a variety of meteorological indices, appeared to give the most accurate predictions, as it maximizes the utilized information content of external variables without running the risk of overfitting and multicollinearity between the regressors. How well the models perform with climate model data is to be tested using a broad regional climate model ensemble for the area. The main criteria for success will be proper reproduction of the meteorological indices by the climate models and a relationship between regressors and low flow indices that remains more or less stationary over the years. The station-based temporal models have shown that performance in a future validation period is weaker than in the previous calibration period, which potentially indicates a non-stationary behavior, but could also be related to overfitting or omittance of important variables. Due to the minor calibration lengths available for this study, the latter case is probable. In comparison with the statistical approaches classical hydrological modeling appeared to yield better low flow predictions when calibrated explicitly on the low flow periods. Also, they showed no decrease in performance with distance to the calibration period, which suggests that omittance of information is an important issue with the statistical models. Statistical model fitting using the entire observation period would expectedly raise the predictive power, since more external variables could be included without the risk of overfitting. The final level of uncertainty, which arises due to model error and application of climate model data, needs to be assessed in great detail, as done e.g. by (Parajka et al., 2016).

Regionalization of means of low flow time series for the study area appeared challenging. Especially the geostatistical approaches tested did not yield a satisfying spatial prediction. Best results were obtained by MLR methods under inclusion of aggregated meteorological indices as regressors additional to physiographic catchment descriptors. These indices were capable of reproducing regional differences in low flow indices and could account for scaling effects, which allowed comprising of all catchment sizes within a single regional model. Through averaging and aggregation over various spatial and temporal scales, the meteorological indices appeared to represent relevant catchment processes better than the set of physiographic catchment descriptors that was used in this study. For spatial prediction of low flow quantiles the index method appeared superior to direct quantile regionalization. The spatial prediction can be possibly improved by a combination of geostatistical approaches like top kriging with input of external variables. A TK model with external drift would be able to additionally account for the similarity of catchments resulting from partial overlaps.

The combination of temporal and spatial prediction within a single model appeared not only feasible but overall superior to modeling both components individually. The best method that could be identified is the spatiotemporal modeling of annual low flow index values based on panel data regression models fitted to complete time series of all stations in a region. Modeling annual values instead of long term statistics has the additional advantage that all aspects of the low flow distribution are predicted for the future, including changes in mean, variance, skewness etc. Equally to the temporal models, the predictive power will potentially increase if applied to the full observation period rather than a calibration period only. The panel data regression model will also be capable of considering inconsistent time series lengths, so that all available observations in the area can be included into the model fitting procedure, making it even more powerful. When compared directly, panel data regression of annual index values outperformed regionalization of station-wise estimated future low flows. However, in order to evaluate the models' full abilities for climate change impact assessment they also need to be applied to actual climate model data. Only plausible future meteorological index distributions will yield meaningful distributions of low flow indices.

A crucial point that holds for all the applied methods in general is the fact that statistical modeling of time series data comes with a vast number of preconditions, pre-processing steps and restrictions to achieve proper inference. Issues in regional non-stationary modeling arise especially from serial autocorrelation and spatial cross-correlation. Temporal dependence could be identified in several cases, even though the observed time series of index values were considered on a coarse yearly resolution. It thus needs to be paid attention that any autocorrelation can be explained by the predictors, which happened to be the case in most of the analyzed time series. Otherwise models should be selected that allow handling of this deficiency. Spatial correlation has been given less attention to in this study. Analyses of the residuals of regional regression models have not exhibited any obvious correlation structure, which was accepted as proof for independence. More detailed investigations on cross-

correlations need to be carried out, ideally using entire observed time series rather than average values only, to reveal actual correlation structures. Also, fitting of regional probability distributions should be subject to independence constraints. Regional low flow frequency analyses should therefore ideally be applied to a set of independent catchments. Cross-correlation within rivers could be reduced through subtraction of incoming flow of upstream catchments. Several attempts have been made for this work, none yielding satisfactory results but unrealistic low flow values. The procedures may have enhanced effects of human interference.

Anthropogenic influence is probably the most critical problem in the study area. This holds true for all approaches tested in this study, as none of them can handle alteration within the river basins. Present interference may be contained in the temporal models by relating the meteorological input to an influenced low flow discharge. Any change in the magnitude of this influence will drastically effect the prediction of the MLR models. Spatial models, on the other hand, are affected by present alterations, as these result in regional low flow differences that cannot be explained by natural factors. Thus, careful handling of the models via proper assessment of the degree of anthropogenic impact, as well as regular model update in case of changing usage is advised. Ideal would be the inclusion of indices that describe the magnitude of anthropogenic interference. Such values, however, are hard to derive and often reversely related to low flow magnitude, which aggravates modeling.

There is a range of possibilities for development and improvement of the applied methods. Panel data tools proved quite successful and could be applied for temporal and regional assessment of change points and detection of anthropogenic influence, but also for delineation of regions with similar temporal development. Panel data regression is also capable of fitting time-variant regression parameters, which could aid the prognosis of future low flows tremendously in a non-stationary setting. L-moment regression could also be coupled with geostatistical techniques for regionalization of higher order moments based on proximity rather than catchment descriptors. Finally, interrelationships between the different low flow indices need to be assessed, in order to improve temporal and spatial prediction of all relevant aspects. Some indices could be much easier predicted in time than in space, and vice versa, so that a weighted combination of various indices in spatiotemporal models my potentially increase the overall predictability of indices in both time and space.

Finally, all tested approaches should be compared with continuous hydrological modeling in order to assess the actual value of the simple statistical methods and to bring physical meaning to the relationships found in the latter. Hydrological models can simulate long- and short-term storage behavior, as well as soil conditions and various other low-flow relevant processes. A detailed sensitivity analysis of model parameters of the hydrological model with respect to low flows could yield a better insight into the actual low-flow determining processes in a catchment, especially when parameter optimization occurs merely for specific low flow events, like in this

study. A detailed analysis could be carried out for relating lead times and base periods of dry and wet events used in the statistical models to important storage conditions in the hydrological model to achieve a better understanding of the mere black-box approach and to better identify truly relevant input variables. This step would lead to compliance with the ideas of the PUB initiative, which promotes understanding of catchment processes for regionalization.

Whether the tested approaches will be equally applicable to other regions and other streamflow metrics needs to be assessed. The overall high performance of the models independent from size, elevation or position of the considered catchments is a good prospective for a wider application.

References

Agarwal A., Maheswaran R., Kurths J., Khosa R. (2016) Wavelet Spectrum and Self-Organizing Maps-Based Approach for Hydrologic Regionalization -a Case Study in the Western United States. Water Resources Management 30:4399-4413. DOI: 10.1007/s11269-016-1428-1.

Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, MR 0423716, Doi:10.1109/TAC.1974.1100705.

Archfield S.A., Vogel R.M. (2010) Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. Water Resources Research 46. DOI: Artn W10513 10.1029/2009wr008481.

Archfield S.A., Pugliese A., Castellarin A., Skøien J.O., Kiang J.E. (2013) Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach? Hydrology and Earth System Sciences 17:1575-1588. DOI: 10.5194/hess-17-1575-2013.

Baltagi B.H. Econometric analysis of panel data. Fifth Edition.

BGR (2013) Bodenübersichtskarte der Bundesrepublik Deutschland 1:1000000. Bundesanstalt für Geowissenschaften und Rohstoffe, Berlin.

Box G.E.P., Cox D.R. (1964) An Analysis of Transformations. Journal of the Royal Statistical Society Series B-Statistical Methodology 26:211-252.

Burn D.H. (1990) An Appraisal of the Region of Influence Approach to Flood Frequency-Analysis. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 35:149-165. DOI: Doi 10.1080/02626669009492415.

Burn D.H., Sharif M., Zhang K. (2010) Detection of trends in hydrological extremes for Canadian watersheds. Hydrological Processes 24:1781-1790. DOI: 10.1002/hyp.7625.

Castellarin A. (2007) Probabilistic envelope curves for design flood estimation at ungauged sites. Water Resources Research 43. DOI: Artn W04406 10.1029/2005wr004384.

Castellarin A. (2014) Regional prediction of flow-duration curves using a three-dimensional kriging. Journal of Hydrology 513:179-191. DOI: 10.1016/j.jhydrol.2014.03.050.

Castiglioni S., Castellarin A., Montanari A. (2009) Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. Journal of Hydrology 378:272-280. DOI: 10.1016/j.jhydrol.2009.09.032.

Castiglioni S., Castellarin A., Montanari A., Skøien J.O., Laaha G., Blöschl G. (2011) Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging. Hydrology and Earth System Sciences 15:715-727. DOI: 10.5194/hess-15-715-2011.

Chokmani K., Ouarda T.B.M.J. (2004) Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research 40. DOI: Artn W12514 10.1029/2003wr002983.

Chow G.C. (1960) Tests of Equality between Sets of Coefficients in 2 Linear Regressions. Econometrica 28:591-605. DOI: Doi 10.2307/1910133.

Clausen B., Pearson C.P. (1995) Regional Frequency-Analysis of Annual Maximum Streamflow Drought. Journal of Hydrology 173:111-130. DOI: Doi 10.1016/0022-1694(95)02713-Y.

Cloke H.L., Pappenberger F. (2009) Ensemble flood forecasting: A review. Journal of Hydrology 375:613-626. DOI: 10.1016/j.jhydrol.2009.06.005.

Coles S. (2001) An introduction to statistical modeling of extreme values Springer, London.

Croissant, Y., Millo, G., Toomet, O., Kleiner, C., Zeileis, A., Henningsen, A., Andronic, L., Schoenfelder, N. (2016) Package „plm": Linear Models for Panel Data.

Cunderlik J.M., Burn D.H. (2003) Non-stationary pooled flood frequency analysis. Journal of Hydrology 276:210-223. DOI: 10.1016/S0022-1694(03)00062-3.

Cunderlik J.M., Ouarda T.B.M.J. (2006) Regional flood-duration-frequency modeling in the changing environment. Journal of Hydrology 318:276-291. DOI: 10.1016/j.jhydrol.2005.06.020.

Dai A.G. (2013) Increasing drought under global warming in observations and models (vol 3, pg 52, 2013). Nature Climate Change 3:171-171. DOI: 10.1038/Nclimate1811.

Dalrymple, T. (1960) Flood Frequency Analysis. U.S. Geol. Surv. Water Supply Pap., 1543-A.

Dawson C.W., Abrahart R.J., Shamseldin A.Y., Wilby R.L. (2006) Flood estimation at ungauged sites using artificial neural networks. Journal of Hydrology 319:391-409. DOI: 10.1016/j.jhydrol.2005.07.032.

de Wit M.J.M., van den Hurk B., Warmerdam P.M.M., Torfs P.J.J.F., Roulin E., van Deursen W.P.A. (2007) Impact of climate change on low-flows in the river Meuse. Climatic Change 82:351-372. DOI: 10.1007/s10584-006-9195-2.

Delgado J.M., Merz B., Apel H. (2014) Projecting flood hazard under climate change: an alternative approach to model chains. Natural Hazards and Earth System Sciences 14:1579-1589. DOI: 10.5194/nhess-14-1579-2014.

Demuth S. (1993) Untersuchungen zum Niedrigwasser in West-Europa, Freiburger Schriften zur Hydrologie. pp. 205.

Douglas E.M., Vogel R.M., Kroll C.N. (2000) Trends in floods and low flows in the United States: impact of spatial correlation. Journal of Hydrology 240:90-105. DOI: Doi 10.1016/S0022-1694(00)00336-X.

Durbin J., Watson G.S. (1950) Testing for Serial Correlation in Least Squares Regression .1. Biometrika 37:409-428. DOI: Doi 10.2307/2332391.

Durbin J., Watson G.S. (1951) Testing for Serial Correlation in Least Squares Regression .2. Biometrika 38:159-178. DOI: Doi 10.2307/2332325.

Durbin J., Watson G.S. (1971) Testing for Serial Correlation in Least Squares Regression .3. Biometrika 58:1-&.

Durrans S.R., Tomic S. (1996) Regionalization of low-flow frequency estimates: An Alabama case study. Water Resources Bulletin 32:23-37.

DWA (2009) Regionalisierung von Niedrigwasserkenngrößen. DWA-Themen HW 3.1 – August 2009. Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall e. V., Hennef.

Eng K., Milly P.C.D., Tasker G.D. (2007) Flood regionalization: A hybrid geographic and predictor-variable region-of-influence regression method. Journal of Hydrologic Engineering 12:585-591. DOI: 10.1061/(Asce)1084-0699(2007)12:6(585).

Fangmann, A. (2012) Time series analysis for drought and low flow indicators in Lower Saxony. Master Thesis. Leibniz Universität Hannover.

Fangmann, A., Belli, A., Haberlandt, U. (2013) Trends in beobachteten Abflusszeitreihen in Niedersachsen. Hydrologie und Wasserbewirtschaftung 57(5):196-205. DOI: 10.5675/HyWa_2013,5_1

Farsadnia F., Kamrood M.R., Nia A.M., Modarres R., Bray M.T., Han D., Sadatinejad J. (2014) Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. Journal of Hydrology 509:387-397. DOI: 10.1016/j.jhydrol.2013.11.050.

Fernandez C., Vega J.A., Fonturbel T., Jimenez E. (2009) Streamflow drought time series forecasting: a case study in a small watershed in North West Spain. Stochastic Environmental Research and Risk Assessment 23:1063-1070. DOI: 10.1007/s00477-008-0277-8.

Forzieri G., Feyen L., Rojas R., Florke M., Wimmer F., Bianchi A. (2014) Ensemble projections of future streamflow droughts in Europe. Hydrology and Earth System Sciences 18:85-108. DOI: 10.5194/hess-18-85-2014.

Goldfeld S.M., Quandt R.E. (1965) Some Tests for Homoscedasticity. Journal of the American Statistical Association 60:539-547. DOI: Doi 10.2307/2282689.

GRDC (2016) The Global Runoff Data Centre, 56068 Koblenz, Germany.

Gudmundsson L., Tallaksen L.M., Stahl K. (2011) Spatial cross-correlation patterns of European low, mean and high flows. Hydrological Processes 25:1034-1045. DOI: 10.1002/hyp.7807.

Gustard A., Bullock A., Dixon J.M. (1992) Low flow estimation in the United Kingdom, in: I. o. Hydrology (Ed.), Institute of Hydrology. pp. 88.

Haberlandt U., Klocking B., Krysanova V., Becker A. (2001) Regionalisation of the base flow index from dynamically simulated flow components - a case study in the Elbe River Basin. Journal of Hydrology 248:35-53. DOI: Doi 10.1016/S0022-1694(01)00391-2.

Hall M.J., Minns A.W. (1999) The classification of hydrologically homogeneous regions. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 44:693-704. DOI: Doi 10.1080/02626669909492268.

Hartigan, J. A., Wong, M. A. (1979) Algorithm AS 136: A k-means clustering algorithm. Applied Statistics 28(1):100–108.

Hisdal H., Stahl K., Tallaksen L.M., Demuth S. (2001) Have streamflow droughts in Europe become more severe or frequent? International Journal of Climatology 21:317-+. DOI: Doi 10.1002/Joc.619.

Holmes M.G.R., Young A.R., Gustard A., Grew R. (2002) A region of influence approach to predicting flow duration curves within ungauged catchments. Hydrology and Earth System Sciences 6:721-731.

Hosking, J.R.M., (1985) Algorithm AS 215: Maximum-likelihood estimation of the parameter of the generalized extreme-value distribution, Appl. Stat. 34:301–310.

Hosking J.R.M., Wallis J.R. (1997) Regional frequency analysis : an approach based on L-moments Cambridge University Press, Cambridge.

Hothorn, T., Zeileis, A., Farebrother, R.W., Cummins, C., Millo, G., Mitchell, D. (2011) Package "lmtest": Testing Linear Regression Models.

Hrachowitz M., Soulsby C., Tetzlaff D., Speed M. (2010) Catchment transit times and landscape controls-does scale matter? Hydrological Processes 24:117-125. DOI: 10.1002/hyp.7510.

Huang N.E., Shen Z., Long S.R., Wu M.L.C., Shih H.H., Zheng Q.N., Yen N.C., Tung C.C., Liu H.H. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear

and non-stationary time series analysis. Proceedings of the Royal Society a-Mathematical Physical and Engineering Sciences 454:903-995.

Institute of Hydrology (1980) Low Flow Studies Report No. 1: Research Report. Wallingford, GB.

Jarvis A., Reuter, H.I., Nelson, A., Guevara, E. (2008) Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT).

Kendall, M.G. (1975): Rank Correlation Methods. Griffin, London.

Kim, D., Oh, H.-S. (2014) Package "EMD": Empirical Mode Decomposition and Hilbert Spectral Analysis.

Koehler, G., M. Schwab, M. von Hauff, K. Kluth, W. Finke and J. U. Belz (2006): Niedrigwasserperiode 2003 in Deutschland. Ursachen – Wirkungen – Folgen. BfG-Mitteilung Nr. 27, Koblenz.

Koenker R., Bassett G. (1978) Regression Quantiles. Econometrica 46:33-50. DOI: Doi 10.2307/1913643.

Koenker R., Dorey V. (1994) A Remark on Algorithm as-229 - Computing Dual Regression Quantiles and Regression Rank Scores. Applied Statistics-Journal of the Royal Statistical Society Series C 43:410-414. DOI: Doi 10.2307/2986030.

Koenker R.W., Dorey V. (1987) Computing Regression Quantiles. Applied Statistics-Journal of the Royal Statistical Society Series C 36:383-393. DOI: Doi 10.2307/2347802.

Koenker, R.W., Portnoy, S., Ng, P.T., Zeileis, A., Grosjean, P., Ripley, B. D. (2017) Package "quantreg": Quantile Regression.

Krige D.G. (1951) A statistical approach to some mine valuation and allied problems on the Witwatersrand, University of the Witwatersrand, Johannesburg.

Kroll C.N., Vogel R.M. (2002) Probability distribution of low streamflow series in the United States. Journal of Hydrologic Engineering 7:137-146. DOI: 10.1061/(Asce)1084-0699(2002)7:2(137).

Laaha G., Blöschl G. (2006a) A comparison of low flow regionalisation methods - catchment grouping. Journal of Hydrology 323:193-214. DOI: 10.1016/j.jhydrol.2005.09.001.

Laaha G., Blöschl G. (2006b) Low flow estimation in Austria. Climate Variability and Change - Hydrological Impacts 308:117-+.

Laaha G., Blöschl G. (2006c) Seasonality indices for regionalizing low flows. Hydrological Processes 20:3851-3878. DOI: 10.1002/hyp.6161.

Laaha G., Skøien J.O., Blöschl G. (2014) Spatial prediction on river networks: comparison of top-kriging with regional regression. Hydrological Processes 28:315-324. DOI: 10.1002/hyp.9578.

Laaha G., Skøien J.O., Nobilis F., Blöschl G. (2013) Spatial Prediction of Stream Temperatures Using Top-Kriging with an External Drift. Environmental Modeling & Assessment 18:671-683. DOI: 10.1007/s10666-013-9373-3.

Laaha G., Parajka J., Viglione A., Koffler D., Haslinger K., Schoner W., Zehetgruber J., Blöschl G. (2016) A three-pillar approach to assessing climate impacts on low flows. Hydrology and Earth System Sciences 20:3967-3985. DOI: 10.5194/hess-20-3967-2016.

Labat D. (2010) Cross wavelet analyses of annual continental freshwater discharge and selected climate indices. Journal of Hydrology 385:269-278. DOI: 10.1016/j.jhydrol.2010.02.029.

Labat D., Godderis Y., Probst J.L., Guyot J.L. (2004) Evidence for global runoff increase related to climate warming. Advances in Water Resources 27:631-642. DOI: 10.1016/j.advwatres.2004.02.020.

LBEG: Geologische Übersichtskarte von Niedersachsen 1 : 500 000. – digitaler Datensatz des Niedersächsischen Bodeninformationssystems (NIBIS®); Landesamt für Bergbau, Energie und Geologie (LBEG), Hannover.

Leclerc M., Ouarda T.B.M.J. (2007) Non-stationary regional flood frequency analysis at ungauged sites. Journal of Hydrology 343:254-265. DOI: 10.1016/j.jhydrol.2007.06.021.

Levene, H. (1960) Robust tests for equality of variances. In: Olkin, I., Hotelling, H. et al. (editors): Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, p. 278-292.

Ley R., Casper M.C., Hellebrand H., Merz R. (2011) Catchment classification by runoff behaviour with self-organizing maps (SOM). Hydrology and Earth System Sciences 15:2947-2962. DOI: 10.5194/hess-15-2947-2011.

Lima C.H.R., Lall U. (2010) Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow. Journal of Hydrology 383:307-318. DOI: 10.1016/j.jhydrol.2009.12.045.

Lin G.F., Chen L.H. (2006) Identification of homogeneous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology 324:1-9. DOI: 10.1016/j.jhydrol.2005.09.009.

Lin J.Y., Cheng C.T., Chau K.W. (2006) Using support vector machines for long-term discharge prediction. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 51:599-612. DOI: DOI 10.1623/hysj.51.4.599.

Lindstrom G., Johansson B., Persson M., Gardelin M., Bergstrom S. (1997) Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology 201:272-288. DOI: Doi 10.1016/S0022-1694(97)00041-3.

Ling H.B., Xu H.L., Shi W., Zhang Q.Q. (2011) Regional climate change and its effects on the runoff of Manas River, Xinjiang, China. Environmental Earth Sciences 64:2203-2213. DOI: 10.1007/s12665-011-1048-2.

Liu W.T., Juarez R.I.N. (2001) ENSO drought onset prediction in northeast Brazil using NDVI. International Journal of Remote Sensing 22:3483-3501. DOI: Doi 10.1080/01431160010006430.

Longobardi A., Villani P. (2013) A statistical, parsimonious, empirical framework for regional flow duration curve shape prediction in high permeability Mediterranean region. Journal of Hydrology 507:174-185. DOI: 10.1016/j.jhydrol.2013.10.019.

Lopez J., Frances F. (2013) Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates. Hydrology and Earth System Sciences 17:3189-3203. DOI: 10.5194/hess-17-3189-2013.

Mann H.B., Whitney D.R. (1947) On a Test of Whether One of 2 Random Variables Is Stochastically Larger Than the Other. Annals of Mathematical Statistics 18:50-60. DOI: DOI 10.1214/aoms/1177730491.

Marechal D., Holman I.P. (2005) Development and application of a soil classification-based conceptual catchment-scale hydrological model. Journal of Hydrology 312:277-293. DOI: 10.1016/j.jhydrol.2005.02.018.

Matheron, G. (1963) Principles of geostatistics. Economic Geol. 58:1246–1266. DOI: 10.2113/gsecongeo.58.8.1246.

McCabe G.J., Wolock D.M. (2002) A step increase in streamflow in the conterminous United States. Geophysical Research Letters 29. DOI: Artn 218510.1029/2002gl015999.

McKee, T.B., Doesken, N.J., Kleist, J. (1993) The relationship of drought frequency and duration to time scales. 8th Conference on Applied Climatology. American meteorological society, Boston, USA.

Merz R., Blöschl G. (2005) Flood frequency regionalisation-spatial proximity vs. catchment attributes. Journal of Hydrology 302:283-306. DOI: 10.1016/j.jhydrol.2004.07.018.

Merz R., Blöschl G., Humer G. (2008) National flood discharge mapping in Austria. Natural Hazards 46:53-72. DOI: 10.1007/s11069-007-9181-7.

Meyer, D. (2017) Package "e1071": Misc Functions of the Department of Statistics, Probability Theory Group (Formerly E1071), TU Wien.

Modarres R. (2007) Streamflow drought time series forecasting. Stochastic Environmental Research and Risk Assessment 21:223-233. DOI: 10.1007/s00477-006-0058-1.

Modarres R. (2008) Regional frequency distribution type of low flow in north of Iran by L-moments. Water Resources Management 22:823-841. DOI: 10.1007/s11269-007-9194-8.

Nash, J.E., Sutcliffe, J.V. (1970) River flow forecasting through conceptual models part I – A discussion of principles. Journal of Hydrology 10(3):282–290. DOI: 10.1016/0022-1694(70)90255-6.

Nathan R.J., Mcmahon T.A. (1990a) Identification of Homogeneous Regions for the Purposes of Regionalisation. Journal of Hydrology 121:217-238. DOI: Doi 10.1016/0022-1694(90)90233-N.

Nathan R.J., Mcmahon T.A. (1990b) Practical Aspects of Low-Flow Frequency-Analysis. Water Resources Research 26:2135-2141.

Nds. MU (2012): Landschaft. www.umwelt.niedersachsen.de. Niedersächsisches Ministerium für Umwelt, Energie und Klimaschutz, Hannover, Germany [http://www.umwelt.niedersachsen.de/portal/live.php?navigation_id=2710&article_id=7971&_psmand=10, date of last access: June 29, 2012].

Nelder J.A., Mead R. (1965) A Simplex-Method for Function Minimization. Computer Journal 7:308-313.

NLS (2007): Nutzungsarten der Bodenfläche: Flächenerhebung zum 31.12.2004. Statistische Berichte Niedersachsen. Niedersächsisches Landesamt für Statistik, Hannover, Germany.

NLWKN (2012) Globaler Klimawandel: Wasserwirtschaftliche Folgenabschätzung für das Binnenland. Oberirdische Gewässer Band 33. Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz Betriebsstelle Hannover-Hildesheim.

Ochola W.O., Kerkides P. (2003) A markov chain simulation model for predicting critical wet and dry spells in Kenya: Analysing rainfall events in the Kano Plains. Irrigation and Drainage 52:327-342. DOI: Doi 10.1002/lrd.94.

Ouarda T.B.M.J., Shu C. (2009) Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resources Research 45. DOI: Artn W11428 10.1029/2008wr007196.

Ouarda T.B.M.J., Charron C., St-Hilaire A. (2008) Statistical Models and the Estimation of Low Flows. Canadian Water Resources Journal 33:195-205.

Ouarda T.B.M.J., Cunderlik J.M., St-Hilaire A., Barbet M., Bruneau P., Bobee B. (2006) Data-based comparison of seasonality-based regional flood frequency methods. Journal of Hydrology 330:329-339. DOI: 10.1016/j.jhydrol.2006.03.023.

Parajka J., Blaschke A.P., Blöschl G., Haslinger K., Hepp G., Laaha G., Schoner W., Trautvetter H., Viglione A., Zessner M. (2016) Uncertainty contributions to low-flow projections in Austria. Hydrology and Earth System Sciences 20:2085-2101. DOI: 10.5194/hess-20-2085-2016.

Prudhomme C., Wilby R.L., Crooks S., Kay A.L., Reynard N.S. (2010) Scenario-neutral approach to climate change impact studies: Application to flood risk. Journal of Hydrology 390:198-209. DOI: 10.1016/j.jhydrol.2010.06.043.

Pugliese A., Farmer W.H., Castellarin A., Archfield S.A., Vogel R.M. (2016) Regional flow duration curves: Geostatistical techniques versus multivariate regression. Advances in Water Resources 96:11-22. DOI: 10.1016/j.advwatres.2016.06.008.

Renard B., Lang M., Bois P., Dupeyrat A., Mestre O., Niel H., Sauquet E., Prudhomme C., Parey S., Paquet E., Neppel L., Gailhard J. (2008) Regional methods for trend detection: Assessing field significance and regional consistency. Water Resources Research 44. DOI: Artn W08419 10.1029/2007wr006268.

Rodriguez-Iturbe I. (1969) Estimation of Statistical Parameters for Annual River Flows. Water Resources Research 5:1418-+. DOI: Doi 10.1029/Wr005i006p01418.

Salinas J.L., Laaha G., Rogger M., Parajka J., Viglione A., Sivapalan M., Blöschl G. (2013) Comparative assessment of predictions in ungauged basins - Part 2: Flood and low flow studies. Hydrology and Earth System Sciences 17:2637-2652. DOI: 10.5194/hess-17-2637-2013.

Samaniego L., Bardossy A., Kumar R. (2010) Streamflow prediction in ungauged catchments using copula-based dissimilarity measures. Water Resources Research 46. DOI: Artn W02506 10.1029/2008wr007695.

Sanborn S.C., Bledsoe B.P. (2006) Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon. Journal of Hydrology 325:241-261. DOI: 10.1016/j.jhydrol.2005.10.018.

Schnier S., Cai X.M. (2014) Prediction of regional streamflow frequency using model tree ensembles. Journal of Hydrology 517:298-309. DOI: 10.1016/j.jhydrol.2014.05.029.

Schwarz, G. E. (1978) Estimating the dimension of a model. Annals of Statistics 6(2):461–464, MR 468014, Doi:10.1214/aos/1176344136.

Shapiro S.S., Wilk M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika 52:591-&. DOI: Doi 10.2307/2333709.

Shi P., Chen X., Qu S.M., Zhang Z.C., Ma J.L. (2010) Regional Frequency Analysis of Low Flow Based on L Moments: Case Study in Karst Area, Southwest China. Journal of Hydrologic Engineering 15:370-377. DOI: 10.1061/(Asce)He.1943-5584.0000206.

Shu C., Ouarda T.B.M.J. (2008) Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. Journal of Hydrology 349:31-43. DOI: 10.1016/j.jhydrol.2007.10.050.

Sivapalan M. (2003) Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrological Processes 17:3163-3170. DOI: 10.1002/hyp.5155.

Skøien, J. O., R. Merz, and G. Blöschl (2006), Top-kriging—Geostatistics on stream networks, Hydrol. Earth Syst. Sci., 10, 277–287.

Skøien J.O., Blöschl G. (2007) Spatiotemporal topological kriging of runoff time series. Water Resources Research 43. DOI: Artn W09419 10.1029/2006wr005760.

Skøien, J.O. (2016) Package "rtop": Interpolation of Data with Variable Spatial Support.

Smakhtin V.U. (2001) Low flow hydrology: a review. Journal of Hydrology 240:147-186. DOI: Doi 10.1016/S0022-1694(00)00340-1.

Smola A.J., Kondor R. (2003) Kernels and regularization on graphs. Learning Theory and Kernel Machines 2777:144-158. DOI: Doi 10.1007/978-3-540-45167-9_12.

Smola A.J., Scholkopf B. (2004) A tutorial on support vector regression. Statistics and Computing 14:199-222. DOI: Doi 10.1023/B:Stco.0000035301.49549.88.

Snelder T.H., Lamouroux N., Leathwick J.R., Pella H., Sauquet E., Shankar U. (2009) Predictive mapping of the natural flow regimes of France. Journal of Hydrology 373:57-67. DOI: 10.1016/j.jhydrol.2009.04.011.

Srinivas V.V., Tripathi S., Rao A.R., Govindaraju R.S. (2008) Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology 348:148-166. DOI: 10.1016/j.jhydrol.2007.09.046.

Stahl K., Hisdal H., Hannaford J., Tallaksen L.M., van Lanen H.A.J., Sauquet E., Demuth S., Fendekova M., Jodar J. (2010) Streamflow trends in Europe: evidence from a dataset of near-natural catchments. Hydrology and Earth System Sciences 14:2367-2382. DOI: 10.5194/hess-14-2367-2010.

Tallaksen L.M., Lanen H.A.J.v. (2004) Hydrological drought : processes and estimation methods for streamflow and groundwater Elsevier, Amsterdam ; London.

Theil, H. (1950a): A rank-invariant method of linear and polynomial regression analysis, I. Proceedings of the Koninklijke Nederlandse Academie van Wetenschappen 53: 386–392.

Theil, H. (1950b): A rank-invariant method of linear and polynomial regression analysis, II. Proceedings of the Koninklijke Nederlandse Academie van Wetenschappen 53: 521–525.

Theil, H. (1950c): A rank-invariant method of linear and polynomial regression analysis, III. Proceedings of the Koninklijke Nederlandse Academie van Wetenschappen 53: 1397–1412.

Toth E. (2013) Catchment classification based on characterisation of streamflow and precipitation time series. Hydrology and Earth System Sciences 17:1149-1159. DOI: 10.5194/hess-17-1149-2013.

Tsakiris G., Nalbantis I., Cavadias G. (2011) Regionalization of low flows based on Canonical Correlation Analysis. Advances in Water Resources 34:865-872. DOI: 10.1016/j.advwatres.2011.04.007.

Umweltbundesamt (2009) CORINE Land Cover (CLC2006). Umweltbundesamt, Dessau.

van Loon A.F., Laaha G. (2015) Hydrological drought severity explained by climate and catchment characteristics. Journal of Hydrology 526:3-14. DOI: 10.1016/j.jhydrol.2014.10.059.

van Vliet M.T.H., Ludwig F., Zwolsman J.J.G., Weedon G.P., Kabat P. (2011) Global river temperatures and sensitivity to atmospheric warming and changes in river flow. Water Resources Research 47. DOI: Artn W02544 10.1029/2010wr009198.

Vezza P., Comoglio C., Rosso M., Viglione A. (2010) Low Flows Regionalization in North-Western Italy. Water Resources Management 24:4049-4074. DOI: 10.1007/s11269-010-9647-3.

Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V., Chervonenkis, A. (1974) Theory of Pattern Recognition. Nauka, Moscow. (in Russian); German translation: Theorie der Zeichenerkennung. Akademie Verlag, Berlin, 1979.

Vicente-Serrano S.M., Begueria S., Lopez-Moreno J.I. (2010) A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. Journal of Climate 23:1696-1718. DOI: 10.1175/2009JCLI2909.1.

Vogel R.M., Kroll C.N. (1990) Generalized Low-Flow Frequency Relationships for Ungaged Sites in Massachusetts. Water Resources Bulletin 26:241-253.

Vogel R.M., Kroll C.N. (1992) Regional Geohydrologic-Geomorphic Relationships for the Estimation of Low-Flow Statistics. Water Resources Research 28:2451-2458. DOI: Doi 10.1029/92wr01007.

Vogel R.M., Yaindl C., Walter M. (2011) Nonstationarity: Flood Magnification and Recurrence Reduction Factors in the United States. Journal of the American Water Resources Association 47:464-474. DOI: 10.1111/j.1752-1688.2011.00541.x.

Von Storch V.H. (1995): Misuses of statistical analysis in climate research. In: Analysis of Climate Variability: Applications of Statistical Techniques, von Storch H, Navarra A (eds). Springer-Verlag, Berlin. 11–26.

Vrugt J.A., Robinson B.A. (2007) Improved evolutionary optimization from genetically adaptive multimethod search. Proceedings of the National Academy of Sciences of the United States of America 104:708-711. DOI: 10.1073/pnas.0610471104.

Wallner M., Haberlandt U., Dietrich J. (2013) A one-step similarity approach for the regionalization of hydrological model parameters based on Self-Organizing Maps. Journal of Hydrology 494:59-71. DOI: 10.1016/j.jhydrol.2013.04.022.

Ward P.J., Eisner S., Florke M., Dettinger M.D., Kummu M. (2014) Annual flood sensitivities to El Nino-Southern Oscillation at the global scale. Hydrology and Earth System Sciences 18:47-66. DOI: 10.5194/hess-18-47-2014.

Webster R., Oliver M.A. (2007) Geostatistics for environmental scientists. 2nd ed. ed. Wiley, Chichester.

Westerberg I.K., Wagener T., Coxon G., McMillan H.K., Castellarin A., Montanari A., Freer J. (2016) Uncertainty in hydrological signatures for gauged and ungauged catchments. Water Resources Research 52:1847-1865. DOI: 10.1002/2015WR017635.

Wilcoxon F. (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 1:80-83. DOI: Doi 10.2307/3001968.

WMO (2008) Manual on Low-flow Estimation and Prediction. WMO, No. 1029

Wood A.W., Maurer E.P., Kumar A., Lettenmaier D.P. (2002) Long-range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research-Atmospheres 107. DOI: Artn 442910.1029/2001jd000659.

Yaeger M., Coopersmith E., Ye S., Cheng L., Viglione A., Sivapalan M. (2012) Exploring the physical controls of regional patterns of flow duration curves - Part 4: A synthesis of empirical analysis, process modeling and catchment classification. Hydrology and Earth System Sciences 16:4483-4498. DOI: 10.5194/hess-16-4483-2012.

Yao H.X., Scott L., Guay C., Dillon P. (2009) Hydrological impacts of climate change predicted for an inland lake catchment in Ontario by using monthly water balance analyses. Hydrological Processes 23:2368-2382. DOI: 10.1002/hyp.7347.

Yue S., Pilon P., Phinney B., Cavadias G. (2002) The influence of autocorrelation on the ability to detect trend in hydrological series. Hydrological Processes 16:1807-1829. DOI: 10.1002/hyp.1095.

Zaidman M.D., Keller V., Young A.R., Cadman D. (2003) Flow-duration-frequency behaviour of British rivers based on annual minima data. Journal of Hydrology 277:195-213. DOI: 10.1016/S0022-1694(03)00089-1.

Zealand C.M., Burn D.H., Simonovic S.P. (1999) Short term streamflow forecasting using artificial neural networks. Journal of Hydrology 214:32-48. DOI: Doi 10.1016/S0022-1694(98)00242-X.

Zhang Q., Qi T.Y., Singh V.P., Chen Y.D., Xiao M.Z. (2015) Regional Frequency Analysis of Droughts in China: A Multivariate Perspective. Water Resources Management 29:1767-1787. DOI: 10.1007/s11269-014-0910-x.

Zrinji Z., Burn D.H. (1994) Flood Frequency-Analysis for Ungauged Sites Using a Region of Influence Approach. Journal of Hydrology 153:1-21. DOI: Doi 10.1016/0022-1694(94)90184-8.

Appendix

Appendix A: Data

Appendix A1: List of stations used for temporal (T), spatial (S) and spatiotemporal (ST) analysis.

ID	Station	River	Operator	Area [km²]	T	S	ST
3183101	Sudendorf	Bever	NLWKN	122		x	
3346103	Schwege	Oedingberger Bach	NLWKN	44.7		x	
3437108	Beesten	Große Aa	NLWKN	407		x	
3611103	Auburg	Aubach	NLWKN	16.8		x	
3614105	Wissingen	Wierau	NLWKN	63.1		x	
3615104	Lüstringen	Hase	NLWKN	209		x	
3616102	Gretesch	Belmer Bach	NLWKN	25.8		x	
3618104	Haste	Nette	NLWKN	53.6		x	
3619101	Eversburg	Hase	NLWKN	344	x	x	
3623101	Georgsmarienhütte	Düte	NLWKN	47.3		x	x
3633101	Bramsche	Hase	NLWKN	682		x	
3634106	Neuenkirchen	Vördener Aue	NLWKN	78.3		x	
3637101	Bersenbrück	Hase	NLWKN	958		x	
3639112	Augustmühle	Hase	NLWKN	1165		x	
3643101	Gut Lage	Lager Hase	NLWKN	191		x	x
3644105	Telbrake	Füchteler Bach	NLWKN	65.3		x	
3644116	Addrup	Fladderkanal	NLWKN	228		x	
3647101	Uptloh	Lager Hase	NLWKN	506		x	x
3648102	Osteressen	Blocksmühlenbach	NLWKN	24.3		x	
3655101	Bunnen	Große Hase	NLWKN	1778		x	
3658105	Lodberger Straße	Loeninger Muehlenbach	UHV	70.9		x	
3659102	Düenkamp	Große Hase	NLWKN	1923		x	
3659103	Hölze	Große Hase	NLWKN	1927		x	
3669120	Aselage	Hahnenmoorkanal	NLWKN	295		x	
3671101	Herzlake	Hase	NLWKN	2246	x	x	
3672106	Augustenfeld	Südradde	NLWKN	81.7		x	x
3672108	Holter Mühle	Südradde	NLWKN	119		x	
3674108	Andrup-Lage	Lager Bach	NLWKN	127		x	
3675101	Haselünne	Hase	NLWKN	2549		x	
3676106	Lotten	Lotter Beeke	NLWKN	86.9		x	
3678104	Haverbeck	Hase-Altarm	NLWKN	58.6		x	
3687102	Westerlohmühlen I	Iradde	NLWKN	156		x	x
3691102	Bokeloh	Hase	NLWKN	2975		x	
3692102	Teglingen II	Kleine Beeke	NLWKN	23.6		x	
3692105	Teglingen I	Teglinger Bach	NLWKN	64.7		x	
3723105	Apeldorn	Nordradde	NLWKN	127		x	
3731104	Versen	Goldbach	NLWKN	36.1		x	
3732102	Wesuwe	Wesuwer Schloot	NLWKN	20.3		x	
3736102	Dankern	Mersbach	NLWKN	36.2		x	
3749101	Walchum	Walchumer Schlot	NLWKN	72.8		x	
3754101	Dersum	Dersumer Schlot	NLWKN	27.9		x	
3776102	Brual	Brualer Schlot	NLWKN	19.5		x	

3813104	Breddenberg	Ohe	NLWKN	59.76		x
3821102	Bischofsbrück	Marka	OOWV	25.9		x
3821105	Gut Ellerbrock	Marka	OOWV	68.9		x
3824102	Heetberg	Markhauser Moorgraben	OOWV	22.8		x
3829101	Neuscharrel	Marka	NLWKN	141		x	x
3863101	Neuburlage	Burlage-Langholter Tief	NLWKN	61.5		x
3881105	Stedingsmühlen	Soeste	NLWKN	75.4		x	x
3881113	Pehmertange	Soeste	OOWV	183		x
3881127	Kampe	Soeste	NLWKN	408		x
3882101	Aschhausen	Halfsteder Bäke	NLWKN	26.7		x
3884115	Ihorst	Große Norderbäke	NLWKN	51.9		x
3926104	Bagband	Bagbander Tief	NLWKN	47.6		x
4175102	Heldra	Heldrabach/Haselbach	WSD	4302	x	
4199103	Letzter Heller	Werra	WSD	5487	x	
4310101	Hann.- Münden	Weser	WSD	12440	x x	
4364106	Steimke A	Ahle	NLWKN	79.9		x	x
4364109	Allershausen	Rehbach I	NLWKN	52.3		x
4369101	Vernawahlshausen	Schwülme	NLWKN	281		x	x
4389102	Bodenfelde	Reiherbach I	NLWKN	30.8		x
4390101	Wahmbeck	Weser	WSD	13000	x	
4511101	Karlshafen	Weser	WSD	14790	x	
4536104	Holzminden Ha	Hasselbach	NLWKN	15.6		x
4538103	Negenborn	Forstbach	NLWKN	39		x	x
4545104	Oelkassen	Lenne	NLWKN	65.1		x	x
4569106	Welsede	Emmer	NLWKN	509		x	x
4572113	Afferde F	Fluthamel	NLWKN	180		x
4575104	Hameln-Wehrbergen	Weser	WSD	17090	x	
4599101	Vlotho	Weser	WSD	17620	x	
4665101	Wetter	Else	NLWKN	145		x
4713102	Porta	Weser	WSD	19160	x	
4723104	Achum	Bückeburger Aue	NLWKN	86.5		x
4741102	Volksdorf	Gehle	NLWKN	32.4		x
4767109	Heide OP	Große Aue	NLWKN	1016		x	x
4768111	Sieden	Siede	NLWKN	163		x
4769112	Steyerberg	Große Aue	NLWKN	1446		x
4813127	Warmenau Kl.A	Kleine Aller	NLWKN	141		x
4819102	Brenneckenbrück	Aller	NLWKN	1638	x	x
4821103	Altenau O	Okerstausee	Harzwasserwerke	31.2	x	
4821118	Oker	Abzucht	NLWKN	31.5		x
4821125	Vienenburg R	Radau	NLWKN	57.5		x	x
4823104	Schladen O	Oker	NLWKN	363	x	
4825109	Ohrum	Oker	NLWKN	813	x	
4826108	Wendessen	Altenau	NLWKN	118		x	x
4828104	Süpplingen	Schunter	NLWKN	55.1		x
4828126	Hondelage	Schunter	NLWKN	396		x
4828132	Niedersickte	Wabe	NLWKN	40.1		x	x
4828140	Harxbüttel	Schunter	NLWKN	592		x	x
4829102	Gr.Schwülper	Oker	NLWKN	1734	x	

4834114	Neuhaus S	Schwarzwasser	NLWKN	210		x	x
4836116	Bargfeld	Schmalwasser	NLWKN	40.2		x	
4836118	Luttern	Lutter	NLWKN	145		x	
4836128	Beedenbostel	Aschau	NLWKN	148		x	
4836129	Lachendorf	Lachte	NLWKN	433		x	x
4837102	Celle	Aller	WSD	4374	x x		
4843105	Broistedt	Fuhse	NLWKN	181		x	
4845103	Peine	Fuhse	NLWKN	360		x	x
4848111	Eltze	Erse	NLWKN	179		x	x
4849104	Wathlingen	Fuhse	NLWKN	812		x	
4854107	Immensen	Burgdorfer Aue	NLWKN	92.4		x	
4854112	Aligse	Burgdorfer Aue	NLWKN	180		x	
4863115	Poitzen	Írtze	NLWKN	198		x	
4866111	Reiningen	Wietze	NLWKN	159		x	
4872119	Meitze	Wietze	NLWKN	242		x	x
4872128	Wieckenberg	Wietze	NLWKN	399		x	x
4875102	Marklendorf	Aller	WSD	7209	x		
4881137	Gartemühle	Garte	NLWKN	86.3		x	
4881142	Göttingen	Leine	NLWKN	633		x	
4881166	Leineturm	Leine	NLWKN	990		x	
4882120	Westerode	Nathe	NLWKN	32.2		x	
4882129	Rollshausen S	Suhle	NLWKN	85.2		x	
4882139	Erikabrücke	Oder	Harzwasserwerke	43.6	x		
4882148	Kupferhütte	Lutter	Harzwasserwerke	16		x	
4882161	Pionierbrücke	Sieber	Harzwasserwerke	44.5	x	x	x
4882164	Herzberg I	Sieber	Harzwasserwerke	68.7	x	x	x
4882171	Lindau O	Oder	NLWKN	376		x	
4882173	Berka R	Rhume	NLWKN	895		x	x
4882176	Riefensbeek	Söse	Harzwasserwerke	24.2	x		
4882195	Berka S	Söse	NLWKN	210	x	x	x
4882196	Elvershausen	Rhume	NLWKN	1115		x	x
4882198	Northeim	Rhume	NLWKN	1176		x	
4884104	Relliehausen	Ilme	NLWKN	63.6		x	
4884108	Dassel	Spüligbach	NLWKN	37.2		x	
4884110	Markoldendorf	Ilme	NLWKN	149		x	x
4884122	Kuventhal	Krummes Wasser	NLWKN	61.8		x	x
4885116	Gandersheim	Gande	NLWKN	95.4		x	x
4885118	Greene	Leine	NLWKN	2916	x	x	
4885133	Godenau	Glene	NLWKN	37.9		x	
4885142	Thüste	Thüster Beeke	NLWKN	19.6		x	
4885148	Benstorf	Aue	NLWKN	26.9		x	
4885150	Mehle	Saale	NLWKN	136		x	x
4885154	Poppenburg	Leine	NLWKN	3463		x	
4885162	Hallerburg	Haller	NLWKN	104		x	
4886115	Lindthal	Innerste	Harzwasserwerke	98.1		x	
4886122	Hohenrode	Innerste	NLWKN	212		x	
4886129	Sehlde	Neile	NLWKN	59.9		x	
4886143	Gr.Rhüden	Nette	NLWKN	125		x	x
4886155	Derneburg	Nette	NLWKN	309		x	

4886164	Östrum	Riehe	NLWKN	74.3		x	
4886165	Bad Salzdetfurth	Lamme	NLWKN	127		x	
4886168	Heinde	Innerste	NLWKN	897		x	
4886171	Söhre	Beuster	NLWKN	32.5		x	
4886175	Groß Giesen	Innerste	NLWKN	1005		x	
4887121	Oberricklingen	Ihme	NLWKN	86.7		x	
4887123	Herrenhausen	Leine	WSD	5304	x	x	
4888121	Rodenberg	Rodenberger Aue	NLWKN	154		x	x
4888137	Düendorf	Südaue	NLWKN	189		x	
4888139	Wunstorf	Westaue	NLWKN	558		x	
4889102	Neustadt	Leine	WSD	6043		x	
4889117	Averhoy	Auter	NLWKN	122		x	
4889128	Schwarmstedt	Leine	WSD	6443	x	x	
4892106	Hörsten	Meiße	NLWKN	79.6		x	x
4894107	Tetendorf	Böhme	NLWKN	110		x	
4894119	Brock	Böhme	NLWKN	285		x	x
4894120	Westerharl Süd	Bomlitz	NLWKN	60.8		x	
4894136	Hollige	Böhme	NLWKN	538		x	
4895101	Rethem	Aller	WSD	14730	x x		
4896111	Frankenfelderbruch	Alpe	NLWKN	164		x	x
4898107	Lehringen	Lehrde	NLWKN	100		x	x
4911101	Intschede	Weser	WSD	37720	x x		
4921104	Okel S	Süstedter Bach	UHV	48.4		x	
4921109	Sudweyhe	Süstedter Bach	NLWKN	125		x	
4922106	Steimke H	Hache	NLWKN	84.3		x	
4922107	Barrien	Hache	NLWKN	110		x	
4922109	Weyhe	Hache	NLWKN	115		x	
4924104	Fahrenhorst	Hombach	Harzwasserwerke	43.6		x	
4924105	Leeste Homb.	Hombach	Harzwasserwerke	50		x	
4926109	Kirchseelte	Klosterbach	NLWKN	108		x	
4941106	Wümme	Wümme	NLWKN	93.1		x	x
4941115	Lauenbrück F	Fintau	NLWKN	96		x	
4941116	Lauenbrück B 75	Wümme	NLWKN	248		x	
4942104	Westervesede Veerse	Veerse	NLWKN	61.5		x	
4942108	Veerse	Veerse	NLWKN	110		x	
4943104	Rotenburg	Wümme	NLWKN	469		x	
4944110	Hastedt	Rodau	NLWKN	137		x	
4944120	Worth	Wiedau	NLWKN	149		x	
4945108	Hellwege-Schleuse	Wümme	NLWKN	908		x	x
4945122	Sottrum	Wieste	NLWKN	94		x	
4946102	Schnakenmühlen	Wörpe	NLWKN	22.8		x	
4946105	Grasberg	Wörpe	NLWKN	93.8		x	x
4948105	Ahrensdorf	Giehler B	NLWKN	79.6		x	
4961107	Angelbeck	Hunte	NLWKN	42		x	
4961112	Bohmte	Hunte	NLWKN	179		x	
4961121	Hunteburg	Elze	NLWKN	54.6		x	
4961125	Schäferhof	Hunte	NLWKN	403		x	
4961177	Dümmer Auslauf	Hunte	NLWKN	426		x	

4963101	Hoopen OP	Hunte	NLWKN	772	x	x
4964115	Düste	Wagenfelder Aue	NLWKN	200	x	
4965116	Colnrade OP	Hunte	NLWKN	1318	x	x
4965142	Huntlosen	Hunte	NLWKN	1714	x	x
4966102	Gut Lethe	Lethe	OOWV	22.6	x	
4966112	Oberlethe	Lethe	NLWKN	160	x	
4968102	Düwelshoop	Haaren	NLWKN	21.5	x	
4976103	Dorfhagen	Drepte	NLWKN	40	x	
4983103	Deelbrügge	Lune	NLWKN	98.4	x	
4984103	Stubben	Billerbeck	NLWKN	38.4	x	
4986101	Dohren	Dohrener Bach	NLWKN	11.1	x	
4992106	Köhlener Brücke	Geeste	NLWKN	77.3	x	
4992112	Hainmühlen	Obere Wittgeeste	NLWKN	15.1	x	
5648104	Zorge	Zorge	NLWKN	31.9	x	
5648107	Wieda	Wieda	NLWKN	14.1	x	
5934130	Köhlen	Köhlener Mühlenbach	NLWKN	41	x	
5935105	Neu Darchau	Elbe	WSD	132000	x	
5941119	Niendorf II S	Stederau	NLWKN	285	x	
5941123	Niendorf II W	Wrestedter Bach	NLWKN	60	x	
5942120	Hansen	Gerdau	NLWKN	308	x	
5945114	Klein Hesebeck	Röbbelbach	NLWKN	148	x	
5945124	Bargdorf	Barum-Bienenbütteler Mühlenbach	NLWKN	68	x	
5945125	Bienenbüttel	Ilmenau	NLWKN	1434	x	x
5945139	Häcklingen	Hasenburger Mühlenbach	NLWKN	88.3	x	
5946106	Marienau	Neetze	NLWKN	85	x	
5946112	Süttorf	Neetze	NLWKN	174	x	
5948130	Wulfsen	Luhe	Hamburger Wasserwerke	66	x	
5952115	Döhle S.Aue	Schmale Aue	Hamburger Wasserwerke	30	x	
5952124	Marxen	Schmale Aue	Hamburger Wasserwerke	157	x	
5952127	Jehrden	Seeve	NLWKN	408	x	x
5958103	Langeloh	Este	NLWKN	40.9	x	x
5958112	Emmen	Este	NLWKN	184	x	x
5963101	Oersdorf	Aue (Lühe)	NLWKN	28.1	x	
5965101	Daudieck	Aue (Lühe)	NLWKN	132	x	
5972105	Schwinge	Schwinge	NLWKN	61.1	x	
5972109	Hagen	Steinbeck	NLWKN	16	x	
5981111	Ramshausen	Ramme	NLWKN	72.3	x	x
5981121	Weertzen	Oste	NLWKN	307	x	
5982105	Brüttendorf	Aue-Mehde	NLWKN	74.7	x	
5983105	Twistenbostel	Twiste	NLWKN	61	x	
5983110	Rockstedt	Oste	NLWKN	611	x	x
5983115	Minstedt	Oste	NLWKN	680	x	
5985108	Gräpel	Gräpeler Mühlenbach	NLWKN	13	x	
9286107	Samern	Vechte	NLWKN	405	x	

9286120	Drievorden	Ahlder Bach	NLWKN	34.6		x	
9286121	Engden I	Engdener Bach	NLWKN	18.8		x	
9286127	Wehr Neuenhaus	Vechte	NLWKN	702	x	x	x
9286132	Holt und Haar	Puntbecke	NLWKN	17.1		x	
9286139	Lage gesamt		NLWKN	611		x	
9286144	Tinholt W	Vechte	NLWKN	1395		x	
9286161	Haselaar	Emlichheimer Entlastungskanal	NLWKN	93.7		x	
9286162	Emlichheim	Vechte	NLWKN	1731	x	x	

Appendix A2: GRDC data set used for time-series extrapolation.

ID	Station	River	Country	Area [km²]
6119030	OLORON-SAINTE-MARIE (OLORON-SNCF)	GAVE D'OLORON	FR	1085
6140400	DECIN	ELBE RIVER	CZ	51123
6142100	MORAVICANY	MORAVA	CZ	1559
6142200	BRATISLAVA	DANUBE RIVER	SK	131331
6229100	VASSBOTTEN	ENNINGDALSAELVEN	SE	624
6229500	VARGOENS KRV	GOETA AELV	SE	46886
6233100	ASBRO 3	VISKAN	SE	2160
6233170	AENGABAECKS KRV	LAGAN (SWEDEN)	SE	5480
6233205	ERSBO	VAESTERDALAELVEN	SE	1104
6233250	TORSEBRO KRV (POWERSTATION)	HELGE A	SE	3665
6233366	SKAERSBODA	SOLGENAN	SE	157
6233510	GRANAKER	VINDELAELVEN	SE	11851
6233650	SOLLEFTEA KRV	ANGERMANAELVEN	SE	30638
6233680	SORSELE 2	VINDELAELVEN	SE	6056
6233750	BODENS KRV (+ VATTENVERK, TRANGFORS)	LULEAELVEN	SE	24924
6233901	KALLIO 2	MUONIOAELVEN, MUONIONJOKI	SE	14477
6233910	KUKKOLANKOSKI OEVRE	TORNEAELVEN, TORNIONJOKI	SE	33930
6335020	REES	RHINE RIVER	DE	159300
6335060	KOELN	RHINE RIVER	DE	144232
6335291	GAILDORF	KOCHER	DE	726
6335301	SCHWEINFURT - NEUER HAFEN	MAIN	DE	12715
6335410	OBERLAUCHINGEN	WUTACH	DE	627
6335500	WUERZBURG	MAIN	DE	14031
6336050	COCHEM	MOSELLE RIVER	DE	27088
6337514	BODENWERDER	WESER	DE	15924
6338120	GREVEN	EMS	DE	2842
6340120	DRESDEN	ELBE RIVER	DE	53096
6340140	BARBY	ELBE RIVER	DE	94060

6340150	WITTENBERGE	ELBE RIVER	DE	123532
6340430	PORSCHDORF 1	LACHSBACH	DE	267
6340620	WECHSELBURG 1	ZWICKAUER MULDE	DE	2107
6342200	KEMPTEN	ILLER	DE	955
6342513	LANDSBERG	LECH	DE	2295
6342620	REGENSTAUF	REGEN	DE	2658
6342800	HOFKIRCHEN	DANUBE RIVER	DE	47496
6342900	ACHLEITEN	DANUBE RIVER	DE	76653
6342930	KOCHEL	LOISACH	DE	685
6343100	WASSERBURG	INN	DE	11983
6343500	BURGHAUSEN	SALZACH	DE	6649
6343560	UNTERJETTENBERG	SAALACH	DE	927
6435060	LOBITH	RHINE RIVER	NL	160800
6607650	KINGSTON	THAMES	GB	9948
6607651	KINGSTON (NATURALISED DISCHARGE)	THAMES	GB	9948
6607830	FEILDES WEIR	LEE RIVER	GB	1036
6731165	VIKSVATN	GAULAR	NO	505
6731170	LOVATN	LOELV	NO	231
6731175	HORNINDALSVATN	EIDSELV	NO	378
6731200	BULKEN	VOSSO	NO	1102
6731400	LANGNES	GLAMA	NO	40540
6731401	ELVERUM	GLAMA	NO	15426
6731403	SOLBERGFOSS	GLAMA	NO	40540
6731450	LOSNA	LOSNA	NO	11210
6731570	NYBERGSUND	KLARA	NO	4410
6731610	FUSTVATN	FUSTA	NO	520
6731907	MALANGSFOSS	MAALSELV	NO	3239
6742200	ORSOVA (1971:DROBATA-TURNU SEVERIN)	DANUBE RIVER	RO	576232
6854104	MUROLEENKOSKI	KOKEMAENJOKI	FI	6102
6854107	KITUSJAERVI - OUTLET	KOKEMAENJOKI	FI	546
6854590	LAKE LENTUA OUTLET	OULUJOKI	FI	2045
6854600	RAASAKKA (NEAR THE MOUTH)	IIJOKI	FI	14191
6854755	KARUNKI	TORNEAELVEN, TORNIONJOKI	FI	NA
6854900	SKATILA (LANSORSUND)	KYRONJOKI	FI	4833
6855280	VUOSJARVI, HUOPANANKOSKI	KYMIJOKI	FI	2186
6935051	BASEL, RHEINHALLE	RHINE RIVER	CH	35897
6935052	BASEL, SCHIFFLAENDE	RHINE RIVER	CH	35905
6935054	REKINGEN	RHINE RIVER	CH	14718
6935055	NEUHAUSEN, FLURLINGERBRUECKE	RHINE RIVER	CH	11887
6935060	MONTIER (LA CHARRUE)	BIRSE	CH	183
6935145	DOMAT/EMS	RHINE RIVER	CH	3229
6935300	UNTERSIGGENTHAL, STILLI	AARE	CH	17601
6935310	MELLINGEN	REUSS	CH	3382

6935320	EMMENMATT	EMME	CH	443
6935400	ANDELFINGEN	THUR (CH)	CH	1696
6935412	APPENZELL	SITTER	CH	74
6939050	CHANCY, AUX RIPES	RHONE	CH	10323
6939200	PORTE DU SCEX	RHONE	CH	5244
6939540	GSTEIG	LUETSCHINE	CH	379
6943100	MARTINSBRUCK	INN	CH	1945
6948100	BELLINZONA	TICINO	CH	1515
6974150	SMALININKAI	NEMUNAS - NEMAN	LT	81200

Appendix B: Temporal low flow modeling

Appendix B1: Non-stationary GEV extrapolation results for the Q_{95}, Q_{80} and NM30Q.
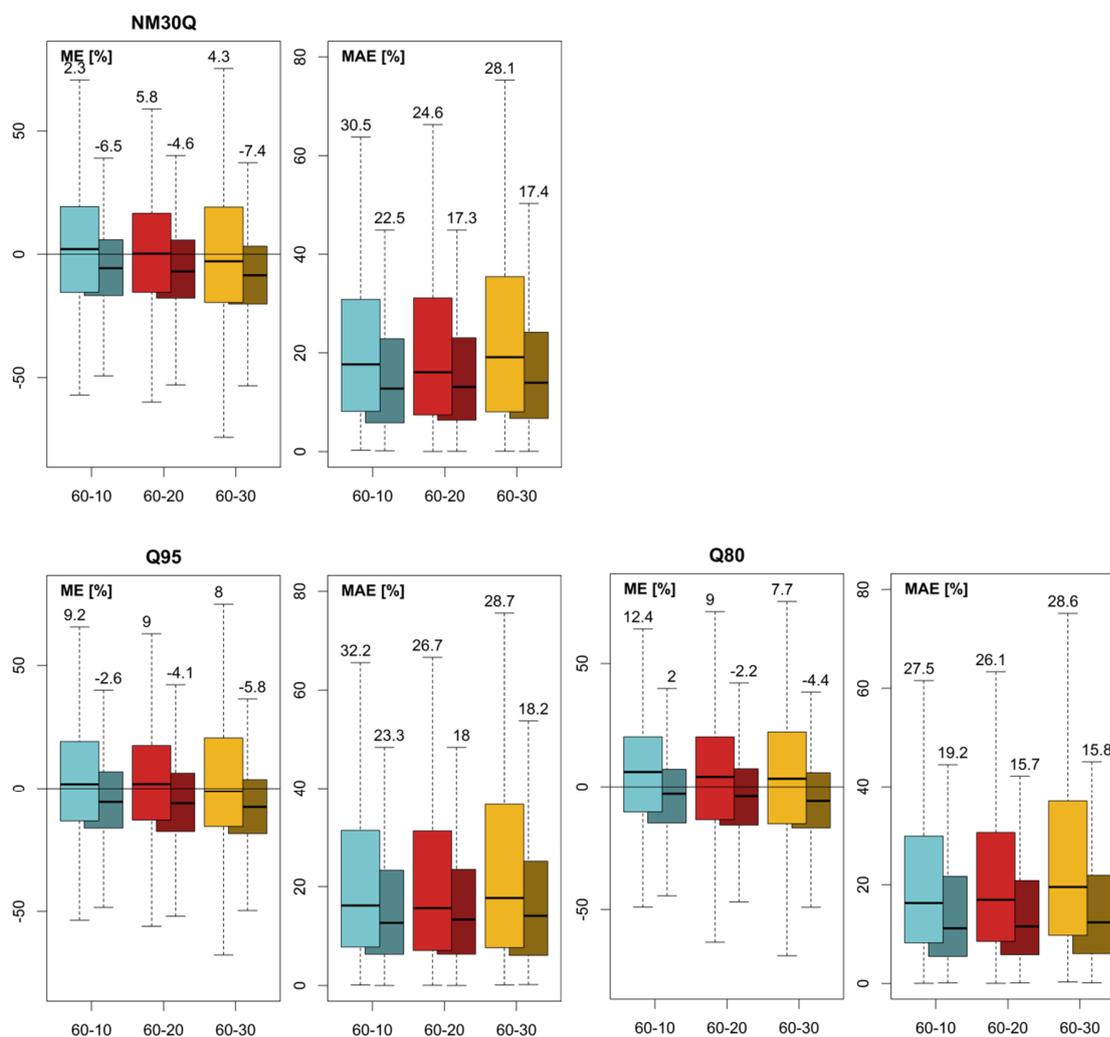
Appendix B2: SVR extrapolation results for various indices.
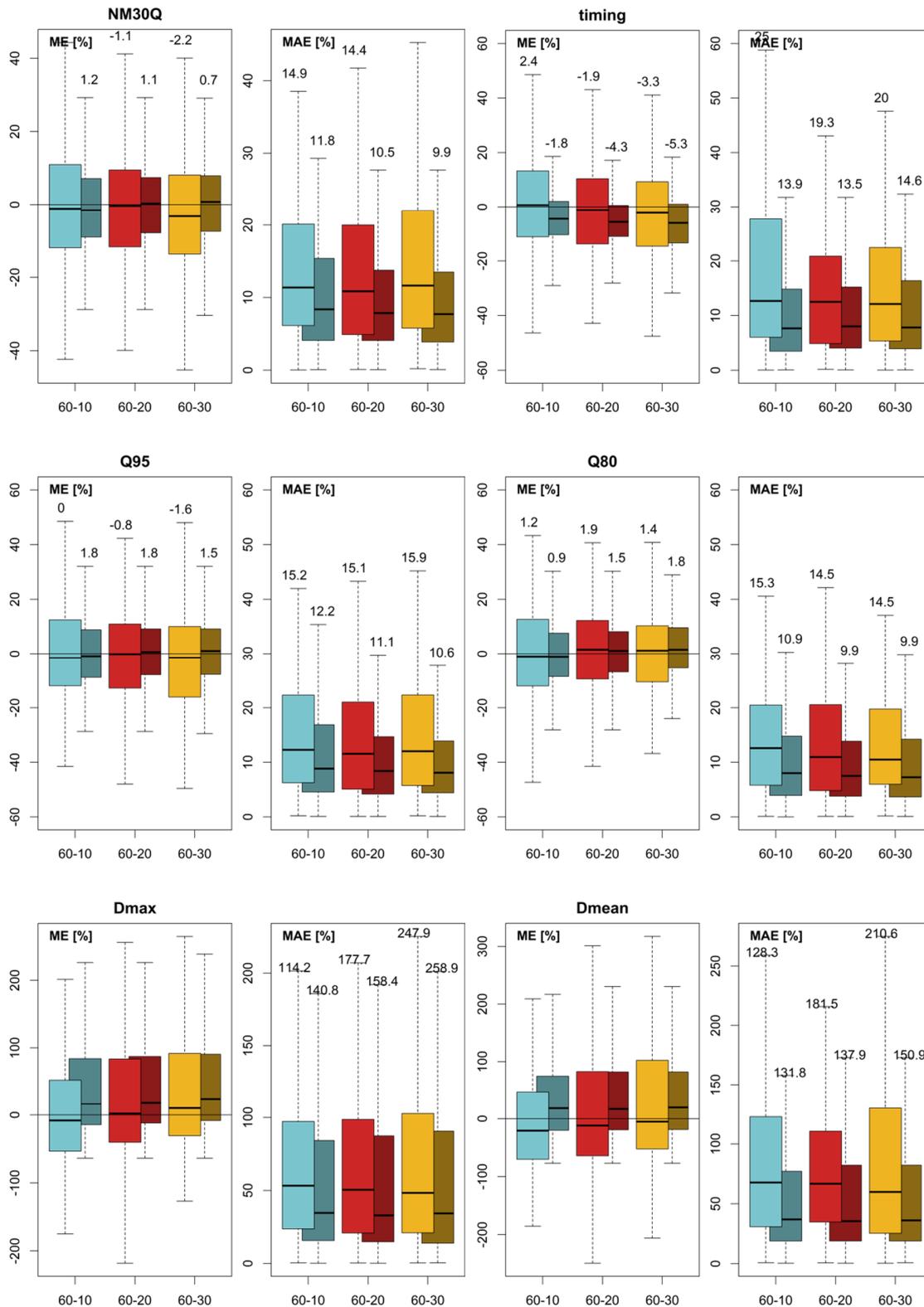
Appendix B3: GLS-R station models for estimation of annual NM7Q for 2 runs at 28 stations including SVR parameters.

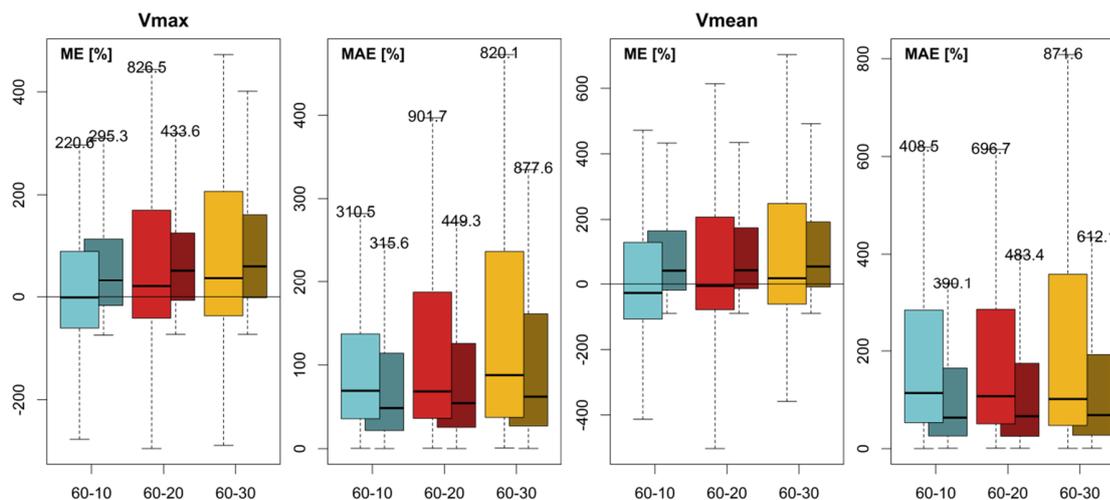| 3619101		
	[1]	1.28 Intercept + 0.15 SPEI 6-0 + 0.23 SPEImin 12
		Epsilon: 0.5 Cost: 11
	[2]	2.88 Intercept + 0.45 PETPmean 6-0 - 0.2 Gmean 3-11
		Epsilon: 0.5 Cost: 1
3671101		
	[1]	5.76 Intercept + 1.78 SPEImin 12 + 0.63 SPEI 6-0
		Epsilon: 0.8 Cost: 16
	[2]	3.79 Intercept + 1.47 SPEI 3-0 + 1.22 P70 6-12
		Epsilon: 0.7 Cost: 16
4175102		
	[1]	4.8 Intercept + 2.13 SPEI 6-0 + 2.23 P80 6-2
		Epsilon: 0.8 Cost: 16
	[2]	11.09 Intercept + 11.41 PETPmean 6-0 - 5.42 Pmean 6-0
		Epsilon: 0.8 Cost: 11
4199103		
	[1]	17.56 Intercept + 3.23 SPEI 3-1 + 2.7 SPEI 6-0
		Epsilon: 0.5 Cost: 6
	[2]	16.83 Intercept + 3.44 SPEI 6-0 + 14.81 P40 3-2
		Epsilon: 0.2 Cost: 16
4310101		
	[1]	72.15 Intercept - 12.63 DSDWSDmean 6-1 + 8.39 SPEI 6-1
		Epsilon: 0.3 Cost: 6
	[2]	19.16 Intercept + 9.29 PETPmean 6-3 + 11.3 PETPmean 6-1
		Epsilon: 0.4 Cost: 11
4390101		
	[1]	56.11 Intercept + 10.34 SPEI 6-2 + 8.25 SPEI 3-1
		Epsilon: 0.1 Cost: 6
	[2]	33.72 Intercept + 6.46 SPEI 3-3 + 14.23 PETPmean 6-1

		Epsilon: 0.1 Cost: 11
4511101		
	[1]	103 Intercept - 25.55 DSDWSDmean 6-1 + 9.56 SPEI 6-3
		Epsilon: 0.1 Cost: 6
	[2]	20.63 Intercept + 14.05 PETPmean 6-1 + 11.73 PETPmean 6-3
		Epsilon: 0.4 Cost: 6
4575104		
	[1]	108.57 Intercept + 13.52 SPEI 6-1 - 20.41 DSDWSDmean 6-1
		Epsilon: 0.1 Cost: 11
	[2]	72.93 Intercept + 13.3 SPEI 6-0 + 8.23 SPEI 3-2
		Epsilon: 0.2 Cost: 16
4713102		
	[1]	109.71 Intercept + 19.5 SPEI 6-1 - 15.04 DSDWSDmean 3-2
		Epsilon: 0.1 Cost: 6
	[2]	55.99 Intercept + 17.13 PETPmean 6-0 + 15.45 SPEI 6-1
		Epsilon: 0.1 Cost: 16
4819102		
	[1]	-0.43 Intercept + 1.39 Pmean 6-0 +
		Epsilon: 0.8 Cost: 11
	[2]	7.19 Intercept - 0.91 SPImin 12- - 8.45 ETPmean 12-0
		Epsilon: 0.1 Cost: 6
4821103		
	[1]	0.11 Intercept + 0.04 SPEImin 12
		Epsilon: 0.7 Cost: 6
	[2]	0.05 Intercept + 0.04 P50 3-1 +
		Epsilon: 0.6 Cost: 11
4823104		
	[1]	1.93 Intercept + 0.57 SPEI 3-3
		Epsilon: 1 Cost: 6
	[2]	1.74 Intercept + 0.86 P40 3-1
		Epsilon: 0.9 Cost: 6
4825109		
	[1]	-0.59 Intercept + 0.02 Pmean 6-1 + 1.13 Pmean 6-2
		Epsilon: 0.5 Cost: 1
	[2]	0.88 Intercept + 0.91 PETPmean 6-1
		Epsilon: 0.5 Cost: 1
4829102		
	[1]	1.68 Intercept + 1.22 Pmean 3-2 + 0.79 SPEI 6-0
		Epsilon: 0.8 Cost: 16
	[2]	10.06 Intercept + 1.3 SPEI 6-2 - 5.98 ETPmean 6-1
		Epsilon: 0.3 Cost: 6
4837102		
	[1]	10.61 Intercept + 2.21 SPEI 6-2 + 2.41 SPEImin 12-
		Epsilon: 0.1 Cost: 6
	[2]	3.71 Intercept + 2.22 PETPmean 3-2 + 3.78 PETPmean 6-1
		Epsilon: 0.3 Cost: 6
4875102		
	[1]	17.19 Intercept - 0.45 DSDWSDmean 6-1 + 6.68 SPEI 6-1
		Epsilon: 0.1 Cost: 6
	[2]	-2.48 Intercept + 7.99 PETPmean 6-2 + 1 WSDmax 3-13

Epsilon: 0.1 Cost: 1

4882139		
	[1]	0.17 Intercept + 0.07 SPEImin 12
		Epsilon: 0.4 Cost: 6
	[2]	0.1 Intercept + 0.1 P40 3-1
		Epsilon: 0.5 Cost: 16
4882161		
	[1]	0.37 Intercept + 0.11 indexc
		Epsilon: 0.1 Cost: 11
	[2]	0.51 Intercept + 0.05 SPEI 3-0 - 0.16 DSDWSDmean 6-0
		Epsilon: 0.6 Cost: 1
4882164		
	[1]	0.47 Intercept + 0.16 SPEImin 12
		Epsilon: 0.6 Cost: 11
	[2]	0.37 Intercept + 0.14 SPEI 3-1
		Epsilon: 1 Cost: 16
4882176		
	[1]	0.11 Intercept + 0.04 SPEImin 12
		Epsilon: 0.2 Cost: 1
	[2]	0.03 Intercept + 0.02 PETPmean 3-1 + 0.09 P40 6-1
		Epsilon: 0.2 Cost: 11
4882195		
	[1]	-0.31 Intercept + 0.3 P80 6-3 + 0.29 P70 6-0
		Epsilon: 0.5 Cost: 16
	[2]	1.58 Intercept + 0.4 SPEI 6-0
		Epsilon: 0.4 Cost: 16
4885118		
	[1]	8.68 Intercept + 2.91 SPEI 6-1 + 2.31 Pmean 3-2
		Epsilon: 0.1 Cost: 11
	[2]	8.63 Intercept + 1.47 SPEI 6-0 + 3.11 PETPmean 6-1
		Epsilon: 0.7 Cost: 16
4887123		
	[1]	21.76 Intercept + 4.09 SPEI 6-1 + 3.43 SPEImin 12
		Epsilon: 0.2 Cost: 1
	[2]	8.37 Intercept + 3.89 Pmean 3-1 + 3.96 SPEI 6-0
		Epsilon: 0.3 Cost: 16
4889128		
	[1]	26.96 Intercept + 4.87 SPEI 6-1 + 3.84 SPEImin 12-
		Epsilon: 0.2 Cost: 16
	[2]	25.06 Intercept + 3.98 SPEI 6-0 + 4.33 SPEI 3-1
		Epsilon: 0.1 Cost: 16
4895101		
	[1]	53.6 Intercept + 11.01 SPEI 6-1 + 8.96 SPEImin 12
		Epsilon: 0.1 Cost: 1
	[2]	14.41 Intercept + 24.8 indexc
		Epsilon: 0.1 Cost: 16
4911101		
	[1]	148.94 Intercept + 35.77 SPEI 6-1 + 18.09 SPEI 3-1
		Epsilon: 0.1 Cost: 6
	[2]	133.28 Intercept + 23.09 SPEI 6-1 + 26.08 SPEI 3-1

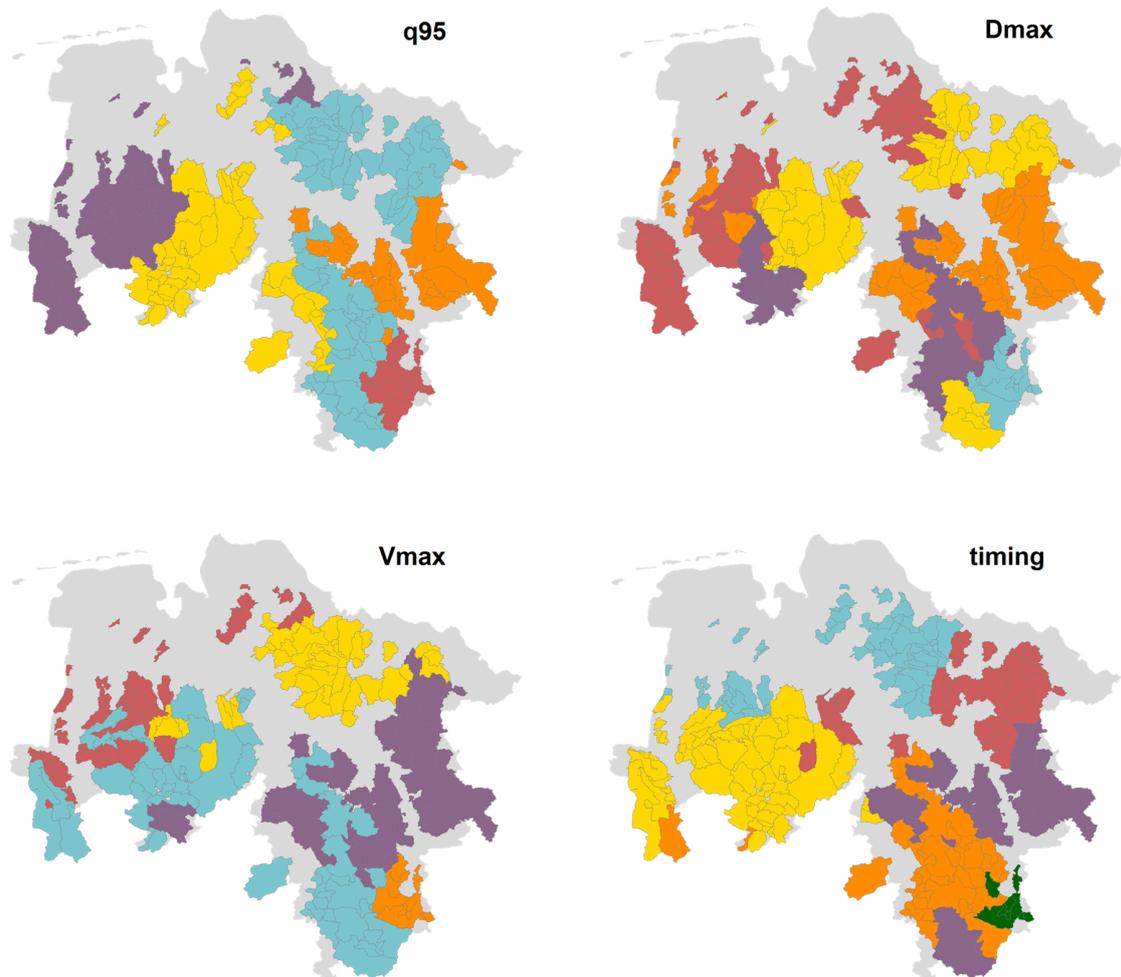|		Epsilon: 0.3 Cost: 6
9286127		
	[1]	1.28 Intercept + 0.62 indexc
		Epsilon: 0.5 Cost: 16
	[2]	-0.06 Intercept + 0.61 SPEImin 1210.29 ETPmean 3-9
		Epsilon: 0.8 Cost: 16
9286162		
	[1]	2.74 Intercept + 1.45 indexc
		Epsilon: 0.3 Cost: 11
	[2]	3.54 Intercept + 1.1 SPEImin 12
		Epsilon: 0.5 Cost: 1

Appendix C: Spatial low flow modeling

Appendix C1: K-means clusters for the low flow indices q95, Dmax, Vmax and low flow timing.

Appendix C2: Scatterplots of the regionally estimated low flow indices q95, Dmax, Vmax and timing vs. the respective observation.
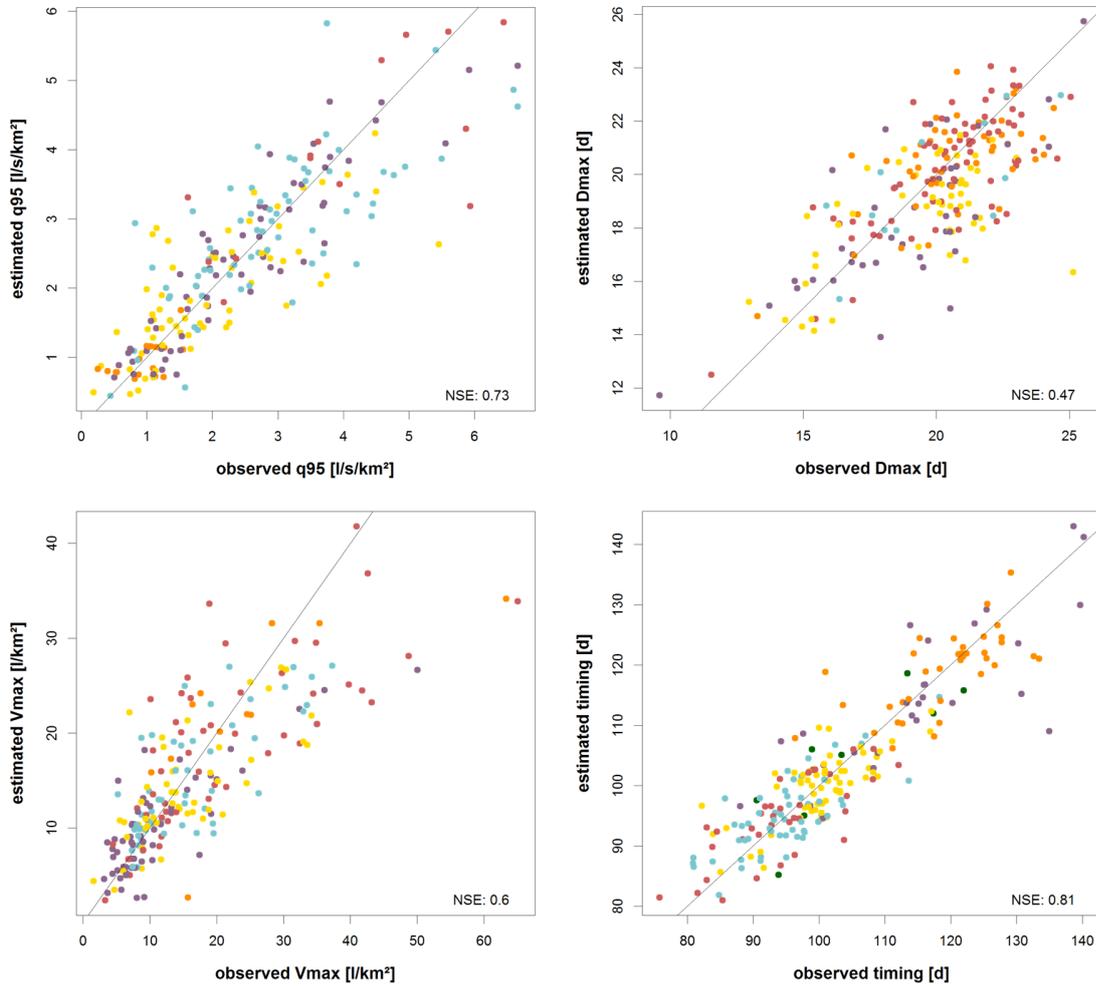
Appendix C3: Regional MLR models for the low flow indices q95, Dmax, Vmax and timing.

Q_{95}

[1]   -8.03 Intercept + 3.55 Minimum SPI 3-13 + 1.08 Maximum P80 3-3
      + 0.82 Maximum DSDWSDmax 6-2 + 0.52 Minimum HWDmean 3-8 + 0.03 Urban
      + 0.55 Range P60 12-4-

[2]   -3.33 Intercept - 0.62 Minimum Tmin 12-7 - 0.63 Maximum P90 6-12
      + 0.46 Range P80 6-10 + 1.58 Mean WSD 3-2

[3]   -22.87 Intercept + 1.35 Range Pmean 3-8 + 0.07 Urban 0.58 Minimum G 3-0
      + 2.31 Maximum HWDmean 12-3 - 0.68 Maximum WSDmean 3-17

[4]   4.05 Intercept - 0.68 Maximum T 3-11 - 3.26 Mean HWD 6-12 + 0.4 Minimum DSDmean 12-9

[5]   -2.4 Intercept + 0.21 Maximum Pmax 6-1

D_{max}

[1]   2.38 Intercept - 0.05 Range DSDmax 3-5 + 0.62 Minimum HWDmean 6-10
      + 1.48 Maximum HWDmean 3-3 + 0.49 Range WSDmean 3-13

[2]   4.76 Intercept - 0.22 Range DSD/WSDmax 3-13 + 0.39 Minimum P50 6-13
      + 730.45 Maximum P40 12-1 - 195.9 Maximum P40 6-10 + 0.37 Range WSDmean 12-7

[3]   8.44 Intercept - 0.17 Maximum WSDmax 12-10 - 0.8 Maximum DSDmean 3-12
      - 0.03 Urban - 0.34 Range DSDmean 6-2 - 0.98 Maximum P60 6-1
      + 0.13 Maximum WSDmax 12-11 + 0.47 Range WSDmean 6-11
      - 0.56 Range WSDmean 12-4 + -0.2 Minimum HWDmean 6-11

[4]   3.02 Intercept + 0.42 Range P80 6-11 - 5.1 Maximum P50 12-4

[5]   6.22 Intercept - 1.76 Range WSDmean 12-11

V_{max}

[1]   -14.85 Intercept + 0 Umfang 9.66 Mean HWD 3-11 - 10.66 Range P50 12-8
      + 5.94 Range P50 6-0 + 1.3 Minimum P60 6-6 - 6.15 Maximum DSDWSDmean 3-4
      + 1.2 Minimum P70 3-15 + 1.69 Range DSDWSDmean 3-1

[2]   -20.74 Intercept + 0.72 Range DSD/WSDmean 3-9 - 8.85 Maximum aSPEI 3-2
      - 4.78 Maximum DSDWSDmean 3-5 + 7.8 Mean HWD 6-15
      - 1.35 Range SPEI 6-10 + 0.08 Minimum WSD 6-1

[3]   22.33 Intercept - 1.58 Maximum DSDmax 3-0 - 2.03 Range P80 6-2
      - 9.03 Maximum P60 6-1 + 415.56 Maximum P50 3-3 - 6.04 Maximum P60 3-6

[4]   0.35 Intercept + 0.82 Range P70 3-3

[5]   36.91 Intercept - 13.89 Maximum DSDmean 3-14

timing

[1]   7.48 Intercept + 0.82 Mean DSDmax 3-12 - 2.25 Range Tmean 12-8

[2]   2.33 Intercept + 0.76 Mean DSD/WSDmax 6-14 - 1.69 Minimum P50 6-7
      + 4.28 Mean HWD 6-9 - 0.26 Range P90 3-9 + 0.37 Maximum WSDmax 3-1
      + 0.19 Range P90 12-16 + 0.31 Range DSDmean 12-9

[3]   43.94 Intercept - 1.41 Minimum G 6-6 - 2.79 Maximum HWDmean 12-3
      + 17.38 Maximum P50 6-8 - 0.25 Range P90 6-3

[4]   8.46 Intercept + 0.14 Range HWDmax 6-10 + 0.01 Silt

        - 1.06 Range WSDmean 6-3 - 0.07 Range WSDmax 6-13

[5]   9.4 Intercept - 14.8 Mean SPEI 3-11

[6]   5.0 Intercept + 1.05 Minimum Tmin 3-0 – 0.14 Maximum DSDmean 3-6

Appendix D: Spatio-temporal low flow modeling

Appendix D1: Spatiotemporal models for direct NM7q (ST-1) and L-moment (ST-3) prediction for the 5 k-means clusters.

ST-1

[1]   10.19 Intercept + 1.01 PETPmean$_{6-1}$ - 7.56 *Minimum DSDmean$_{12-1}$
         + 1.24 *Maximum DSDWSDmax$_{6-10}$ + 2.33 *Mean HWDmax$_{3-3}$

[2]   4.91 Intercept + 0.77 PETPmean$_{6-1}$ + 0.22 WSDmean$_{3-6}$ - 1.41 Range DSDmean$_{3-14}$

[3]   2.54 Intercept + 0.58 PETPmean$_{6-1}$ - 1.05 *Minimum DSDmax$_{6-8}$ + 0.24 SPEI$_{6-11}$ + 0.32 P50$_{3-2}$
         + 36.93 *Minimum P50$_{6-1}$

[4]   0.7 Intercept + 0.6 SPEI$_{3-1}$ + 0.28 *Maximum HWDmax$_{3-9}$ - 2.27 ETPmean$_{12-3}$
         - 0.09 *Range WSDmax$_{12-3}$ + 0.61 P50$_{12-13}$

[5]   -10.14 Intercept + 0.48 SPEI$_{12}$- + 5.28 *Mean HWDmax$_{3-8}$ - 0.72 *Minimum HWDmean$_{3-8}$
         + 0.31 P50$_{6-0}$ + 0.17 SPEI$_{3-5}$

ST-3

L1

[1]   -2.18 Intercept + 0.33 *Maximum Pmax$_{12-0}$ - 0.67 *Maximum DSD/WSDmax$_{6-0}$
         - 1.1 *Range DSD/WSDmean$_{3-1}$ - 0.29 Mean HWDmean$_{3-14}$

[2]   -11.14 Intercept + 3.23 *Minimum PETPmean$_{3-0}$ - 0.01 *Minimum HWDmax$_{6-0}$

[3]   -1.25 Intercept + 0.53 *Maximum Pmax$_{6-10}$ - 1.61 *Maximum DSD/WSDmax$_{12-17}$
         + 0.3 *Maximum Pmax$_{3-0}$ + 0.44 Mean SPEI$_{3-2}$ - 5.19 *Mean HWD$_{6-10}$

[4]   1.54 Intercept + 0.53 Mean SPEI$_{3-2}$ + 0.24 Mean HWD$_{3-9}$ - 0.35 *Range P90 $_{6-0}$

[5]   -7.40 Intercept + 4.43 *Mean HWDmax$_{3-8}$ - 0.09 *Mean DSDmax$_{3-3}$

L2

[1]   -2.57 Intercept + 0.46 Range SPEI$_{3-1}$ + 0.1 Range WSDmax$_{3-13}$
         + 0.18 Maximum HWDmean$_{3-13}$ + 0.05 Range WSDmean$_{3-8}$

[2]   -2.78Intercept + 0.04 Mean Pmax$_{6-3}$ - 0.2 Mean Tmin$_{3-6}$ - 0.91 Minimum HWDmean$_{6-7}$

[3]   -4.32Intercept + 1.25 Mean SPEI$_6$ + 0.06 Mean Gmean$_{6-1}$ + 0.45 *Mean WSDmax$_{6-9}$

[4]   -0.96 Intercept + 0.9 Mean SPEI$_6$ + 0.13 Mean P70$_{12-5}$ + 0.11 Range Tmean$_{6-3}$

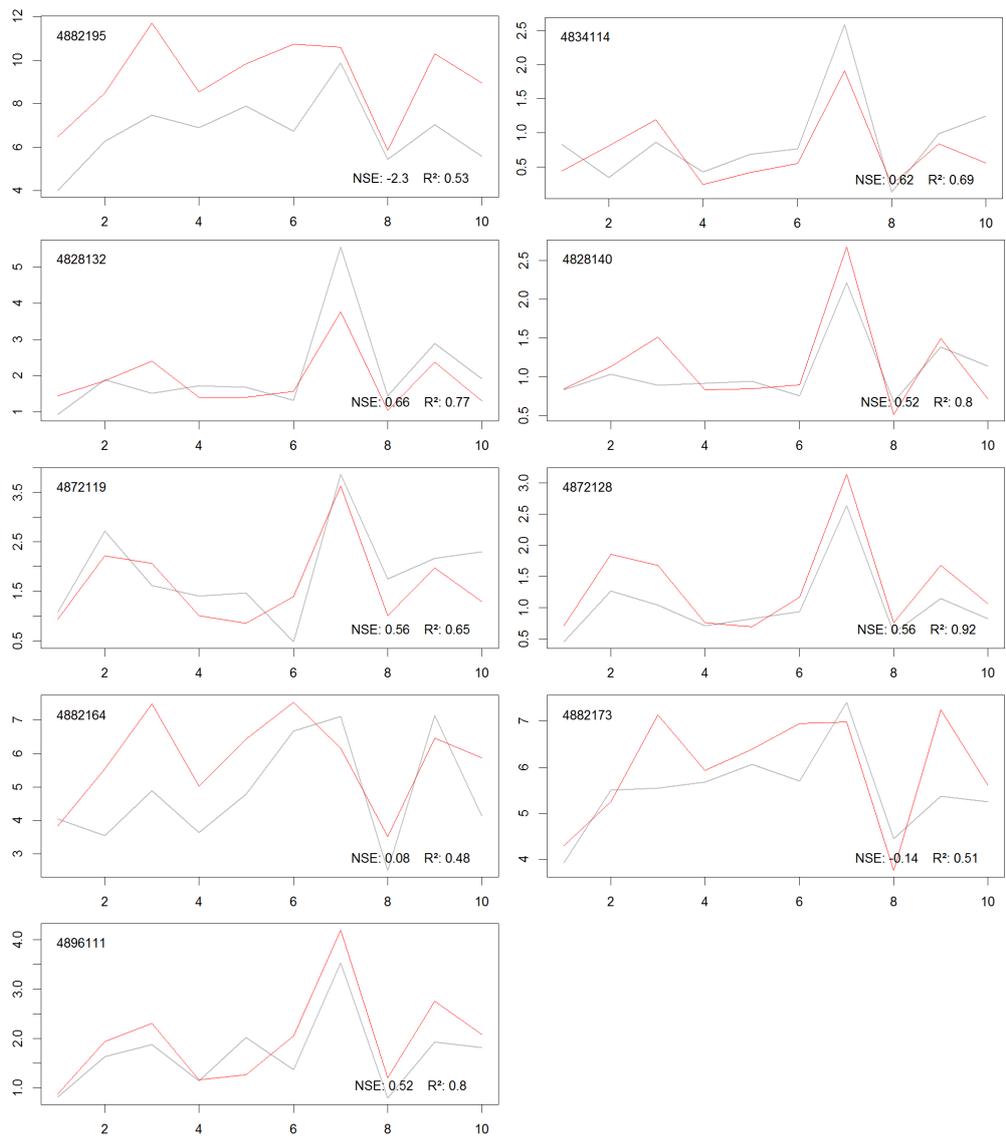[5]   -0.97 Intercept + 0.45 Mean SPEI$_{3-3}$ + 0.23 Mean Gmean$_{3-3}$

* stationary variables

Appendix D2: Simulated NM7q during combined cross and split validation for the period 1995-2005 using the ST-1 model.
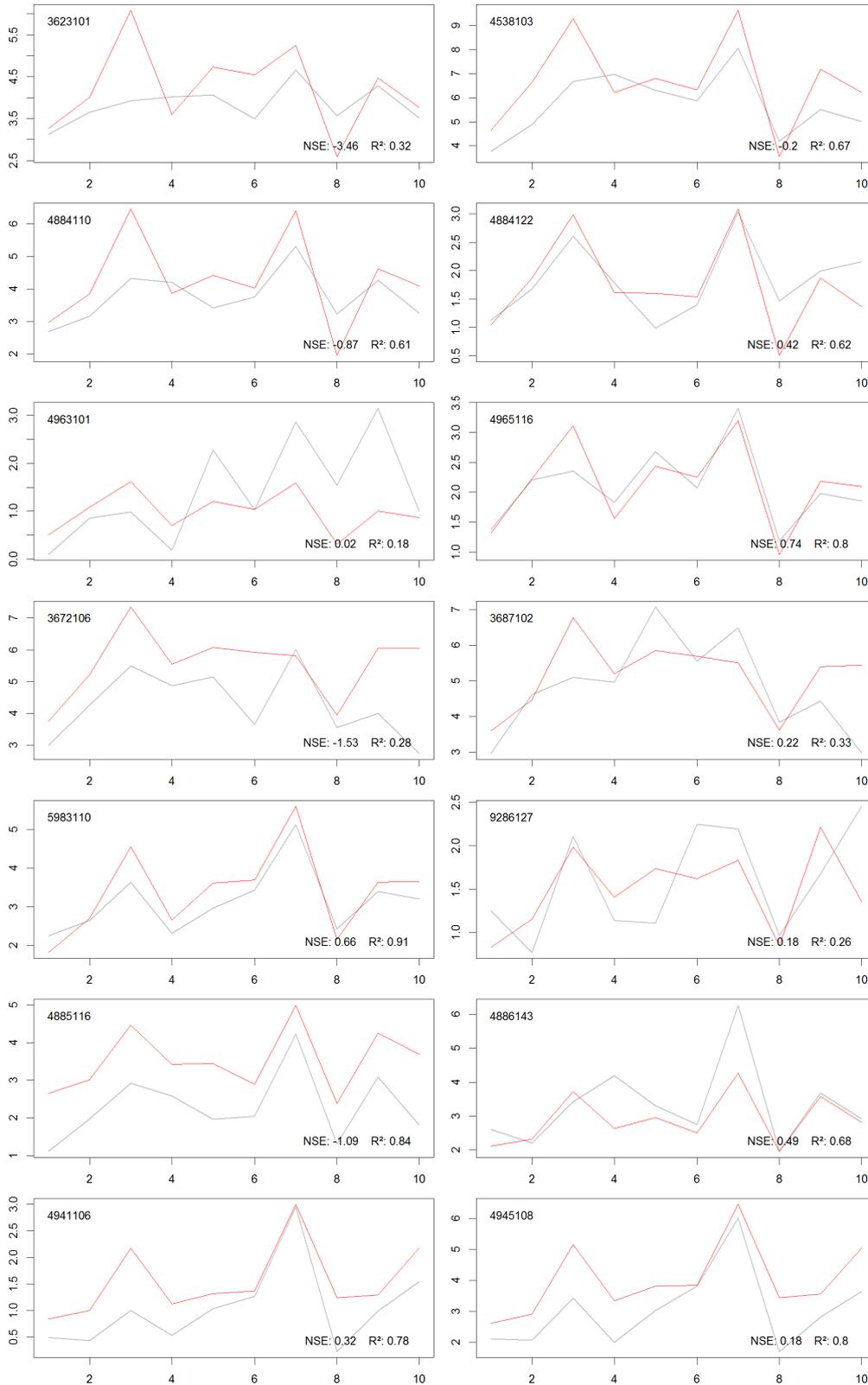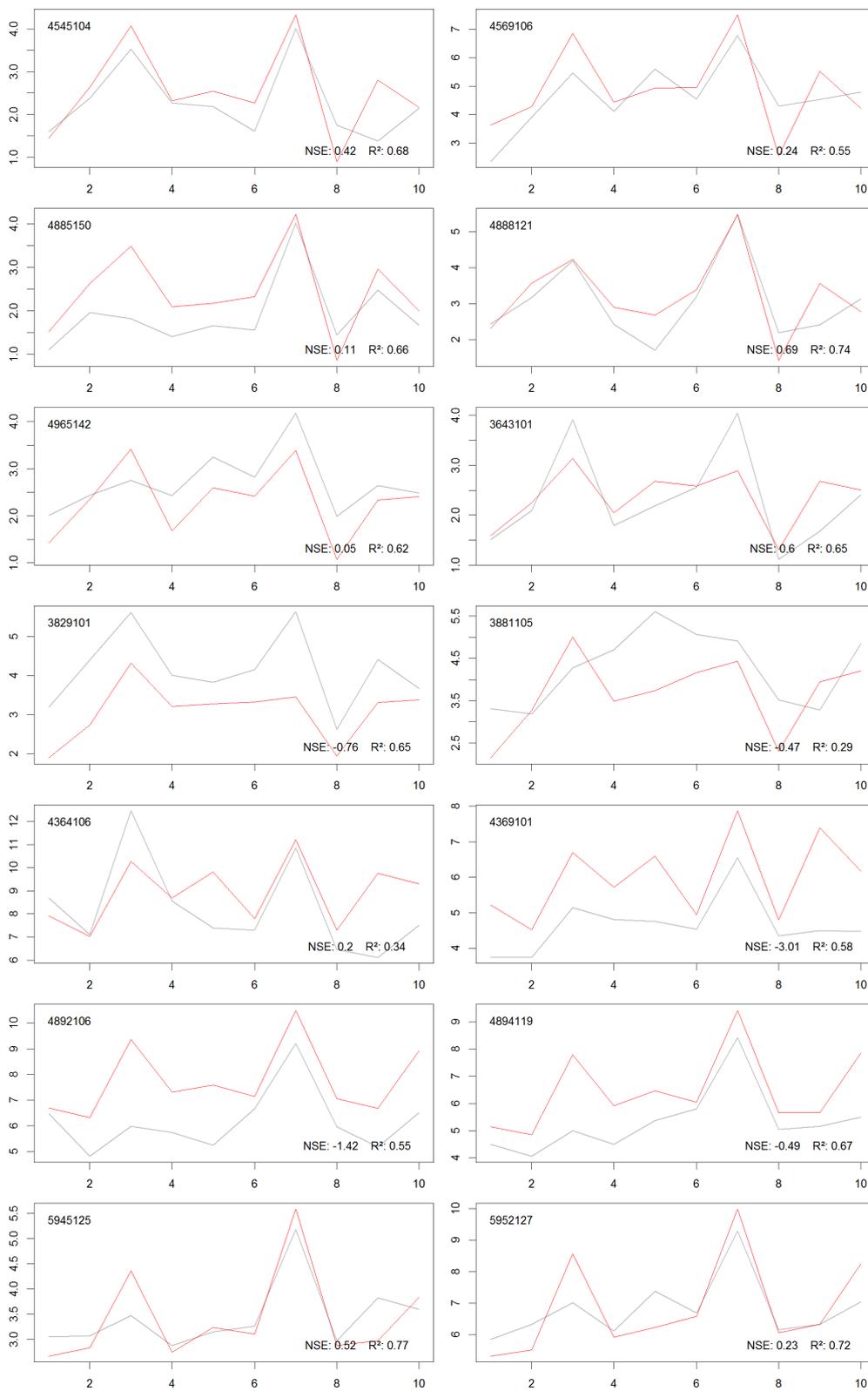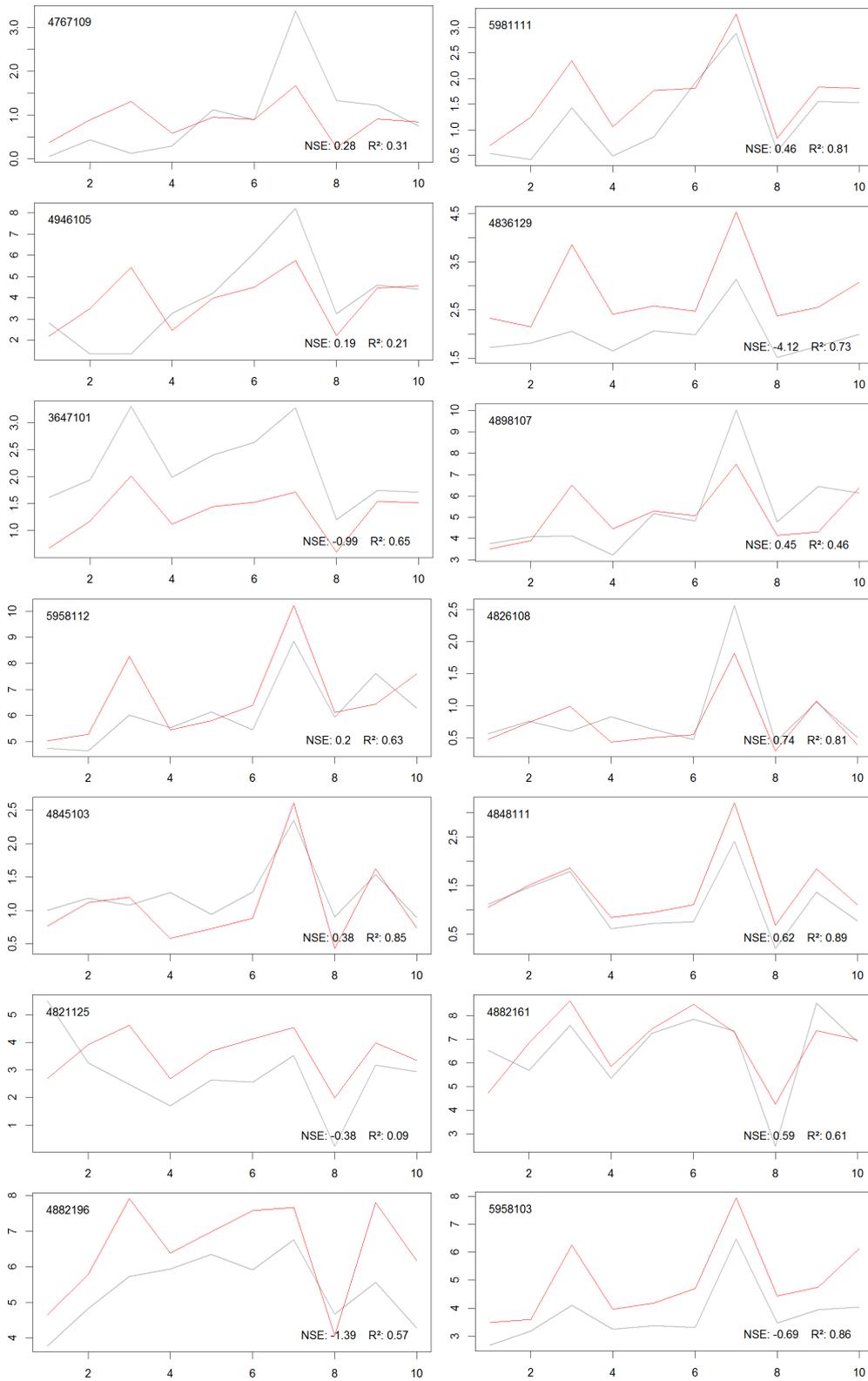
Appendix D3: Regression tree for identification of homogeneous regions for the set of 51 stations in the period 1966 - 1995.

Maximum $P_{mean\ 12-2} < 2.91$

Range $T_{max\ 3-6} < 4.24$ Range $DSD_{max\ 6-2} < 14.78$

0.4621   0.2844

Range $WSD_{mean\ 12-9} < 0.79$ Max $DSD/WSD_{mean\ 3-0} < 1.78$

Fiel Capacity < 21.44 Range $WSD_{mean\ 12-4} < 1.10$

0.1893   0.3013

0.1134   0.1559   0.2165   0.1698