Freie Universität Berlin, Department of Earth Sciences

Dealing with missing data in hydrology

- Data analysis of discharge and groundwater time-series in Northeast Germany

Dissertation submitted to the Department of Earth Sciences

Freie Universität Berlin, Germany,

for the academic degree of Doctor of Natural Sciences (Dr. rer. nat.)

YONGBO GAO

February 10th, 2017

Berlin

Dealing with missing data in hydrology
- Data analysis of discharge and groundwater time-series in Northeast Germany

Dissertation submitted to the Department of Earth Sciences of Freie Universität Berlin, Germany, for the academic degree of Doctor of Natural Sciences (Dr. rer. nat) in Hydrology

Erstgutachter : Prof. Dr. Christoph Merz
Freie Universität Berlin, Fachbereich Geowissenschaften, Institut für Geologische Wissenschaften

Zweitgutachter: Prof. Dr. Michael Schneider
Freie Universität Berlin, Fachbereich Geowissenschaften, Institut für Geologische Wissenschaften

Drittgutachter: Prof. Dr. Gunnar Lischeid
Universität Potsdam, Institut für Erd- und Umweltwissenschaften
Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF), Institut für Landschaftswasserhaushalt

Die Disputation erfolgte am: 10. Februar 2017

Erklärung

Hiermit erkläre ich, Yongbo Gao, dass diese Arbeit ausschließlich auf Grundlage der angegebenen Hilfsmittel und Hilfen selbstständig von mir verfasst wurde. Diese Arbeit wurde nicht in einem früheren Promotionsverfahren eingereicht.

Berlin, den 09. Januar 2017
Yongbo Gao

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Dr. Christoph Merz, Prof. Dr. Gunnar Lischeid, Prof. Dr. Michael Schneider for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Besides my supervisors, I would like to thank Marcus Fahle, Tobias Hohenbrink, Steven Böttcher, Philipp Rauneker, Steffen Gliege, Jin Sun for all their comments, supporting and encouragements. Special thanks to Björn Thomas for collecting and extracting the meteorological and catchment properties data of the catchments in the Federal State of Brandenburg, Germany. Furthermore, the whole working group at the Leibniz Centre for Agricultural and Landscape Research (ZALF) contributed to my work with their help.

My sincere thanks also goes to China Scholarship Council (CSC) who provided me an opportunity to achieve something I have never dreamed in my life before. Without their precious support it would not be possible to conduct this research.

Last but not least, I would like to thank my family: my parents for supporting me spiritually and financially throughout writing this thesis. In particular, I am grateful to have Stefan Wagner for enlightening me the first glance of research.

Abstract

Hydrological missing data is a common issue for hydrologists as it poses a serious problem for many statistical approaches in hydrology which require complete data sources since missing data is often harmful beyond reducing statistical power. For reasons of convenience, researchers often resort to simple solutions to deal with missing data such as simply discarding observations characterized by missing data or by replacing missing data with a statistical methodology. Despite its convenience, discarding is suboptimal as it reduces the quality of the conclusion to be drawn when analyzing the data. Actually, a variety of statistical techniques are available to treat missing data. My research is about finding the right techniques to deal with missing data problems in Hydrology and distinguishing in which certain circumstances which method works better.

First, various imputation methods available to the hydrological researchers have been reviewed, including arithmetic mean imputation, Principal Component Analysis (PCA), regression-based methods and multiple imputation methods.

Due to the time-series nature of hydrological data often requires more flexible non-linear model, we therefore put an emphasis on time-series regressions approaches that exploit the time series nature of hydrological data. Auto Regressive Conditional Heteroscedasticity (ARCH) models which originate from finance and econometrics and Autoregressive Integrated Moving Average (ARIMA) models are discussed regarding the applicability to hydrological contexts here. I focused the attention on discussing econometric time-series methods as they explicitly model the particular statistical properties of hydrological time-series (autocorrelation and heteroscedasticity) which are mostly neglected in algorithmic machine learning approaches.

Second, the performances of imputation techniques which are widespread and easy to use but ignore the time series nature of hydrological data and imputation techniques exploiting their time series nature are compared. By running a hydrological model - Hydrologiska Byråns Vattenbalansavdelning (HBV) model we generated 5 different discharge time series that exhibit different patterns of volatility to analyze.

The combination of Mean Squared Error (MSE) and Nash Sutcliff efficiency (NSE) as performance measures demonstrates that econometric time series models such as Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Conditional Heteroscedasticity (ARCH) model outperform alternative imputation approaches such as mean imputation or Ordinary Least Squares (OLS) based regression methods.

Furthermore, we examined how the inclusion of information beyond the time-series of the variable of interest itself can improve imputation results. Extensions of these models to incorporate additional exogenous regressors are readily available with ARIMAX and ARCHX models. Using discharge data from Brandenburg in the northeast of Germany, we compare the imputation performance of univariate ARIMA and ARCH models which have been shown well in hydrological settings before with the performance of extended model version.

These results shown that the models' performance can be further enhanced by the inclusion of exogenous regressors such as precipitation, potential evapotranspiration or discharge measures from neighboring research areas. In particular, the inclusion of discharge measures of neighboring areas has a bigger effect on imputation quality. Moreover, the choice between ARIMA/X and ARCH/X is less important than the choice of additional regressors.

Despite they overall encouraging findings there are, however, on the conceptual level, our results have been obtained using data from only one catchment area (Brandenburg) and the results might differ for data obtained from other catchments. More comprehensive validation of our results using data from different settings therefore seems to be warranted.

Zusammenfassung

Hydrologische Daten sind oft durch fehlende Messwerte und Datenlücken gekennzeichnet, was die Anwendung statistischer Methoden zur Datenanalyse erschwert. Gängige statistische Ansätze erforden vollständige Datensätze und fehlende Werte können nicht nur ihre Präzision reduzieren sondern auch zu verzerrten bzw. falschen Ergebnissen führen. Einfache Ansätze zum Umgang mit fehlenden Daten beinhalten das Löschen von Beobachtungen mit fehlenden Werten in einer oder mehreren Variablen oder auch das Ersetzen fehlender Werte durch den Einsatz statistischer Prognosemethoden (Imputation). Das Löschen von Beobachtungen ist oft suboptimal, da es die Qualität statistischer Schlussfolgerungen reduziert. Aus diesem Grund wurde eine Reihe von Verfahren entwickelt, die fehlende Werte mit prognostizierten Werten ersetzen und somit die Daten komplettieren. Die vorliegende Arbeite soll einen Überblick über alternative Imputationsverfahren geben und deren Anwendbarkeit in verschiedenen hydrologischen Problemstellungen evaluieren.

Im ersten Teil der Arbeit wird die Problematik fehlender Werte und ihr Einfluss auf die Anwendbarkeit statistischer Analyseverfahren dargestellt. Darauf aufbauend werden gebräuchliche Imputationsverfahren vorgestellt und vergleichend diskutiert. Dabei wird auf verschiedene Verfahren eingegangen, wie etwa das Ersetzen fehlender Werte durch den Mittelwert aller Beobachtungen, die Hauptkomponentenanalyse (PCA) oder regressionsbasierte Imputationsverfahren. Schwerpunkt dieser Untersuchungen ist die Darstellung zeitreihenbasierter Verfahren, da hydrologische Daten in der Regel als Zeitreihen vorliegen. Insbesondere werden die in der Volkswirtschaftslehre gebräuchlichen Autoregressive Integrated Moving Average (ARIMA) und Autoregressive Conditional Heteroscedasticity (ARCH) Modelle vorgestellt. Diese Modelle wurden ausgewählt, da sie explizit die Zeitreihencharakteristiken von Daten (Autokorrelation und Heteroskedastizität) modellieren, die in alternativen Verfahren oft vernachlässigt werden.

Im zweiten Teil der Arbeit werden die Ergebnisse von einfachen und weitverbreiteten Imputationsverfahren den Ergebnissen zeitreihenbasierter Imputationsverfahren

gegenübergestellt. Dies erfolgt in einem hydrologischen Kontext. Mittels des Hydrologiska Byråns Vattenbalansavdelning (HBV) Modells werden zuerst fünf verschiedene Abflusszeitreihen generiert, die durch unterschiedliche Volatilität gekennzeichnet sind. Anschließend werden zufällig generierte Werte in diesen Zeitreihen mittels alternativen Imputationsverfahren approximiert und die Imputationsqualität mittels des Mean Squared Error (MSE) und des Nash Sutcliffe Efficiency (NSE) Kriteriums verglichen. Die Ergebnisse belegen, dass ökonometrische Zeitreihenmodelle wie etwa das Autoregressive Integrated Moving Average (ARIMA) und das Autoregressive Conditional Heteroscedasticity (ARCH) Modell alternativen Methoden überlegen sind.

Im dritten Teil der Arbeit werden Generalisierungen der vorgestellten ARIMA und ARCH Modelle in einem ähnlichen Kontext evaluiert. Diese multivariaten Modelle (ARIMAX und ARCHX) beziehen neben der Zeitreihe der abhängigen Variablen weitere, unabhängige Variablen in das Modell mit ein. Basierend auf exemplarischen Abflusszeitreihen aus Brandenburg wird die Imputationsleistung von ARIMA und ARCH Modellen mit denen der erweiterten Modelle verglichen. Die Ergebnisse zeigen, dass die Präzision von Zeitreihenmodellen zur Modellierung gemessenen Abflusses durch die Einbeziehung zusätzlicher unabhängiger Variablen wie Niederschlag, potentielle Verdunstung oder Abfluss in Nachbarregionen erhöht werden kann. Insbesondere zeigt sich, dass die Berücksichtigung von Abflussdaten aus Nachbarregionen einen größeren Effekt auf die Imputationspräzision, hat als die anderen Variablen. Der Unterschied zwischen ARIMA/X und ARCH/X hingegen ist weniger bedeutend als die Wahl zusätzlicher Regressoren.

Trotz viel-versprechender Erkentnisse auf konzeptioneller Ebene ist anzumerken, dass die hier vorgestellten Ergebnisse auf den Daten eines Einzugsgebiets basieren und für Daten weiterer Einzugsgebiete variieren können. Eine umfassendere Validierung der Ergebnisse auf Basis von Daten unterschiedlicher Einzugsgebiete erscheint daher zukünftig sinnvoll.

Contents

List of Figure

List of Tables

1 General introduction

1.1 Scientific background and outline

Like almost all fields of science, hydrology has benefited to a large extent from the tremendous improvements of scientific instruments that are needed to collect data and an increase in available computational power and storage capabilities over the last decades (Bradbury et al., 1999). These technological developments have facilitated the development and the application of more advanced models of hydrological phenomena: both computationally intense simulation models as well as statistical approaches including determination of the flow duration curve, autocorrelation function, spectrum analysis, extreme value analysis based on the generalized extreme value distribution of annual blocks are ubiquitous nowadays to analyze and predict hydrological phenomena. Observed real-world data, however, is typically needed to calibrate (in the case of simulation models) and to estimate (in the case of statistical models) these models before meaningful predictions can be derived (Marques et al., 2006; Smith, 1989; Yanik & Avci, 2004).

Data that is required for model estimation and is usually collected in observation stations and stored in databases that are available for research purposes. It is a well-known fact, however, that numerous hydrological and research databases contain missing values for multiple and often idiosyncratic reasons (Elshorbagy, Simonovic, & Panu, 2002). They include failure of observation stations, incomparable measurements or manual data entry procedures that are prone to errors and also equipment errors (Johnston, 1999). Missing data is a challenge for empirical research in general and for hydrology in particular as it reduces the power and the precision of statistical research methods (Roth, Switzer, & Switzer, 1999). In addition to a reduction in the power of these methods, missing data can also lead to biased estimates of the relations between two or more variables (Pigott, 2001). Both problems – reduction in power and bias of estimates – can lead to inaccurate conclusions in analyses of datasets that contain missing data (Graham, 2009). Missing data is also a frequent problem in deterministic hydrological modeling which relies on observed historical data to model complex relations between variables relating to weather conditions and geographic surroundings (Gill, Asefa, Kaheil, & McKee, 2007).

Against this backdrop it needs to be highlighted imputation methods which attempt to 'fix' datasets characterized by missing data by replacing them with inserting numerical values have improved dramatically over the last decades (Peugh & Enders, 2004). The availability of more sophisticated imputation methods allows researchers replacing missing values with imputed values rather than excluding them from the analysis entirely (Saunders et al., 2006).

It is the aim of this thesis to provide an overview over alternative imputation methods that are available to the researcher and evaluate the performance of selected imputation methods in a given hydrological settings. In this attempt, particular emphasis is laid on the fact that hydrological data can be characterized as time-series data in which statistical patterns such as autocorrelation or seasonality emerge over time and can be exploited for imputation purposes. Note that the three chapters of this thesis originally written as review and research articles are waiting publication or submission in international, peer-reviewed journals.

The goal of Chapter 2 first defines patterns of the "missingness" or rather incompleteness of the data as it has important implications on the choice of particular imputation methods. It has to be diagnosed whether data in some observations is missing randomly or whether the observed incompleteness follows a particular pattern (R. Little & Rubin, 1987). Following Rubin (1976), Chapter 2 classifies distinguishes three so-called missing data mechanisms that describe relationship between measured variables and the probability of a missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). When missing data are MNAR, however, results from statistical analyses will be iased and there is little what imputation techniques can do to ease the problem (Donders, van der Heijden, Stijnen, & Moons, 2006). After these basic discussion, Chapter 2 then presents an overview of different imputation methods that allow to address MCAR and MAR situations. They include arithmetic mean imputation, Principal Component Analysis (PCA), regression-based methods and multiple imputation methods. A particular focus, however, is laid on time-series regressions approaches such as Autoregressive Integrated Moving Average (ARIMA) models and Auto Regressive Conditional Heteroscedasticity (ARCH) models which originate from finance and econometrics.

Chapter 3 compares the performance of the imputation techniques which have been introduced in Chapter 2. The aim is analyze to what extent methods that are widespread and

easy to use but ignore the time series nature of the data can be outperformed by imputation techniques that exploit the time series nature of hydrological data. In particular, Chapter 3 focuses on the performance of advanced statistical techniques such as Autoregressive Moving Average/Autoregressive Integrated Moving Average (ARMA/ARIMA) models and Autoregressive Conditional Heteroscedasticity (ARCH) time series models. The basic idea of Chapter 3 is to use discharge time series data that can be found typically in hydrological applications as reference data. We resort to using output discharge data obtained from a hydrological model - Hydrologiska Byråns Vattenbalansavdelning (HBV) model so that to obtain the reference data that does not suffer from missing values itself. Furthermore, we use HBV model to create reference discharge time series with specific characteristics. In order to evaluate different imputation methods, a certain fraction of the observations of the reference data is made missing. These missing values will then be replaced by approximations obtained from different imputation methods. Comparing the reference time series data with the imputed time series will allow me to draw conclusions regarding the performance of different imputation methods. The findings of Chapter 3 reveal that imputation methods that neglect the time series nature of the underlying reference data perform significantly worse than imputation methods that exploit this feature of the data. Moreover, advanced time series methods such as ARCH significantly outperform relatively simple time series method such as the preceding value imputation.

Chapter 4 exclusively focuses on time-series models that have been shown to outperform cross-sectional imputation methods and departs from the observation that both ARIMA and ARCH exclusively exploit time-dependencies in the time-series of a given variable while ignoring other available information that might improve the quality of imputation (Degiannakis & Xekalaki, 2004; Stergiou, Christou, & Petrakis, 1997). Including additional information (additional variables) in a statistical model of the dependent variable, however, might improve its quality therefore also affect imputation performance. In this chapter, we analyze to what extent extensions of the ARIMA and the ARCH model that incorporate additional exogenous regressors - ARIMAX and ARCHX models - outperform ARIMA and ARCH models that exclusively exploit the time-series properties of the dependent variable. In particular, we compare imputations for various shares of missing values in a time-series of daily discharge derived from alternative time-series model: In a first step, we impute missing values from ARIMA and ARCH models that exclusively rely on the observed time-series of

discharge. Second, we approximate missing values using extended ARIMAX and ARCHX models that include additional exogenous regressors such as precipitation, potential evapotranspiration or discharge measured from neighboring catchment areas. Finally, we compare the results from the different imputations in order to determine which approach yields the best results. We show that models including additional regressors seem to outperform more parsimonious models significantly.

The dissertation concludes with chapter 5 which synthesizes the main results and the conclusions regarding different imputation methods and their application in hydrology.

1.2 Regional framework

The spatial scope of all Chapters in this dissertation is the federal state of Brandenburg located in Northeast Germany (Figure 1.1). The whole area is 29,479 km^2 excluding Berlin in its center, has a population of 2.5 million. In this region, forest area contributes to 35% of the whole area. Agricultural land is another main land use type with 34% cropland and 9% pasture. With a mean annual precipitation of 557 mm and a mean annual temperature of 8.7 ℃ (period: 1960-1990; German Weather Service, 2012), it is one of the areas in with the lowest climatic water balance in Germany. Due to high climatic water demand, the evapotranspiration here is approximately 510 mm per year, only leaving 100 mm per year as runoff (Lischeid & Nathkin, 2011). The runoff exhibits substantial spatial variability, depending on local meteorological conditions. Groundwater flow and groundwater discharge into rivers and channels are the dominating hydrological components of the regional water cycle. About 80 out of 100 mm runoff per year occurs as baseflow, whereas surface runoff plays only a minor role, accounting for less than of total runoff (Merz & Pekdeger, 2011).

Water in Brandenburg is drained by rivers Elbe and Oder to the Northern Sea and Baltic Sea, respectively. According to climate projections, it is located in the transition zone between increasing streamflow in northern Europe and decreasing streamflow in southern Europe.

Figure 1.1: Location of the Federal State of Brandenburg in Germany and the German capital Berlin in the center of Brandenburg.

The whole region is part of a postglacial landscape which formed since the last Pleistocene glaciations. The ground mainly consists of glacio-fluvial deposits with several Quaternary aquifer (Natkhin, Steidl, Dietrich, Dannowski, & Lischeid, 2012). Soils are relatively young and mainly consist of sands and loamy sands, but less permeable soils are also occur (U Schindler & Müller, 2010). Close to rivers and at discharge areas, gley and peat soils evolved. Low gradients in land surface as well as in surface and subsurface flows, a large number of closed depressions and periglacial channels exposing locally raised relief energy, complex hydraulic interaction of different aquifers and a rather unstable but ecologically crucial interplay between groundwater and streams are major hydrological characteristics of this landscape. Moreover, the region exhibits a wide array of anthropogenic impacts on the fresh systems. These include weirs, dams and locks, flood protection which result in extensive use and alteration of regional freshwater quantity and quality.

For the last centuries, the hydrological system in this area has been altered by humans' behavior. Some of the artificial ditches and streams have existed for more than ten decades. More than 100 discharge gauges are maintained by the ministry of the Environment, Health and Consumer Protection of the Federal State of Brandenburg. Hydrological data such as

precipitation, discharge or temperature is typically is collected over time at given intervals. For more detailed description and overview on human impacts and hydrological changes within this landscape we refer to Merz and Pekdeger (2011) and Germer, Kaiser, Bens, and Hüttl (2011)

2 Dealing with missing data in hydrological data

2.1 Introduction

Gap-free time series are a necessary prerequisite for many statistical approaches in hydrology, including determination of the flow duration curve, autocorrelation function, spectrum analysis, extreme value analysis based on the generalized extreme value distribution of annual blocks, etc. The required data are usually collected in observation stations and stored in databases that can subsequently be accessed for research purposes. It is a well-known fact, however, that numerous hydrological and research databases contain missing values (Elshorbagy et al., 2002). The reasons behind missing data are multiple and often idiosyncratic. They include failure of observation stations, incomparable measurements or manual data entry procedures that are prone to errors and also equipment errors (Johnston, 1999)

Researchers have to find a solution to missing data problems as all of the approaches listed above can be applied properly only using complete data. When the available time series are long enough researchers can use a subset of the data that contains complete observations for a certain period. More often, unfortunately, the available data has been collected over shorter observational periods. In these cases, the application of statistical methods that strictly require complete time-series can be severely aggravated. In order minimize the missing data problem, researchers often resort to imputation methods where missing values are replaced with a numerical value that is obtained from a more or less sophisticated statistical method.

Moreover, it is quite common to encounter databases in which up to 50% of all observations contain missing data. It is very difficult to analyze them by using data analysis methods that are based on the assumption that the sample to be analyzed is a random sample from the population (or the entire database) and hence contains complete information (Farhangfar, Kurgan, & Dy, 2008). Over the last decades, imputation methods which attempt to 'fix' datasets characterized by missing data by replacing them with inserting numerical values

.

have improved dramatically (Peugh & Enders, 2004). The rise of more sophisticated imputation methods led many researchers to prefer replacing missing values with imputed values over excluding them from the analysis entirely (Saunders et al., 2006). Since the last 20 years, statisticians introduced imputation methods such as regression-based imputation, data imputation based on principal component analysis (PCA) or maximum likelihood techniques using the 'expectation – maximum' (EM) algorithm as well as 'multiple imputation' (MI). Often, these methods offer more promising solutions depending on the exact application (Soley-Bori, 2013).

In general, the choice of a specific imputation method is determined by the nature of the process generating the original data. For instance, often data is cross-sectional in its nature and a familiar statistical tool such as PCA or linear regression approaches can be used for imputation purposes. In hydrological settings, however, the choice of an appropriate imputation method needs to take into account the most important features of hydrological data: Hydrological data are often time-series data that are often characterized by stable trends over time and a high autocorrelation of the observations. Moreover, hydrological time-series often display random deviations from these trends and these deviations are not constant over time (Guzman et al., 2013). Given these features of the data generating process underlying the hydrological data, imputation of missing values should be based on statistical time-series methods that take into account the time series nature of hydrological data. For instance, singular spectrum analysis (SSA) or autoregressive moving average/autoregressive integrated moving average (ARMA/ARIMA) models have been applied in hydrological settings (Zhang, Wang, He, Peng, & Ren, 2011). One feature of time-series data that has received little attention in hydrological literature so far is non-constant deviations around a trend which is called heteroscedasticity. For this reason, Autoregressive Conditional Heteroscedasticity (ARCH) time-series models which originate from finance and econometrics will be discussed below. So ARCH models may not only be used to explain and characterize observed hydrological time-series but also for the imputation of missing observation in existing datasets which are characterized by non-constant high variability.

The goal of this chapter is to present an overview of different imputation methods that are available to researchers in hydrology. It starts with a brief introduction of the missing data problem. After the discussion of different patterns of missing data, a summary of the most

frequently used imputation methods will be presented. After this overview, imputation methods based on time-series regression methods will be introduced and their benefits for hydrological applications will be highlighted. A special focus will be laid on ARCH models and the discussion to what extent they might be applied to hydrological settings of missing data. The article concludes with a summary of the major findings and implications for hydrological research.

The phenomenon of missing data has been discussed extensively within and beyond statistics (Schafer & Graham, 2002). This is a common problem in empirical studies in social, medical or geographical sciences and occurs for a number of different reasons, including erroneous manual data entry, and equipment errors during the collection of data or also a loss of data due to defective storage technologies (Tannenbaum, 2009).

Missing data is a challenge for empirical research as it generally reduces the power and the precision of statistical research methods (Roth et al., 1999). In addition to a reduction in the power of these methods, missing data can also lead to biased estimates of the relations between two or more variables (Pigott, 2001). Both problems – reduction in power and bias of estimates – can lead to inaccurate conclusions in analyses of datasets that contain missing data (Graham, 2009). Missing data is also a frequent problem in deterministic hydrological modeling which relies on observed historical data to model complex relations between variables relating to weather conditions and geographic surroundings (Gill et al., 2007).

To date, a variety of different statistical techniques are available to address the problems arising from missing data (Puma, Olsen, Bell, & Price, 2009). An understanding of these methods is increasingly important as having complete and accurate databases is often the prerequisite for the application of increasingly sophisticated statistical methods. Often, it is tempting to follow the simplest way of dealing with missing data which consists of simply discarding (i.e., deleting) observations where information in one or more variables are missing. This approach is also one of the default options for statistical analysis in most software packages and is called “listwise” or “pairwise” deletion (Harrington, 2008). Despite its convenience, this method is practical only when the data contains only a relatively small portion of observations with missing data. If only a negligible share of the observations contains missing data, the analysis of the remaining observations will not lead to serious inference problems (Tsikriktsis, 2005). Nevertheless, it has to be pointed out that deletion of

even only a small share of all observations in a dataset will reduce the statistical power and the accuracy of the analyses undertaken (D. B. Rubin & Little, 2002).

2.2 Patterns of missing data

Before any missing data imputation can be implemented, the most important question that a researcher has to address relates to the underlying patterns of the "missingness" or rather incompleteness of the data. In particular, it has to be diagnosed whether data in some observations is missing randomly or whether the observed incompleteness follows a particular pattern (R. Little & Rubin, 1987). In this context, the classification system of missing data which has been outlined by Rubin (1976) and colleagues remains in widespread use today. Following Rubin (1976), missing data can be seen as a probabilistic process and allows to distinguish three so-called missing data mechanisms that describe relationship between measured variables and the probability of a missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Assume, our data contains one variable of primary interest *Y* and a number of additional variables referred to as a vector *X*. Following this notation, and with *m* being an indicator variable for missing observations in *Y*, i.e., m = 1 if a data point is missing and m = 0 if a data point has been observed, the probability that a value in *Y* is missing can be expressed as a function of *Y* and *X* with

$$Pr\,(\boldsymbol{m = 1} \mid X, Y), \tag{2.1}$$

At first, suppose that the probability of a missing observation in *Y* is completely independent of observed or unobserved measurements of this variable or other variables *X* and also independent of the other observations in the dataset. If this is the case, the absence of a value in a given observation is called missing complete at random (MCAR) (Allison, 2012). This mechanism is what researchers consider as purely random missingness. The case of MCAR missing data causes the least problems for statistical analyses. In a dataset including missing values that are MCAR, the subset of all observations containing the missing data can be deleted. The remaining subset then contains all observations with complete information. This approach often is called listwise/pairwise deletion (McKnight, McKnight, Sidani, & Figueredo, 2007). As the resulting dataset containing only the observations with complete

data is a random sample from the initial data, it can easily be shown that results based on its statistical analysis will be unbiased (D. B. Rubin, 1976). Mathematically, MCAR implies that

$$Pr(\mathbf{m = 1}|X, Y) = Pr(\mathbf{m = 1}), \quad (2.2)$$

Another pattern of missing values is called missing at random (MAR). MAR is a less restrictive assumption regarding the pattern of missing values compared to MCAR. When data are missing at random, the probability of missing data in a variable for a given observation is only related to any other observed variable rather than on *Y* itself (McKnight et al., 2007). This implies

$$Pr(\mathbf{m = 1}|X, Y) = Pr(\mathbf{m = 1}|X), \quad (2.3)$$

Data that contain information MAR requires more attention than data is MCAR: all simple imputation methods for missing data, i.e., listwise and pairwise deletion, arithmetic mean imputation, will give biased results in analyses of the relations between variables in the dataset (Pigott, 2001). Nevertheless, unbiased results can be obtained in the case of data MAR. This requires the application of more sophisticated imputation methods, however, including single and multiple imputations (Donders et al., 2006).

In cases where neither the MCAR nor the MAR assumption holds, data is said to be Missing Not At Random (MNAR) (McKnight et al., 2007). If cases are MNAR, there is a relationship between the variables that include missing data and those for which the values are present and hence the equation is valid

$$Pr(\mathbf{m = 1}|X, Y), \quad (2.4)$$

When missing data are MNAR, results from statistical analyses will be biased and there is little what imputation techniques can do to ease the problem (Donders et al., 2006).

So, it is important to investigate whether the missing pattern is random or not before any statistical test is conducted. For a full discussion, see D. B. Rubin and Little (2002).

2.3 An overview of traditional missing data handling techniques

Dozens of techniques to deal with the missing data problem have been used for decades (Baraldi & Enders, 2010). The more common traditional approaches to deal with missing data include removing the values with incomplete data/deletion or so-called single

imputation methods where missing values are replaced (Peugh & Enders, 2004). Whereas deletion methods reduce the sample size, the purpose of single imputation methods is the retention of the sample size and statistical power in subsequent analyses (Cool, 2000). However, single imputation methods have drawbacks which are addressed by more complicated multiple-imputation methods which are often based on Monte-Carlo-type simulations and require more computational sophistication than simple imputation methods.[1] This section reviews the most important imputation methods. Despite their widespread use, these methods still have shortcomings in their procedures which will be also illustrated in this section.

2.3.1 Listwise and pairwise deletion

The elimination of all observations which have missing data in one or more variables is called listwise deletion (Mcdonald, Thurston, & Nelson, 2000). The primary benefit of listwise deletion is convenience (King, Honaker, Joseph, & Scheve, 1998). This approach has several drawbacks as addressing incomplete data by deleting observations inevitably will reduce the sample size (L. H. Rubin, Witkiewitz, St Andre, & Reilly, 2007). It is a well-established fact in statistics that smaller sample sizes reduce the statistical power and precision of standard statistical procedures (D. B. Rubin & Little, 2002). A reduction in the precision of tests and estimates will render inference (such as hypothesis testing) conservative. A more severe effect could be that it can introduce a systematic bias. If the data is MCAR, a sample excluding observations with missing values will be a random draw from the complete sample and estimates remain unbiased. If, however, the relatively strong assumption of MCAR is violated, the deletion of observations with missing data will bias the value of the estimates of interest. A simulation by Raaijmakers (1999) demonstrated that the statistical power is reduced between 35% (with 10% missing data) and 98% (with 30% missing data) by using listwise deletion.

The elimination of observations on case-by-case basis depending on which variables are used in a specific analysis is called pairwise deletion. It is different than listwise deletion as an observation is deleted only if a variable used in the analysis contains a missing value

[1] As in Section 2.3.3 below, multiple imputation generates multiple datasets containing imputed values which are enhanced by a random error term. The desired statistical analyses are then carried out multiple times on these different datasets and their results aggregated. This approach allows getting more appropriated standard errors on the estimates of the desired parameters.

(Wothke, 2000). For example, if a respondent is missing information on variable A, the respondent's data could be used to calculate other correlations, such as the one between B and C. Pairwise deletion is often an improvement over listwise deletion because it preserves much more information by minimizing the number of cases discarded compared to the listwise deletion (Roth, 1994). Amongst the most important problems of pairwise deletion is a limited comparability of different analyses as the number of observations varies between different pairwise comparisons (Croninger & Douglas, 2005). Moreover, estimates of covariances and correlations might be biased when using pairwise deletion since different parts of the sample are used for each analysis (Kim & Curry, 1977).

Despite their shortcomings, deletion techniques are the default options for missing data techniques in most statistical software packages, and these techniques are probably the most basic methods of handling missing data (Marsh, 1998).

2.3.2 Single imputation methods

Single imputation approaches generate a single replacement for each missing value with suitable values prior to the actual analysis of the data (Enders, 2010). A variety of different missing data imputation methods have been developed over the years, and are readily available in most standard statistical packages. As has been discussed above, all imputation methods produce biased results if the relatively strong MCAR assumption is violated. In particular, imputation is advantageous compared to listwise or pairwise deletion as it generates a complete dataset. Hence, it also makes use of the data that deletion techniques would discard. Nevertheless, as will be discussed below, these methods have potentially drawbacks and even in an ideal MCAR situation most of these approaches generate biased parameter estimation.

Many different single imputation methods have been introduced and applied: arithmetic mean imputation, principle component analysis (PCA) and regression-based imputation are the most commonly known ones and will be briefly introduced below.

Arithmetic mean and median imputation

Arithmetic mean imputation replaces missing values in a variable with the arithmetic mean of the observed values of the same variable (Roth, 1994). Median imputation replaces

missing values with the median value of the observed values of the same variable (McKnight et al., 2007). Both approaches are very convenient since they generate a complete dataset easily (Hawkins & Merriam, 1991). Median imputation is preferable when the distribution of the underlying variable is not symmetric but rather skewed (McKnight et al., 2007).

However, even in situations where the strong MCAR assumption holds, these approaches distort the resulting parameter estimates (Enders, 2010). For instance, they attenuate the standard deviation and the variance of estimates obtained from analyses of mean imputed variables (Baraldi & Enders, 2010). The reason for the reduction of the standard deviation of estimates stems from the fact that the imputed values are identical and at the center of the distribution which reduces the variability of the data (R. J. Little, 1988). This fact also attenuates the magnitude of estimated covariances and correlations between mean-imputed variables and other variables in a dataset (Malhotra, 1987).

Regression-based imputation

Regression-based imputation replaces missing data with predicted values from a regression estimation (Greenland & Finkle, 1995). The basic idea behind this method is using information from all observations with complete values in the variables of interest to fill in the incomplete values which is intuitively appealingly (Frane, 1976). Different variables tend to be correlated in many applications (Allison, 2001). Exploiting information from all observations with complete information is a strategy which regression-based imputation methods share with multiple imputation and maximum likelihood imputation methods, although the former approach does so in a less sophisticated way (Raghunathan, 2004). Note that maximum likelihood imputations not refer to the estimation method used by the regression based imputation methods but rather to the technique of selecting among different values that might be chosen to assess a missing value.

The first step of the imputation process is to estimate regression equations that relates the variable that contains missing data (the dependent variable of the regression) to a set of variables which have complete information across all observations in the dataset (independent variables of the regression). This regression is estimated only for the subset of the data that contains all observations that have complete information both for the

dependent variable and the independent variables. The results of the regression are estimates due to the relation of independent to dependent variables.

The second step exploits this information. Using the regression results from the first step, missing values for the observations that could not have been included in the regression are replaced by predictions obtained from combining the observed values of the independent variables and the estimates from the first step of how they are related to the dependent variables. These predicted values fill in the missing values and produce a complete dataset (Frane, 1976). In the case of *k* variables with *n-r* missing values in the *k-th* variable (*n* being the total number of observations and *r* being the number of complete observations), a linear regression can be estimated based on all *r* complete observations. The regression yields estimated regression's coefficient. Based on these estimates, missing values in the *k-th* variable can be predicted, i.e., imputed with

$$\widehat{y_{t,k}} = \widehat{\beta_0} + \sum_{j=1}^{k-1} \widehat{\beta_J}\, y_{i,j} \;\forall\; i\epsilon[r,n] \tag{2.5}$$

While regression-based imputations most frequently rely on simple linear regressions, it is worth noting that more flexible regression approaches can equally be used and might even be more advantageous depending on the application. In section 2.4, we will discuss to what extent time-series regression approaches can be used in regression-based imputations of hydrological data.

From a methodological viewpoint, regression imputation is superior to mean imputation, but it can lead to predictable biases (van der Heijden, Donders, Stijnen, & Moons, 2006). In particular, regression based imputation methods lead to the opposite problem as mean imputation as missing data is replaced with values that are highly correlated to other variables in the data. Consequently, the application of regression based imputation methods will lead to overestimated correlations and R^2 statistics in subsequent data analysis.

Principle Component Analysis (PCA)–based imputation

Principle component analysis (PCA) originally has been conceived as a multivariate exploratory data analysis technique that can be used to extract patterns from datasets by transforming the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal

component), the second greatest variance on the second coordinate, and so on. Moreover, PCA can be used to compressed high-dimensional vectors into lower dimensional ones (Pandey, Singh, & Tripathi, 2011). The principal idea behind PCA here is to find a smaller dimensional linear representation of data vectors so that the original data can be approximated from the lower dimensional representation with minimal mean square error. Graphically, PCA can then be interpreted of a projection of the original data points on a lower dimensional space which minimizes the reconstruction error (I. T. Jolliffe, 1993).

Formally, PCA can be expressed as follows. Assume that observed data points $x_1, x_2, \dots, x_n \in R^p$ are *p*-dimensional vectors. PCA defines a projection of these data on a *q*-dimensional space (with $q \leq p$) as

$$f(\lambda) = \mu + v_q \lambda \,. \tag{2.6}$$

In this *q*-dimensional model, μ is a vector of the mean values of length *p*, v_q is a $p*q$ matrix with *q* orthogonal unit vectors and λ is the *q*-dimensional projection of each original data vector *x*. A projection of the original data can be found by maximizing the variance of the projection of the original data along the new (reduced) dimensions of the projection space

$$\min_{\mu, \lambda_{1 \dots N}, v_q} \sum_{n=1}^{N} \left\| x_n - \mu - v_q \lambda_n \right\| . \tag{2.7}$$

Here, μ can be interpreted as the intercept of the projection space in the original space, $\lambda_1, \dots, \lambda_n$ are the projection coordinates of the original observations $x_1, \dots, x_n$. Note that PCA can be also be derived from a maximization of the variance of the projected data points along the new dimensions. The results are computationally equivalent.

While originally not devised as an imputation method, PCA can be used to replace missing values in a dataset and hence also as an imputation tool. For this purpose, an iterative PCA algorithm has been proposed by Kiers (1997). The algorithm can be summarized as follows:

1. Missing values are initially replaced by the sample mean.
2. PCA is conducted on the now complete dataset by minimizing the reconstruction error as described above to derive μ, λ_n and v_q.
3. Initially missing values are replaced by imputed values based on the results from step (2) with $x_n = \mu + v_q * \lambda_n$.

4. Steps (2) and (3) are repeated until the imputed values of initially missing values converge.

It can be shown that the iterative PCA corresponds to an expectation-maximization (EM) algorithm and is thus often named EM-PCA algorithm (de Leeuw, 1986; Dempster, Laird, & Rubin, 1977). This approach is computationally more efficient as it does not require the computation of the full covariance matrix. It needs to be stressed that one the biggest disadvantages of PCA is that the choice of the number of dimensions q in PCA needs to be done by the analyst and is not a result of the analysis. This has been identified as a core issue and a very difficult task that has been extensively discussed in the literature. For a treatment of this issue see for instance I. Jolliffe (2002).

2.3.3 Multiple imputation

In order to ease the negative impact of regression imputation mentioned above, more sophisticated approaches have been developed. The principle idea here is to replace each missing item with two or more plausible values, representing a distribution of possibilities. Therefore, these approaches are called multiple imputation(MI) (Graham & Hofer, 2000). Recent advances in computational power made multiple imputation available as relevant procedures are included in standard statistical software packages more frequently. The biggest advantage of multiple imputation is that inference regarding statistics such as correlations error obtained from multiple imputation are not overestimated because they incorporate uncertainty due to missing data (Lee & Carlin, 2010). However, there are some disadvantages in MI. The biggest disadvantage of MI is that it requires more computational effort since both imputation and the subsequent analyses have to be carried out multiple times (D. B. Rubin, 2004). It should be noted, however, that given the advances in computing hardware and software this is not a burden in practice and most statistical software packages nowadays contain routines for MI.

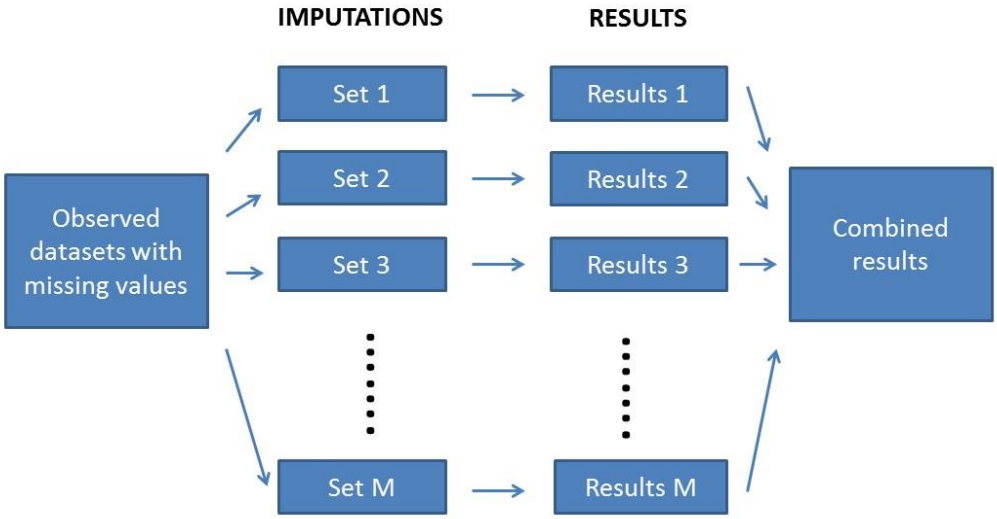

Figure 2. 1: Scheme of Multiple Imputation

While there are different approaches to MI imputation, the underlying sequence of computations steps is similar (Allison, 1999): First, the missing data are imputed by an appropriate model *M* times to produce *M* complete datasets. Most often, regression-based imputation techniques are used in this step. In each of the *M* steps, the predicted values from the regression analysis are varied by a random term of zero mean and a specified standard deviation. After this step, the desired statistical analysis can be carried out on each of the *M* datasets by using standard complete data analysis methods. This yields a set of *M* results of which thereafter average values and standard errors can be computed (Allison, 1999). This approach avoids an underestimation of standard errors and hence often is preferable to single imputation methods.

Despite its desirable properties, multiple imputation requires statistical and computational sophistication. For this reason, the remainder of this chapter focuses on single imputation methods which still seem to be more frequently used in hydrological settings.

2.4 Introduction to time series analysis and its application to imputation of missing values

2.4.1 Overview

A time series in our context is a discrete time-series defined as a series of observations of *Y* where are observations over several consecutive time-periods *t=1,…,T*. y_t might be the amount of discharge from a given measurement station that is measured on a daily basis. Hydrologist might be interested in analyzing run-off over time and how it depends on different types of boundary conditions. The assumption of independence of observations in the dataset then seems far-fetched. In hydrology, it is reasonable to assume that there are periods characterized by high run-off in which today's run-off will be related to the amount of run-off the day before and hence past values are correlated with today's value of run-off.

Data imputation approaches can make efficient use of dependencies between different observations in a time-series that is defined as data resulting from the observation of subjects which are repeatedly measured over a series of time-points (Hedeker & Gibbons, 1997). In contrast to conventional approaches, time-series techniques allow for the assumption that y_t is not independent of preceding observations of *y*. This is called autocorrelation or serial correlation where y_t is a function of a previous value of *y*. The adapted approaches exploit autocorrelation to model if a given phenomenon is not only based on conditions in *t* but also on its own history (for instance Y_{t-1}). In the following, the adaption of PCA to time-series data which is often called Singular Spectrum Analysis (SSA) is discussed before moving to a more comprehensive discussion of time-series regression techniques.

2.4.2 Singular Spectrum Analysis (SSA)

The aim of Singular Spectrum Analysis (SSA) is to decompose a time-series into regular oscillatory components and random noise applying the principles of PCA to time-series data (Hassani, 2007). For this reason, SSA can be considered a time-series version of PCA. SSA, on the other hand, can be applied to univariate time-series y_t with *t*=1,…,*T* in order to separate a signal in a time-series (trends or oscillatory movements) from a noise component that is random. To that end, so-called trajectory matrix is formed from the original data. Consider

the time-series Y=($y_1, y_2, \ldots, y_n$) of length *n* and choose a window length *L* (with $1<L<n$), $K = n - L + 1$ lagged vectors x_j of the original time-series can be generated with $x_j = (y_i, y_{i+1}, \ldots, y_{j+L-1})$ for $j = 1, 2, \ldots, K$. These vectors form the trajectory matrix *X* with

$$X = [X_1, \ldots, X_K]' = \begin{bmatrix} y_1 & y_2 & \cdots & y_L \\ y_2 & y_3 & & y_{L+1} \\ \vdots & & \ddots & \vdots \\ y_K & y_{K+1} & \cdots & y_n \end{bmatrix}. \quad (2.8)$$

In a second step, and similar to PCA, the trajectory matrix is then subjected to a single value decomposition yielding a set of so-called eigentriples which contain the principles components of *Y* (Wall, Rechtsteiner, & Rocha, 2003).

By projecting the principal components back onto the eigenvectors, a time series (referred to as the "reconstructed components") can be recovered in the original time units, each one corresponding to one of the PCs.

This third step of SSA splits the elementary matrices X_i into several groups and sums the matrices within each group. Finally, diagonal averaging transfers each of these matrices into a time series, which is an additive component of the initial series y_t.

It should be noted here, that the choice of window length *L* is a choice that has to be set by the researcher. The choice of *L* is important as it defines the maximum length of the oscillations that can be detected employing SSA. While the literature provides some guidance by providing rules of thumb for the choice of *L*, ultimately any ex ante choice of *L* remains arbitrary and there are no tests available that would allow conducting statistical inference regarding the choice of *L*. In the context of imputation, Kondrashov and Ghil (2006) propose an iterative approach to determine a suitable choice of *L*. In particular, they iteratively produce estimates of missing data points, which are then used to compute a self-consistent lag-covariance matrix and its empirical orthogonal functions. This approach allows to optimize the window length *L* by cross-validation.

2.4.3 Time-series Regression

Autoregressive and moving average models (ARMA, ARIMA)

Similar to linear regression frameworks, for instance, time-series regressions can easily be used for regression-based imputations methods. Imputed values are then derived from a prediction based on time-series regression instead of regression to an external variable. In particular, one can treat time-series prediction as a problem of missing data where the missing data located in the future are predicted based on regression to preceding data (Sorjamaa, Hao, Reyhani, Ji, & Lendasse, 2007).

Different time-series regression methods can be distinguished depending on the assumptions they put on the autocorrelation between different observations of *Y*. The most crucial assumptions related to the number of previous observations of *Y* that are considered in computing the contemporary value of *Y* (the order of the autocorrelation) and whether the correlation between the actual value of *Y* and preceding values is constant or changes over time. It is beyond the scope of this chapter to provide a detailed overview of these methods. Stock and Watson give a thorough treatment of time-series methods (Stock, Watson, & Addison-Wesley, 2007).

Formally, there are different ways of specifying a stochastic process that generates time-series where y_t and y_{t-j} are correlated over time, i.e., autocorrelation between different measures of *y* exists. One possible specification is an autoregressive process AR(*p*) of *p*th order with

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t, \quad (2.9)$$

In (9) epsilon is a random error term that follows a standard normal distribution and is independent over time with $E(\varepsilon t, \varepsilon t-i)=0$ for all $i \neq t$. *p* here denotes the number of lagged values of y_t that are condidered. ε_t is an independent and identically distributed error term with zero mean and constant variance. Commonly used auto-regressive (AR) models make the assumption that autocorrelation is constant over time and depends only on the intervals *j* between the y_t and y_{t-j}.

An alternative specification of a stochastic process that generates autocorrelation in a time-series are moving average (MA) processes in which the contemporary value of y_t is a function of its mean μ and a sequence of random innovations with

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots\ldots + \theta_p\varepsilon_{t-q}, \quad (2.10)$$

While in MA processes y_t is not directly a function of previous values y_{t-q}, autocorrelation between y_t and y_{t-q} is a consequence of the same random innovations ε_{t-q} entering the computation of different y_t.

In time-series modeling, there is often an explicit recognition that time-series models are merely intended to act as an approximation characterizing the dynamic behavior of the underlying series with the intention to approximate autocorrelation structures over a time (Adhikari & Agrawal, 2013). Only in rare circumstances it is intended to provide a "true" model of a time-series. Instead, the focus is often to determine whether a time-series model provides an approximation to observed behavior. While a "true" model may take a large number of lagged terms to provide a proper fit with the specification in (9), it is often possible to fit an observed autoregressive (AR) time-series more parsimoniously by combining it with a moving-average (MA) component consisting of a sum of weighted lags of the error term ε_t (Box and Jenkins 1976). The resulting ARMA model is written as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q}, \quad (2.11)$$

Equation (2.11) is often referred to as an ARMA(p,q) model as it contains a pth-order autoregressive component in the observable time series, y_t, and a qth-order moving average component of the unobservable random shocks ε_t. It is generally assumed that ε_t follows a so-called white-noise process with zero mean $E(\varepsilon_t)$ and constant variance $E(\varepsilon_t^2)=\sigma^2$. Moreover, it needs to be noted that for equation (11) to be a tractable model that can be fitted to data is the requirement of weak stationarity of the underlying time-series y_t. Weak stationarity is given if at least a time-serie's mean, variance and autocovariances are independent of t – whereas higher moments of the distribution of y_t over time might well depend on t. If $E(y_t)=\mu$, $E(y_t - \mu) = \sigma^2$ and $E[(y_t - \mu)(y_{t-j} - \mu)]=\gamma_j$, then a time-series of y_t is said to be weakly stationary. Strict stationarity, on the other hand, would imply that a time-series'

distribution does not depend on t at all and hence $E(y_t)=\mu$ *and* $E(y_t - \mu) = \sigma^2$ and all higher moments are independent of t.

In case a time series y_t is not stationary, stationarity can often be achieved by differencing the time-series one or more times (Box & Jenkins, 1976). If differencing is required the ARMA (p,q) model (Autoregressive Moving Average) becomes an ARIMA (p,d,q) model (Autoregressive Integrated Moving Average) where d denotes the order of differencing, i.e., the number of time y_t is differenced to achieve stationarity.

When fitting ARIMA models the choice of p, d and q can be guided by an inspection of autocorrelation and partial autocorrelation (which measure the correlation between y_t and y_{t-j} after accounting for the correlation between y_t and y_{t-1}, y_{t-2}, ….,y_{t-j+1}) of y_t and ε_t over time. Stationarity is achieved and hence d is determined if autocorrelations between y_t and y_{t-j} become insignificant for increasing j. Moreover, inspection of the partial autocorrelation between y_t and y_{t-j} informs about the order of the AR process p: p should be chosen as the number of lags for which the partial autocorrelation between y_t and y_{t-j} is still significant. In a similar way, the parameter q can be obtained by an inspection of the (partial) autocorrelation of the error terms. A comprehensive procedure to choose the right parameters can be found in Box and Jenkins (1976).

ARMA and ARIMA models can easily be generalized to incorporate the influence of past, current or future values of exogenous factors (x variables) on the observed time-series y_t. These approaches can be extended to ARMAX/ARIMAX by including exogenous variables (Feinberg & Genethliou, 2005). Formally, they can be expressed as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \ldots\ldots + \beta_k x_{t,k} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}\,, \quad (2.12)$$

where β_k denotes the effect of the exogenous variable x_k on the outcome variable y_t. Both ARMA/ARIMA as well as ARMAX/ARIMAX models can be readily estimated using common statistical software packages. The estimates obtained from fitted models can then be used as the basis for predictions used to impute missing values as described for the linear OLS (ordinary least squares) regression above.

Autoregressive conditional heteroscedasticity (ARCH) models

While ARMA/ARIMA models prove to be valid models in many applications, however, the assumption of constant variance of the error terms $E(\varepsilon_t^2)=\sigma^2$ over time might be too restrictive. In finance, for instance, periods of relatively stable stock markets might be followed by periods of crisis and turmoil (Baur & Lucey, 2009) inducing a time-dependent autocorrelation of the error terms with $E(\varepsilon_t^2)=\sigma_t^2$. In stable markets autocorrelation might be relatively high (i.e., prices today will be similar to prices yesterday) and stock price movements are predictable (Fama & French, 1988). In phases of turmoil, however, price movements might be bigger and autocorrelation is lower (Eom, Hahn, & Joo, 2004). In hydrology, the local climate might be characterized by a period of stable conditions followed by change in weather that drastically alters relevant outcomes (Hughes, Cendón, Johansen, & Meredith, 2011). In both examples, the assumption of constant autocorrelation might be too narrow. More realistic would be an assumption of changing variance and hence changing autocorrelation of the observed outcomes over time (heteroscedasticity).

For the reasons stated above, when modeling the outcome variable of interest (y_t), time-series models should focus on its variance and the changes in variance over time. The increased importance of risk and uncertainty considerations in water resources management and hydrological modeling ask for new time series techniques that allow for the modeling of time varying variances.

Auto Regressive Conditional Heteroscedasticity (ARCH) models which originate from finance and econometrics propose a solution to the problem sketched above. Initially proposed by Engle (1982), ARCH models emerged from the observation of volatility clustering in financial markets, in which large changes in prices tend to cluster together. Stock markets often show periods of relative stable trends that are interrupted by periods of turmoil (sometimes caused by crises).

The ARCH model is an extension of more restrictive AR-models with constant autocorrelation of the outcome of interest (Zhu & Wang, 2008). It is a non-linear regression model that in addition to past values of y_t also captures time-varying volatility within the structure of standard time-series models described above. While it is beyond this article to detail the mathematical underpinnings of Engle's work, it should be stressed that ARCH

models are based on the assumption - while holding the unconditional variance of ε_t constant with $E(\varepsilon_t^2) = \sigma^2$ - its conditional variance could follow an AR process of its own with

$$\varepsilon^2 = \zeta + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + \upsilon_t. \quad (2.13)$$

Where υ_t is a white noise process. Based on this specification the ARCH model extends the standard ARMA/ARIMA model to incorporate time-varying volatility. While they require more additional assumptions (see Engle 1982 for technical details), ARCH models and their generalizations have proved useful for modeling flexible time series characterized by non-constant volatility. Moreover, they can generate more accurate forecasts of future volatility and perform better than models that ignore heteroscedascity. For this reason, they could be valuable for hydrological time series modeling in water resource management and flood control applications. Similar to ARMA/ARIMA models, ARCH models can easily be generalized and also allow to model the influence of past, current and future values of exogenous variables x_t on the time-series of interest. Estimation of ARCH is again possible relying on standard statistical software packages and predictions can be used to impute missing values in a time-series.

2.5 Conclusion

Missing data is a common problem in hydrological data and poses a serious problem for many statistical approaches in hydrology. For reasons of convenience, researchers often resort to simple solutions to deal with missing data such as simply discarding observations characterized by missing data or by replacing missing data with a 'naïve' guess (such as the mean of all other observations). Despite their convenience, we have argued that these solutions have severe statistical shortcomings.

Principal component analysis (PCA)-based as well as regression-based imputation methods can improve the accuracy of missing value imputation and reduce statistical problems induced by naïve imputation approaches. However, at the same time they also have disadvantages that should not be neglected. For instance, PCA requires the researcher to choose the number of dimensions on which the higher dimensional data should be projected. However, PCA itself offers only limited guidance on what number of dimensions is optimal and renders this decision in part arbitrary. Frequently, regression-based imputation methods used in practical work are based on linear regression approaches as they are well

understood and easy to implement. In a hydrological setting, however, the assumptions of the linear regression seem to be too restrictive. In particular, the time-series nature of hydrological data as well requires more flexible non-linear models such as the ARIMA and ARCH models that we have discussed above. It should be noted, that there is a multitude of alternative imputation methods based on non-linear regression approaches as well as non-probabilistic algorithms and machine learning approaches such as neural networks, clustering methods or decision tree analysis. While we were not able to cover these approaches in this article, a good introduction to machine learning approaches can be found in Flach (2012). Here we focused our attention on discussing econometric time-series methods as they explicitly model the particular statistical properties of hydrological time-series (autocorrelation and heteroscedasticity) which are mostly neglected in algorithmic machine learning approaches.

It needs to be stressed that there have been few studies concerning imputation of missing data in time series context in hydrology in general. Despite its focus on particular focus on selected methods, our survey clearly shows that there are methodological advances driven by other fields of research that bear relevance for hydrology as well. According to our knowledge, the hydrological community paid little attention to the imputation ability of neither time-series models in general and ARCH models in particular nor other advanced imputation approaches.

We do not address the question of performance advantages of either of these advanced methods in an applied setting. However, we hope that our survey stimulates additional research into these methods and their applicability in hydrology. Whether and to what extent advanced imputation methods lead to more precise hydrological analyses in the presence of incomplete datasets ultimately remains an empirical question. Future research can easily address this question in the context of simulation-based comparisons of different imputation methods within a well-defined hydrological application.

3 Alternative imputation approaches and their performance differences

3.1 Introduction

Complete time series data are a necessary precondition for a variety of statistical approaches that are commonly used in hydrology. For instance, methods such as autocorrelation function, spectrum analysis and extreme value analysis based on the generalized extreme value distribution of annual blocks or principal component analysis all can be applied only to datasets without missing values. Typically, data is usually collected in observation stations over a given period of time (hence time series data) and stored in databases that can subsequently be accessed for research purposes. However, numerous hydrological and research databases contain missing values (Elshorbagy et al., 2002) and are therefore only of limited use to researchers seeking to apply state of the art statistical methods. The reasons behind missing data are multiple and often idiosyncratic. They include failure of observation station, incomparable measurements, manual data entry procedures that are prone to errors and also equipment errors (Johnston, 1999).

Over the last decades, imputation methods which attempt to 'fix' datasets characterized by missing data by replacing them with inserting numerical values have improved dramatically (Peugh & Enders, 2004). The rise of more sophisticated imputation methods led many researchers to prefer replacing missing values with imputed values over excluding them from the analysis entirely (Saunders et al., 2006).

In hydrological settings, the choice of an appropriate imputation method needs to take into account the most important features of hydrological data: Hydrological data are time series data that is often characterized by stable trends over time and a high autocorrelation of the observations. Moreover, hydrological time series often display random deviations from these trends and these deviations are not constant over time (Guzman et al., 2013). Given these features of the data generating process underlying hydrological data, imputation of missing values should be based on statistical time series methods that take into account the time series nature of hydrological data.

In this chapter, we compare the performance of commonly imputation techniques which are widespread and easy to use but ignore the time series nature of the data with imputation techniques exploiting the time series nature of hydrological data. In particular, we are interested in the performance of advanced statistical techniques such as Autoregressive Moving Average/Autoregressive Integrated Moving Average (ARMA/ARIMA) models that have been applied in hydrological settings (Zhang et al., 2011). Moreover, we also evaluate the performance of Autoregressive Conditional Heteroscedasticity (ARCH) time series models which originate from finance and econometrics.

Our performance evaluation is based on hydrological data from the federal state of Brandenburg located in Northeast Germany. We use simulated discharge data that we obtain from a hydrological model for this region as reference data for our performance evaluation. The simulated discharge data does not contain any missing values and reflects typical properties of hydrological data. We randomly delete observations from the reference data and replace the resulting missing values by approximations obtained from different imputation techniques. Comparing the reference data with imputed data allows us to evaluate the performance of the different imputation techniques using the Mean Squared Error (MSE) as well as the Nash Sutcliff Efficiency (NSE) criterion. Our findings indicate that sophisticated time series methods perform significantly better than more commonly used imputation techniques.

The remainder of this chapter proceeds as follows. In Section 3.2, we present the research design of this study in more detail before we move on to a discussion of the research area and the data used in Section 3.3. Section 3.4 is the main part of this chapter in which we evaluate how different imputation techniques perform under different conditions and discuss our findings. Section 3.5 concludes the chapter with a summary of the key findings and a short presentation of the most important implications of this study.

3.2 Working steps

The calibration and application of hydrological researches often require input data with a complete set of observations. In reality, however, observational data is often characterized by missing values. Researchers can replace missing values by applying imputation methods which yield approximations for the missing values derived from the observed data points. There is a multitude of imputation methods available for this purpose and it is not always clear which of the different methods will deliver more satisfactory results in specific applications. We propose a simple research design that allows us to evaluate the performance of different imputation techniques in hydrological settings.

The basic idea of our research design is to use discharge time series data that can be found typically in hydrological applications as reference data. In order to evaluate different imputation methods, we randomly replace a certain fraction of the observations of the reference data with missing values. These missing values will then be replaced by approximations obtained from different imputation methods. Comparing the reference time series data with the imputed time series will allow us to draw conclusions regarding the performance of different imputation methods. Despite this clear structure, it is hard to directly implement this research design for one simple reason: for most of our study regions complete discharge time series for variables of interest hardly exists. Available data often is either characterized by some missing values or with specific values that keep repeating for consecutive days or even weeks – presumably to fill in initially missing values. Therefore, we adjust the basic idea of our research design slightly. In order to obtain reference data that does not suffer from missing values itself, we resort to using output discharge data obtained from a hydrological model. This simulated discharge data is likely to reflect common characteristics of hydrological data found in typical applications. In the following we detail the single steps of our research design which is also summarized in Figure 3.1.

Step 1: Selection of input data for hydrological model

The reference data for the following comparison of different imputation methods is discharge data. As already mentioned above, discharge data without any missing values is hard to find and for this reason we will simulate a time series of discharge data that doesn't suffer from missing values but does reflect typical properties of discharge data (see below). In order to simulate discharge data, we will rely on observed precipitation, temperature and evapotranspiration time series data for 5 years (November 2001 to October 2006) from a given research area.

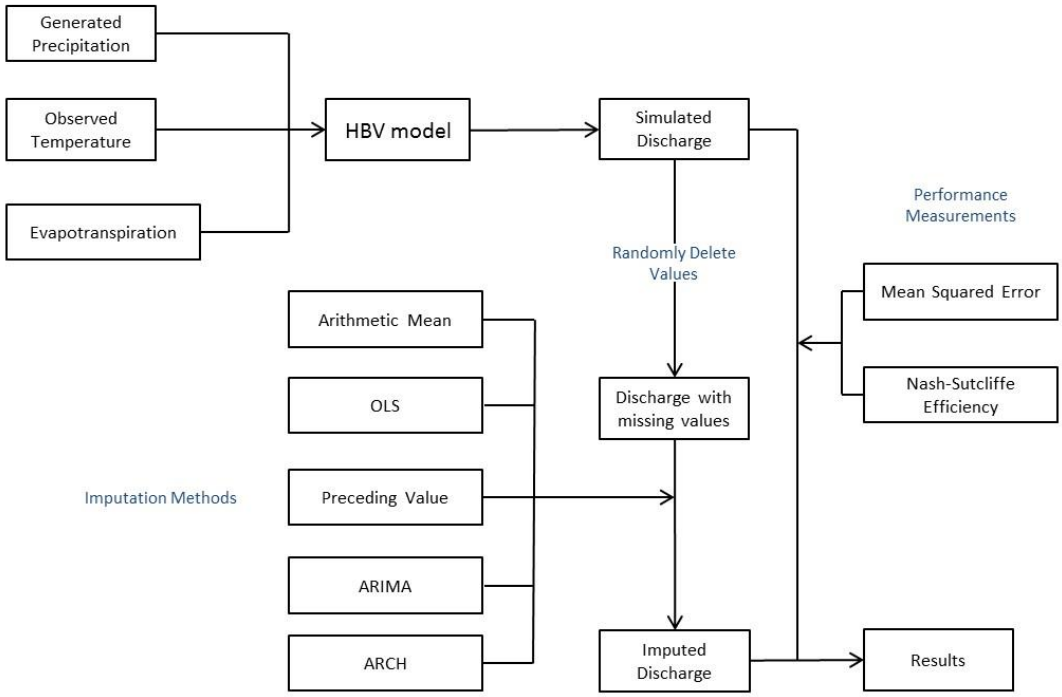

Figure 3. 1: Graphical process of the working steps

Moreover, and in order to learn more about the performance of various imputation methods, we will vary the characteristics of the input data to simulate reference data with different features. In particular, we vary the variance of the original precipitation data (*P_seasonal*) and generate three different precipitation time series: one with low variance (P_low), one with high variance (*P_high*) and one time series where we preserve it's variance but add white noise (*P_noise*). Similarly, we remove seasonality from the original precipitation time series (*P_seasonal*) and obtain precipitation time series without clear seasonality (*P_nonseasonal*).

Step 2: Simulation of reference data

We use the evapotranspiration, observed temperature and the different patterns of the precipitation time series described above as inputs to simulate discharge data by using the Hydrologiska Byråns Vattenbalansavdelning (HBV) hydrology model. The reason we chose precipitation time series to modify because the rainfall is quite dominate during the whole runoff process. By changing different patterns of precipitation time series allows us to generate reference discharge data exhibiting different patterns of variance and seasonality. These differences will help us to identify under which conditions imputation methods might perform differently.

Step 3: Application of imputation methods

We randomly delete a given fraction of observations from the simulated discharge time series obtained in Step 2. In particular, in different steps we delete 5%, 10%, 20%, 30% and 40% of the data. Subsequently, we impute the missing values applying five different imputation techniques to fill the missing values with approximations. We apply imputation techniques commonly used in hydrology – arithmetic mean, ordinary least squares (OLS) and preceding value (PV) – but also imputation techniques that received so far little attention in hydrology – autoregressive integrated moving average (ARIMA) and autoregressive conditional heteroscedasticity (ARCH) models. We are discussing these different imputation methods in chapter 2. The result of step 3 hence is time series including imputed values.

Step 4: Comparison of imputation methods

In the final step, we evaluate the performance of each of the different imputation methods by comparing the imputed time series with the reference time series. We will use different performance criteria (mean squared error and Nash-Sutcliffe efficiency) to determine which imputation method performs best under what conditions.

3.3 Study region and hydrological modeling of discharge data

3.3.1 Study region and input data

Overview

Figure 3. 2: Location of the study area and the gaging station (Federal State of Brandenburg, Germany)

The spatial scope of this study is the federal state of Brandenburg located in Northeast Germany between the rivers Elbe and Oder draining to the Northern Sea and Baltic Sea, respectively (Figure. 3.2). The whole area is 29,479 km2 excluding Berlin in its center. With a mean annual precipitation of 557 mm and a mean annual temperature of 8.7 ℃ (period: 1960-1990; German Weather Service, 2012), it is one of the areas in with the lowest climatic water balance in Germany. Due to high climatic water demand, the evapotranspiration here is approximately 510 mm per year, only leaving 100 mm per year as runoff (Lischeid & Nathkin, 2011). Groundwater flow and groundwater discharge into rivers and channels are the dominating hydrological components of the regional water cycle. About 80 out of 100 runoff per year occurs as ground flow, whereas surface runoff plays only a minor role, accounting for less than of total runoff (Merz & Pekdeger, 2011). Time series of evapotranspiration, observed precipitation, temperature from one of the gaging stations in Bad Wilsnack region in research area was chosen (5 years/ from November 2001 to October 2006) (Figure 3.2).

The whole region is part of a postglacial landscape which formed since the last Pleistocene glaciations. Low gradients in land surface as well as in surface and subsurface flows, a large number of closed depressions and periglacial channels exposing locally raised relief energy, complex interaction of different aquifers and a rather unstable but ecologically crucial interplay between groundwater and streams are major hydrological characteristics of this landscape. Moreover, the region exhibits a wide array of anthropogenic impacts on the fresh systems. These include weirs, dams and locks, flood protection which result in extensive use and alteration of regional freshwater quantity and quality. Due to these specific characteristics, observed discharge time series are disturbed by anthropogenic influences. In order to test different imputation methods relying on more representative reference data, we construct discharge time series by using a hydrological model which is based on observed precipitation, temperature and evapotranspiration. This allows us to simulate discharge as the reference data that is more likely to reflect common characteristics as hydrological time series. For a more detailed description and overview on hydrological changes within this landscape we refer to Merz and Pekdeger (2011) and Germer et al. (2011).

Input data

We will use the HBV model (see below) to simulate a time series of daily discharge Q_s^t which as reference data for the evaluation of different imputation methods. The HBV model requires daily rainfall, temperature and evapotranspiration as input data. These time series have been obtained from Bad Wilsnack region described above for a period of five years (November 2001 to October 2006). ,

Figure 3.3 presents the time series of the evapotranspiration and observed temperature over the observational time period. Note that the left vertical axis contains the temperature scale in degree Celsius whereas the right vertical axis contains the scale for evapotranspiration in mm/day. Both time series are characterized by typical seasonality patterns with low temperatures and low evaporation during winter months.

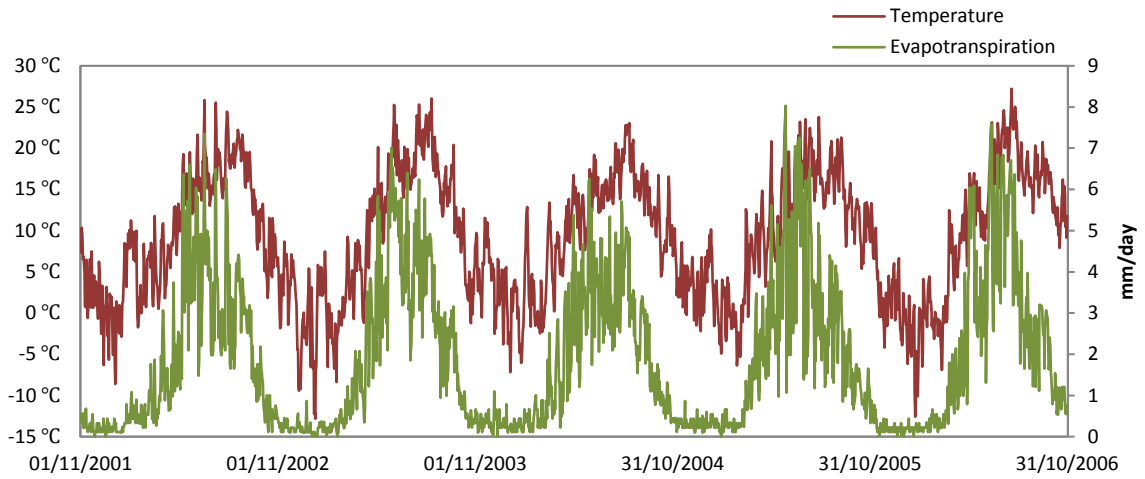

Figure 3. 3: Temperature and Evapotranspiration input data

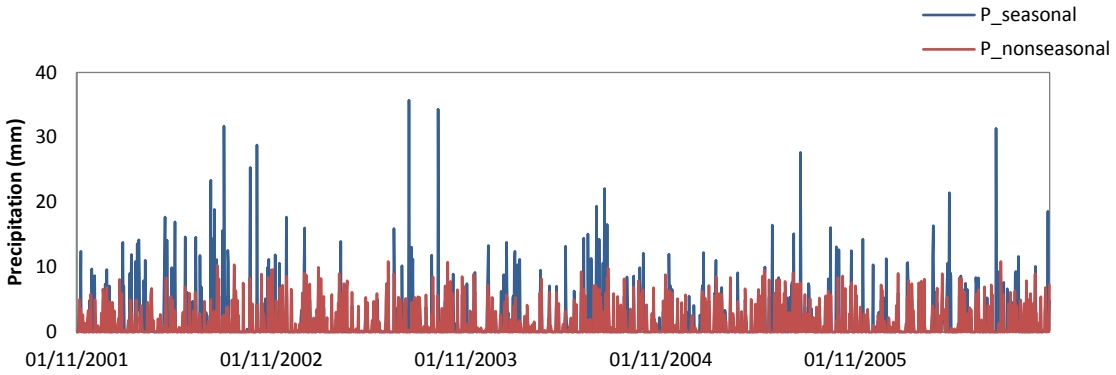

Figure 3. 4: Precipitation input data with/without seasonality

Figure 3.4 presents the time series of the observed precipitation from Bad Wilsnack region where the date between November 2001 and October 2006 on the x-axis and precipitation in mm on the y-axis. Note that Figure 3.4 contains two time series. First, *P_seasonal* is the original time series of precipitation. Second, we de-trended *P_seaosonal* by removing seasonal effects on a monthly base yielding *P_nonseasonal*. We de-trend the time series in order to simulate discharge time series with different structural characteristics using the HBV model. This will allow us to generate insights into performance differences of the imputation methods depending on structural characteristics of the time series to be imputed.

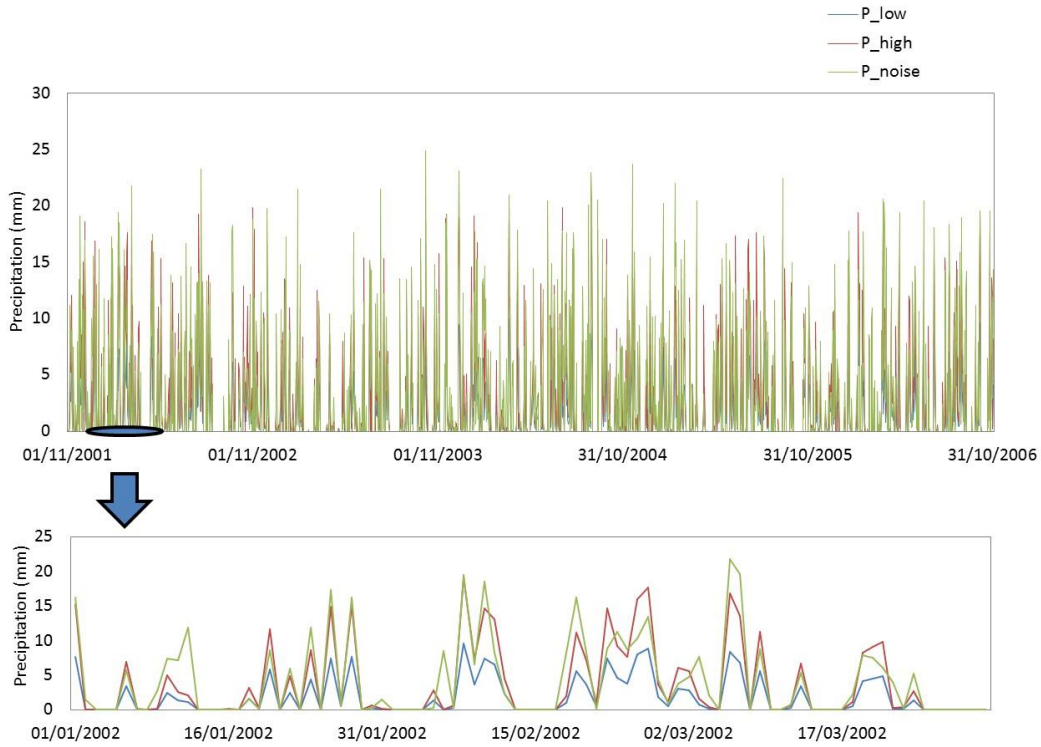

Figure 3. 5: Generated precipitation input data with different variances

Since we are using data from only one region - Bad Wilsnack - in this study, we further manipulate the original precipitation data with regard to its volatility in order to gain further insights how the different imputation methods perform under different conditions. Figure 3.5 presents additional manipulations of the original precipitation data which differ according to the variance. The first manipulation consisted of replacing all values of the original time series that are higher than 10 mm by zero in order to generate a novel time series with low variance (*P_low*). Second, and departing from the derived *P_low*, we increase its variance (and mean) by multiplying *P_low* with a constant multiplier and obtain an additional time series *P_high*. Finally, we preserve *P_high*'s variance but add white noise. White noise here refers to an error term or shock which is drawn from a normal distribution with zero mean and finite variance. Adding independent draws from such a normal distribution to each daily observation yields an additional time series *P_noise* having the same mean as *P_high* but higher variance due to the addition of the random component. Note that Figure 3.5 displays *P_low*, *P_high* and *P_noise* over the full 5 year period (upper half) but also contains a presentation over only three months (January 2002 to March 2002) (lower half). The latter makes typical precipitation patterns and the differences between the three time series visible in a clearer way. Using these different time series as input data for

the HBV model described below allows us to simulate discharge data that reflects different characteristics despite the fact that we work with data from only one catchment.

3.3.2 Hydrological modeling of discharge data

Hydrologiska Byråns Vattenbalansavdelning model

The HBV hydrological model has a long history and the model has found applications in more than 30 countries. Its first application dates back to the early 1970s (Bergström & Forsman, 1973). Originally the HBV model was developed at the Swedish Meteorological and Hydrological Institute (SMHI) for runoff simulation and hydrological forecasting, but the scope of applications has increased steadily (Bergström & Singh, 1995).

Today many versions of the HBV model exist, and new codes are constantly being developed by different groups, see for example Vehvilainen (1986); Killingtveit and Sand (1990); Renner and Braun (1990). The standard at SMHI has long been a version which is best characterized as a semi-distributed conceptual model. Experience has shown, however, that this version has some major drawbacks concerning areal representation, a fact which limits the use of distributed data. There are also a number of physical inconsistencies in this commonly used model, such as the lack of an interception routine and the lack of an elevation correction of evapotranspiration. These inconsistencies became questionable when the model was to be used for climate impact studies.

In 1993 the Swedish Association of River Regulation Enterprises (VASO) and the SMHI initiated a major revision of the structure of the HBV model. The objective of the work was to re-evaluate the existing model and to develop a new model version for hydrological problems related to hydropower production and design. So, the new version of the HBV model, HBV light is based on the same philosophy of simplicity as the original HBV model, but it is more physically reasonable and up-to-date with the current hydrological and meteorological knowledge.

The basic equations are in accordance with the SMHI-version HBV-6 (Bergström, 1992) with only two slight changes. Instead of using initial states the new version uses a ‘warming-up’ period. In the original version, only integer values are allowed for the routing parameter *MAXBAS*. This limitation has been removed in the new version. In order to keep the program

as simple as possible, several functions found in the HBV-6 software were not implemented in the HBV light software. It is possible to use a correction of the long-term mean of potential evaporation values as proposed by Lindström and Bergström (1992). The HBV-light version provides two options which do not exist in the HBV-6 version. The first one is the possibility to include observed groundwater levels into the analysis and the second is the possibility to use a different response routine with a delay parameter.

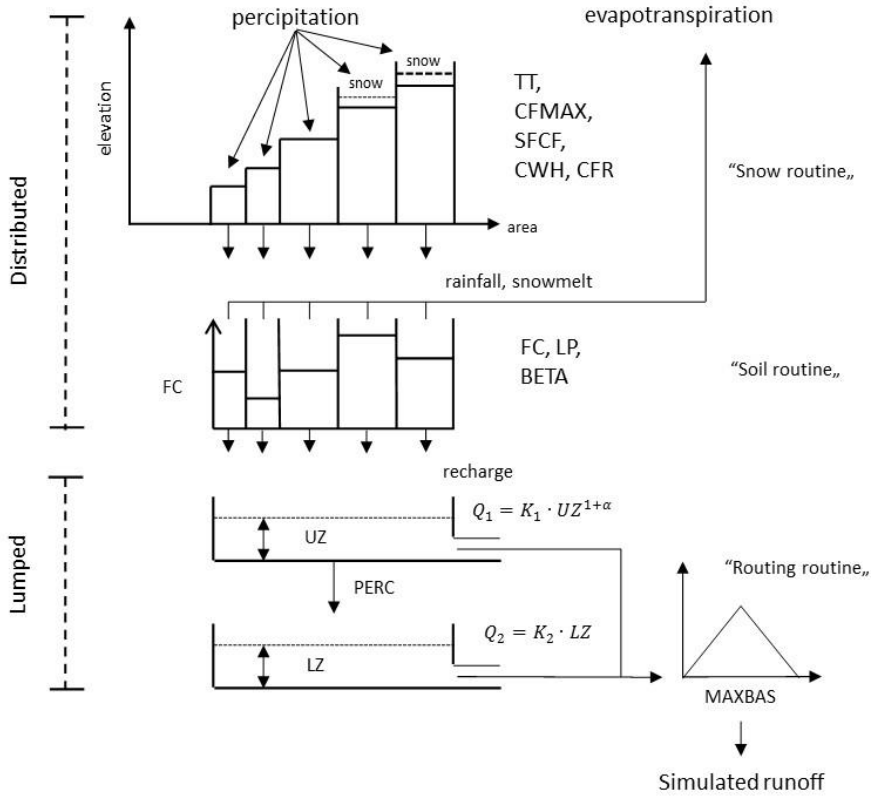

Figure 3. 6: HBV model structure

The model simulates daily discharge using daily rainfall, temperature and potential evaporation as input. Precipitation is simulated to be either snow or rain depending on whether the temperature is above or below a threshold temperature, *TT* (°C). All precipitation simulated to be snow, i.e. falling when the temperature is bellow *TT* (°C), is multiplied by a snowfall correction factor, *SFCF.* Snowmelt is calculated with the degree-day method according to Equation 3.1.

$$melt = CFMAX(T(t) - TT) \tag{3.1}$$

Melt water and rainfall is retained within the snowpack until it exceeds a certain fraction, CWH, of the water equivalent of the snow. Liquid water within the snowpack refreezes according to Equation 3.2

$$refreezing = CFR \cdot CFMAX\big(TT - T(t)\big)\,. \tag{3.2}$$

Rainfall and snowmelt (*P*) are divided into water filling the soil box and groundwater recharge depending on the relation between water content of the soil box (SM (mm)) and its largest value (*FC* (mm)) (Equation 3.3)

$$\frac{recharge}{P(t)} = \left(\frac{SM(t)}{FC}\right)^{BEAT}\,. \tag{3.3}$$

Actual evaporation from the soil box equals the potential evaporation if *SM/FC* is above *LP* while a linear reduction is used when *SM/FC* is below *LP* (Equation 3.4)

$$E_{act} = E_{pot} \cdot min\left(\frac{SM(T)}{FC \cdot LP}, 1\right). \tag{3.4}$$

Groundwater recharge is added to the upper groundwater box and the water percolates from upper to the lower groundwater box. Runoff from the groundwater boxes is computed as the sum of two linear outflows by linear reservoir function (Equation 3.5)

$$Q_{GW_{(t)}} = Q_1 + Q_2 = K_1 \cdot UZ^{1+\alpha} + K_2 \cdot LZ. \tag{3.5}$$

The recession components threshold of upper groundwater box is defined by a linear drainage equation. The runoff is finally transformed by a triangular weighting function to give the simulated runoff (Equation 3.6)

$$Q_{sim(t)} = \sum_{i=1}^{MAXBAS}\left(\int_{i-1}^{i}\frac{2}{MAXBAS} - \left|u - \frac{MAXBAS}{2}\right|\frac{4}{MAXBAS^2}\,du\right)\cdot Q_{GW_{(t-i+1)}}\,. \tag{3.6}$$

Where *P(t), T(t), SM(t),* $Q_{GW(t}$ and $Q_{sim(t)}$ are precipitation, temperature, soil moisture, ground water discharge and simulated discharge at time *t*. *CFMAX*, *CFR*, *FC*, *LP*, K_1, K_2, α and *MAXBAS* are model parameters.

For both the snow and soil routine, calculations are performed for each different elevation zone, but the response routine is a lumped representation of the catchment.

Table 3. 1: Model parameters and feasible ranges

Parameter (Unit)	Explanation	Feasible ranges
Snow routine		
TT (°C)	Threshold temperature	(-2, 0)
CFMAX (mm/°C/d)	Degree-day factor	(0.2, 1)
SFCF	Snowfall correction factor	(1, 4)
CFR	Refreezing coefficient	0.05
CWH	Water holding capacity	0.1
Soil routine		
FC (mm)	Maximum of storage in the soil	(200, 850)
LP (mm)	Threshold for reduction of evaporation	(0.2, 1)
BETA	Shape coefficient	(1, 4)
Response routine		
Alpha	Response box parameter	(0, 0.5)
K1 (1/d)	Recession coefficient (upper storage)	(0.07, 0.2)
K2 (1/d)	Recession coefficient (lower storage)	(0.005, 0.07)
PERC (mm/d)	Percolation from upper to lower response box	(1, 2.5)
Routing routine		
MAXBAS (d)	Transformation function parameter	(2, 5)

Simulation results obtained from the HBV model

Below, we briefly summarize the simulated time series Q_s^t we obtained from applying the HBV model to the original input data obtained from Brandenburg and the derived precipitation time series. In total, we simulated five different discharge time series. Figure 3.7 presents $Q_{seasonal}^t$ as well as $Q_{non-seasonal}^t$ based on the original as well as the de-trended precipitation data. Note, that $Q_{non-seasonal}^t$ unsurprisingly displays much less pronounced seasonality patterns than $Q_{seasonal}^t$. Remaining seasonality effects are due to seasonality in the other input variables, temperature and evapotranspiration.

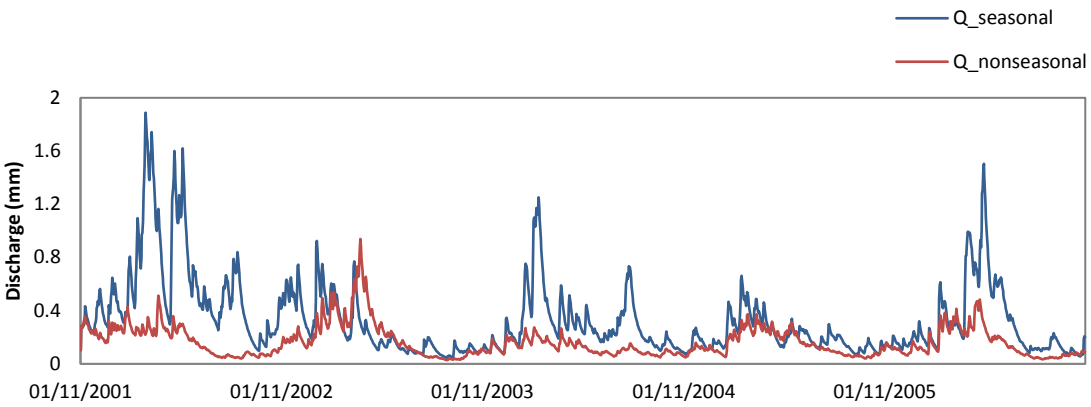

Figure 3. 7: Simulated discharge output data with/without seasonality

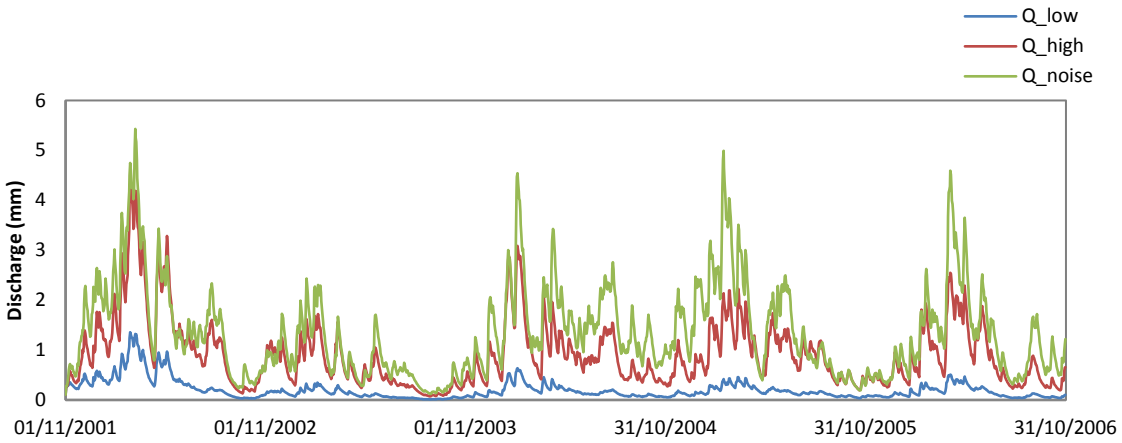

Figure 3. 8: Simulated discharge output data with different variances

Figure 3.8 presents the time series of the simulated discharge data which are based on precipitation inputs with manipulated variance, i.e., Q_{low}^t, Q_{high}^t and Q_{noise}^t. Note, that Q_{low}^t displays much less variance than Q_{high}^t and Q_{noise}^t. Since white noise is added to the input data P_high for creating P_noisy, Q_{noise}^t is characterized by higher fluctuations than Q_{high}^t but preserves its mean.

3.4 Evaluation of imputation methods

3.4.1 Overview on imputation methods used

Before applying different imputation methods to the simulated discharge time series Q_s^t obtained from applying the HBV model to the observed data, we briefly discuss different imputation methods. As described above, we will apply these methods to impute different shares of missing values (5%, 10%, 20%, 30%, 40%) in order to obtain a time series Q_i^t

including imputed values. For the following notation, we denote with Q_m^t the time series of including missing values which will be the basis for our imputation exercises. After the discussion of the different imputation methods used, we assess their performance using the Mean Squared Error (MSE) and the Nash-Sutcliff Efficiency (NSE) criteria which we also introduce below. Pros and cons of all imputation approaches we use in this chapter have been discussed in chapter 2.

Arithmetic mean imputation

A commonly used and simple to implement imputation method for the approximation of missing values is the so-called arithmetic mean imputation. It replaces missing values in a variable with the arithmetic mean of the observed values of the same variable (Roth, 1994). In our context, the missing values are replaced with the arithmetic mean of the non-missing observed values, which is $Q_i^t = \frac{1}{T}\sum_{i=1}^{T} Q_m^t$ with T being the number of non-missing observations here.

Preceding value

An alternative approach to replace missing values is using the last observed preceding value as best predictor for a missing values. Missing values in that case sequentially replaced according to $Q_i^t = Q_m^{t-k}$ where k is the difference in the number of periods between a missing value and the last observed value of Q. If, for instance, two missing values occur subsequently, the second missing value is replaced with $Q_i^t = Q_m^{t-2}$ as Q_m^{t-2} is the last previously observed value.

Ordinary least squares (OLS) regression imputation

Regression-based imputation replaces missing data with predicted values from a regression estimation (Greenland & Finkle, 1995). The basic idea behind this method is using information from all observations with complete values in the variables of interest to fill in the incomplete values which is intuitively appealingly (Frane, 1976). While different regression models can be applied to impute missing values, we start with the most basic regression model – the linear regression.

The first step of the imputation process is to estimate regression equations that relates the variable that contains missing data (the dependent variable of the regression) to a set of

variables which have complete information across all observations in the data set (independent variables of the regression). In our context, we estimate how the non-missing values Q_m^t are related to the observed precipitation data on the same day P_o^t. The regression function we are estimating is then given by $Q_m^t = \beta_0 + \beta_1 P_m^t + \varepsilon_t$, where ε_0 accounts for measurement errors and other unobserved influences on discharge. The regression parameter β_0 and β_1 are estimated only for the subset of the data that contains all observations that have complete information both for the dependent variable and the independent variables using the ordinary least square estimator yielding the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

The second step uses the regression results from the first step and missing values for the observations that could not have been included in the regression are replaced by predictions obtained from combining the observed values precipitation and the estimates from the first step of how it is related to the discharge. These predicted values fill in the missing values and produce a complete data set in which the missing values are replace according to $Q_i^t = \widehat{\beta_0} + \widehat{\beta_1} P_o^t$ for all *t* with missing data.

While regression-based imputations most frequently rely on simple linear regressions, it is worth noting that more flexible regression approaches can equally be used and might even be more advantageous depending on the application. We will discuss more advanced time series regression approaches below.

Auto Regressive Integrated Moving Average Model

Similar to the linear regression framework introduced above, time series regressions can equally be employed for imputations purposes. Imputed values are then derived from a prediction based on time series regression instead of a linear regression.

A time series – such as hydrological data – can be interpreted as a stochastic process where y_t and y_{t-j} are correlated over time, i.e., autocorrelation between different measures of *y* exists. One possible specification is an autoregressive process AR (*p*) of *pth* order with

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t, \quad (3.7)$$

In (3.7) epsilon is a random error term that follows a standard normal distribution and is independent over time with $E(\varepsilon_t, \varepsilon_{t-i}) = 0$ *for all* $i \neq t$.

p here denotes the number of lagged values of y_t that enter the process. ε_t is an identically distributed (iid) error term with zero mean and constant variance. An alternative specification of a stochastic process that generates autocorrelation in a time series are moving average (MA) processes in which the contemporary value of y_t is a function of its mean μ and a sequence of random shocks with

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots\ldots + \theta_p\varepsilon_{t-q}\,. \tag{3.8}$$

The commonly used ARMA model fits an observed autoregressive (AR) time series by combining it with a moving-average (MA) component consisting of a sum of weighted lags of the error term ε_t (Box, Jenkins, Reinsel, & Ljung, 2015). The resulting ARMA model is written as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q}\,. \tag{3.9}$$

Equation (3.9) is often referred to as an ARMA(p,q) model as it contains a *pth*-order autoregressive component in the observable time series, y_t, and a *qth* order moving average component of the unobservable random shocks ε_t. It is generally assumed that ε_t follows a so-called white-noise process with zero mean $E(\varepsilon_t)$ and constant variance $E(\varepsilon_t^2)=\sigma^2$.

It is important to highlight that ARMA models can be fitted to data only if the underlying time series y_t is weakly stationary (see chapter 2). In case a time series y_t is not stationary, stationarity can often be achieved by differencing the time series one or more times (Box & Jenkins, 1976). If differencing is required the ARMA (p,q) model (Autoregressive Moving Average) becomes an ARIMA (p,d,q) model (Autoregressive Integrated Moving Average) where d denotes the order of differencing, i.e., the number of time y_t is differenced to achieve stationarity.

Both ARMA and ARIMA models can be readily estimated using common statistical software packages such as R or STATA. Before the actual estimation, the researcher has to identify whether the time-series is stationary or needs to be differenced in order to induce stationarity. In practice, this can be done by a first visual inspection of the data. More formally, the unit root test of Dickey and Fuller (1979) can be used to test whether differencing is needed. Additionally, the researcher has to determine the appropriate number of lagged terms for the ARMA/ARIMA model, i.e., has to choose p and q. The choice

of these parameters typically is based on an analysis of the Autocorrrelation Function (AFC) and the Partial Autocorrelation Function (PACF). After the model has been specified by the researcher, its parameters need to be estimated. Since ARIMA models typically are non-linear models, parameter estimation requires non-linear model fitting procedures. In the context of ARIMA models, typically maximum likelihood or method of moments approaches are applied. The application of maximum likelihood estimation requires to assume a particular distribution of the error term ε_t (typically a i.i.d. normal distribution). This allows to formulate a joint distribution function expressed in terms of y_t. The maximum likelihood framework chooses the unknown parameters α_p and θ_q in a way joint likelihood function gets maximized conditional the observed data. An alternative estimation approach is to apply a Method of Moments estimator which equates sample moments to population moments and solves for the unknown parameters. Maximum likelihood estimators can be shown to be consistent and to be asymptotically normally distributed which allows the construction of confidence intervals around the point estimates in order to conduct hypothesis testing. Alternatively, Yule and Walker propose a Method of Moments estimator for ARIMA models which, however, can be shown to be not efficient (Brockwell & Davis, 2013). For this reason, in our application, we fit ARIMA (*p,d,q*) models to the data and use the estimates obtained from maximum likelihood approaches as the basis for predictions used to impute missing values as described for the linear OLS regression above.

Autoregressive Conditional Heteroscedasticity Model

ARMA and ARIMA models are based on the assumption of constant variance of the error terms $E(\varepsilon_t^2)=\sigma^2$ over time. This assumption often is too restrictive. In hydrology, the local climate might be characterized by a period of stable conditions followed by change in weather that drastically alters relevant outcomes (Hughes et al., 2011). The assumption of constant autocorrelation is then too narrow. More realistic would be an assumption of changing variance and hence changing autocorrelation of the observed outcomes over time (heteroscedasticity). Auto Regressive Conditional Heteroscedasticity (ARCH) models which originate from finance and econometrics are regression models that in addition to past values of y_t also captures time varying volatility within the structure of standard time series models described above. ARCH models are holding the unconditional variance of ε_t constant with $E(\varepsilon_t^2) = \sigma^2$ but allow its conditional variance to follow an AR process of its own with

$$\varepsilon^2 = \zeta + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + \upsilon_t \,, \tag{3.10}$$

where υ_t is a new white noise process.

Based on this specification the ARCH model extends the standard ARMA/ARIMA model to incorporate time varying volatility. The estimation of ARCH is again possible relying on standard statistical software packages and predictions can be used to impute missing values in a time series. Similar to ARIMA models, the procedure most often used to estimate ARCH models is the maximum likelihood methods. As described above, auxiliary assumptions have to be made (in particular, an i.i.d. normal distribution of the error terms) in order to derive the likelihood function to be maximized conditional on the observed data by choosing the parameters of interest. If the normal distribution of the error terms is hard to justify in practical application, the estimation can be based on more general Quasi Maximum Likelihood (QMLE) approaches which, however, is not efficient. In our case, we fit an ARCH model that extends and ARIMA (*p,d,q*) model by a first-order autoregressive process for the variance of the error term ε_t^2 and estimate the unknown parameters relying on maximum likelihood estimation.

3.4.2 Evaluation of imputation performance

We will evaluate the performance of the two different imputation methods by comparing the imputed time series with the reference time series obtained from the HBV model described above. In particular, we use the Mean Squared Error (MSE) and the Nash-Sutcliffe efficiency (NSE) measure for this purpose. We quickly discuss the two measures below before we discuss the efficiency of our imputation exercise.

Mean Squared Error (MSE)

The Mean Squared Error is a commonly used measure in statistics to assess the quality of an estimator or – as in the case of imputation – a predictor (Harville & Jeske, 1992). The MSE measures the average of the squares of the errors or deviations, i.e., the difference between the predictions and the observed values (Schunn & Wallach, 2005). Note that the MSE can be compared across different models in order to assess which performs better.

Formally, let Q_s^t be the simulated discharge time series (our reference data) and Q_i^t be the time series of discharge including imputed values from one of the imputation methods for the periods $t = 1, \cdots, T$. The MSE is then defined as

$$MSE = \frac{1}{T}\sum_{i=1}^{T}(Q_i^t - Q_s^t)^2. \tag{3.11}$$

A MSE of zero would indicate error-free prediction (imputation) of missing values, but is in reality not to achieve.

Nash-Sutcliffe efficiency (NSE)

Nash and Sutcliffe (1970) proposed an efficiency measure for hydrological models. The Nash-Sutcliffe efficiency is defined as one minus the sum of the squared differences between the predicted Q_i^t and observed values Q_s^t, normalized by the variance of the observed values during the period under investigation with:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_s^t - Q_i^t)^2}{\sum_{t=1}^{T}(Q_s^t - \overline{Q_s})^2}. \tag{3.12}$$

The range of the NSE lies between 1.0 (perfect fit) and $-\infty$. An efficiency of lower than zero indicates that the mean value of the observed time series would have been a better predictor than the model. In this case, the imputation method performs worse than a simple imputation based on the mean of the observed data.

Note that the NSE is related to the MSE. It can be interpreted as dividing MSE by the variance of the observations and subtracting that ratio from 1 with

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_s^t - Q_i^t)^2}{\sum_{t=1}^{T}(Q_s^t - \overline{Q_s})^2} = 1 - \frac{MSE}{\sigma_{Q_s}^2}. \tag{3.13}$$

3.4.3 Results

Mean Squared Error

In a first step, we evaluate how the different imputation mechanisms perform by applying the MSE criterion discussed above before moving on to the NSE results. All results are presented both graphically (see Figures 3.9 to 3.10) and in tables (see Table 3.2). In addition to the simple imputation methods (naïve and mean based imputations), they have been

obtained from the following regression models: First, we fit a linear regression in which we regress Q_S^t on P_o^t and use the obtained parameter estimates in order to predict missing values in Q_S^t. Second, we estimate an ARIMA(1,1,1) model which is based on the differenced time-series (in order to induce stationarity) of Q_S^t and includes the first lag of the dependent variable (AR(1)) as well as a first-order moving average component (MA(1)). Again, we used the obtained parameter estimates to predict missing values in Q_S^t. Finally, the presented ARCH results are derived from the an ARIMA(1,1,1) model which has been extended by a first-order AR process of the squared error term to allow for heteroscedasticity.

Independently of which of the five reference time series Q_S^t we focus, clear patterns from the imputation simulations emerge. First, the MSE monotonously increase in the share of data points that are missing from a data set irrespectively of the imputation technique applied. This is unsurprising, as by definition, a smaller share of missing values implies a higher share of identical values in both the reference time series Q_S^t and the imputed time series Q_I^t and hence a smaller MSE. Moreover, most imputation methods perform better in cases where only few observations are missing as the approximations for the missing values will be based on a relatively larger number of complete observations.

Second, we observe clear performance differences in the different types of imputation techniques used. Most importantly, imputation techniques that ignore the time series character of the data to be imputed perform significantly worse than imputation methods that explicitly take the time series nature of the data into account. In particular, both the results from arithmetic mean imputations as well as the results from OLS-based imputations are characterized by similarly high MSEs relative to the other methods. Imputations techniques that account for the time series nature of the data (preceding value, ARIMA and ARCH) perform significantly better in terms of MSE. In fact, their MSEs are by a factor of 20 to 40 times smaller than the MSEs observed for mean imputation and OLS-based imputation (see Table 3.2). Within the approaches that exploit the time series structure of the data, the flexible ARCH model performs best with its MSEs being clearly smaller than those of the ARIMA model. While the preceding value imputation clearly is superior mean value or OLS-based imputations, it is outperformed by the more sophisticated time series models. Moreover, the outperformance of ARIMA/ARCH models over the preceding value technique

is more pronounced in situations where a large fraction of observations is characterized by missing values (see Figure 3.10).

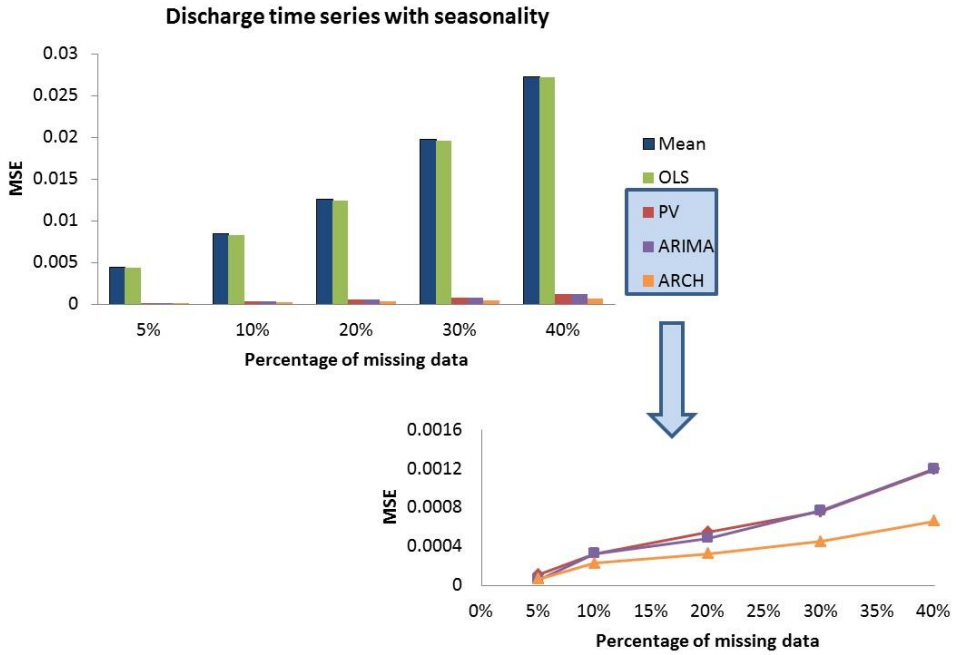

Figure 3. 9: Mean Squared Error of imputation methods for seasonality

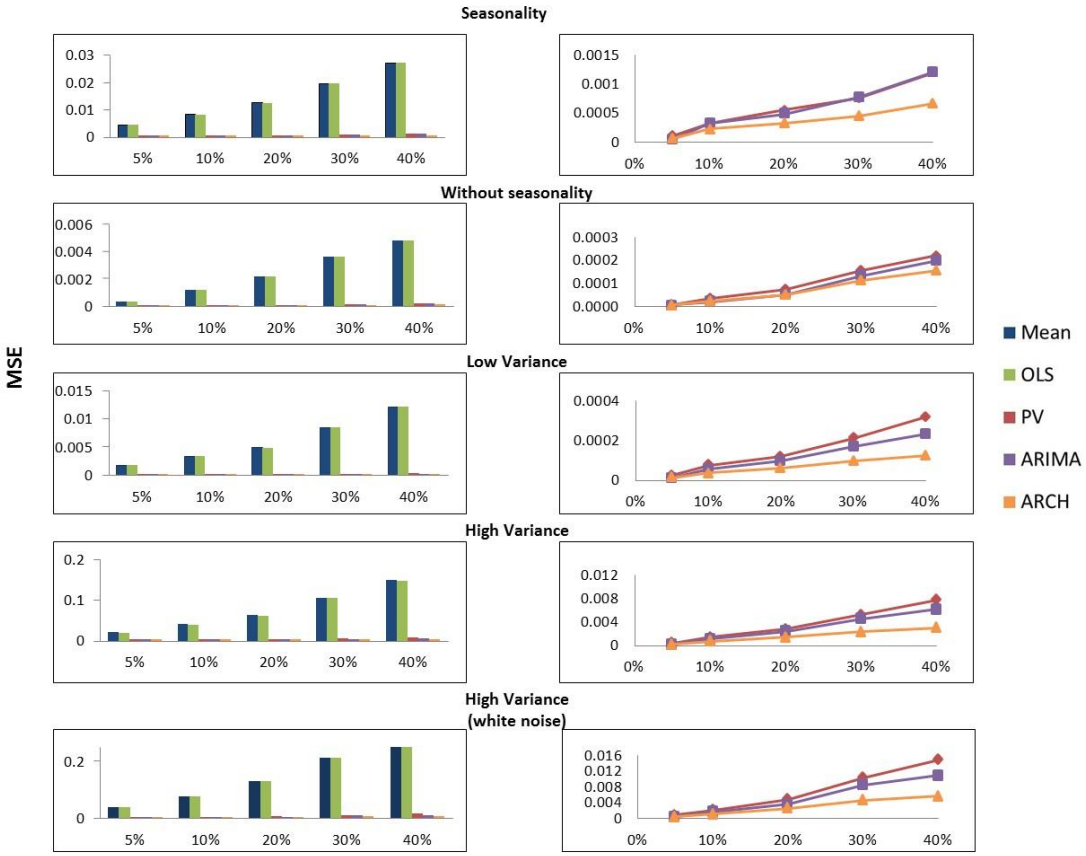

Figure 3. 10: Mean Squared Error of imputation methods for different scenarios

Table 3. 2: Results of Mean Squared Error

Discharge time series without seasonality

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	3.58E-04	3.59E-04	8.49E-06	7.23E-06	6.84E-06
10%	1.23E-03	1.23E-03	3.74E-05	2.21E-05	2.50E-05
20%	2.21E-03	2.21E-03	7.45E-05	5.22E-05	5.21E-05
30%	3.60E-03	3.60E-03	1.57E-04	1.34E-04	1.12E-04
40%	4.79E-03	4.79E-03	2.21E-04	1.99E-04	1.56E-04

Discharge time series with seasonality

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	4.48E-03	4.32E-03	1.14E-04	5.57E-05	6.73E-05
10%	8.50E-03	8.24E-03	3.29E-04	3.29E-04	2.28E-04
20%	1.26E-02	1.24E-02	5.53E-04	4.85E-04	3.30E-04
30%	1.98E-02	1.96E-02	7.61E-04	7.70E-04	4.55E-04
40%	2.73E-02	2.71E-02	1.20E-03	1.20E-03	6.63E-04

Discharge time series with low variance

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	1.75E-03	1.72E-03	2.42E-05	1.03E-05	1.27E-05
10%	3.41E-03	3.35E-03	7.54E-05	5.76E-05	3.71E-05
20%	4.91E-03	4.88E-03	1.19E-04	9.71E-05	6.07E-05
30%	8.45E-03	8.49E-03	2.10E-04	1.72E-04	9.80E-05
40%	1.21E-02	1.22E-02	3.17E-04	2.35E-04	1.23E-04

Discharge time series with high variance

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	2.04E-02	1.99E-02	4.66E-04	2.43E-04	2.64E-04
10%	4.04E-02	3.93E-02	1.49E-03	1.11E-03	7.76E-04
20%	6.29E-02	6.20E-02	2.84E-03	2.39E-03	1.42E-03
30%	1.06E-01	1.06E-01	5.36E-03	4.54E-03	2.35E-03
40%	1.49E-01	1.49E-01	7.75E-03	6.13E-03	3.02E-03

Discharge time series with high variance (white noise)

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	3.84E-02	3.81E-02	7.86E-04	5.25E-04	4.66E-04
10%	7.72E-02	7.69E-02	2.13E-03	1.61E-03	1.05E-03
20%	1.30E-01	1.29E-01	4.78E-03	3.67E-03	2.51E-03
30%	2.13E-01	2.12E-01	1.03E-02	8.44E-03	4.52E-03
40%	3.01E-01	3.00E-01	1.48E-02	1.08E-02	5.68E-03

It is worth noting, that the performance differences across the different imputation methods are independent of the particular characteristics of the reference time series. As discussed above, we evaluated the performance of the different imputation techniques using five reference time series which differ regarding the existence of seasonal trends and their variance. The ranking and the relative difference between the five tested imputation methods is similar across all five reference time series. Not surprisingly, however, comparisons of the results within the different imputation methods reveal that their performance depends significantly on the characteristics of the reference time series. The higher the variance of the reference time series is, the more challenging imputation becomes and MSEs within a given imputation technique increase for reference time series with higher volatility. We also observe that MSEs are higher if seasonal trends are present compared to the MSEs obtained for the reference time series where we removed seasonality.

Nash-Sutcliffe efficiency

In addition to using the MSE criterion, we also evaluate the performance of the different imputation methods by applying the NSE criterion (see Section 3.4.2). Note that given the NSE is a function of the MSE (with $NSE = 1 - \frac{MSE}{\sigma^2_{Q_S}}$) the patterns discussed above hold also when the NSE criterion is applied.

In fact, and most importantly, imputation methods that acknowledge the time series nature of the reference data (preceding value, ARIMA and ARCH) perform significantly better than the other methods (mean imputation and OLS) with the flexible ARCH model achieving NSEs that are closest to the maximum possible (see Figure 3.12 and Table 3.3).[2] The effect of an increasing variance/seasonality on the performance is different when the NSE criterion is applied when compared to the MSE criterion as the NSE criterion uses the reference time series' volatility as normalizing denominator in its definition. As a result, the observed NSE values across different volatility scenarios are less sensitive to changes in the volatility of the underlying reference data than the MSE.

[2] The NSE's value range is bound between minus infinity and 1. NSEs of 1 indicate that the prediction mimics the reference data perfectly.

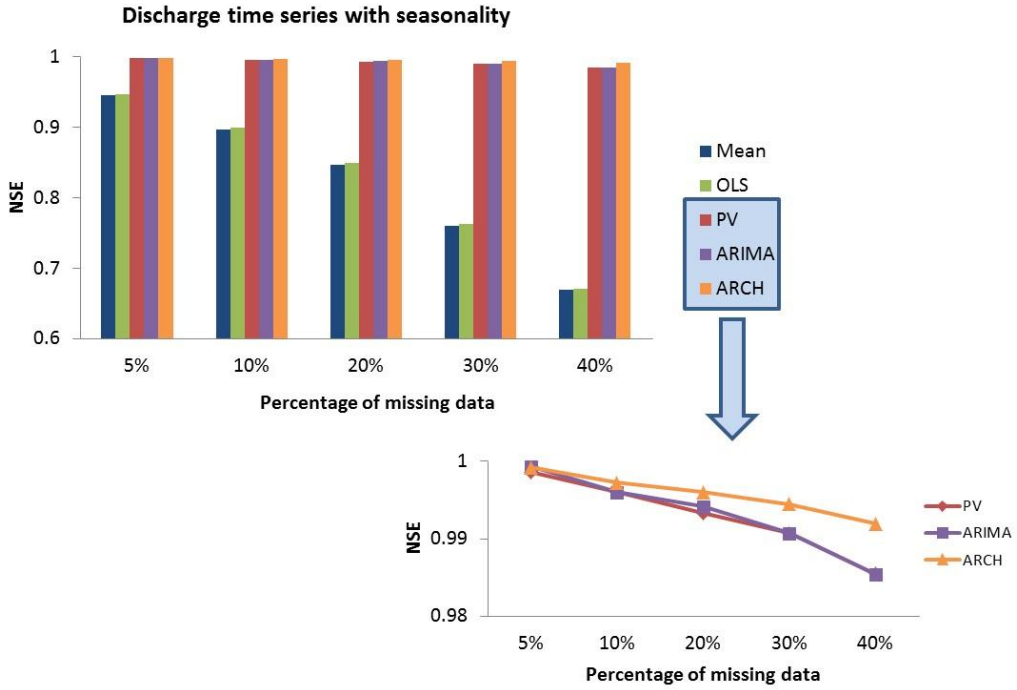

Figure 3. 11: Nash-Sutcliffe efficiency of imputation methods for seasonality

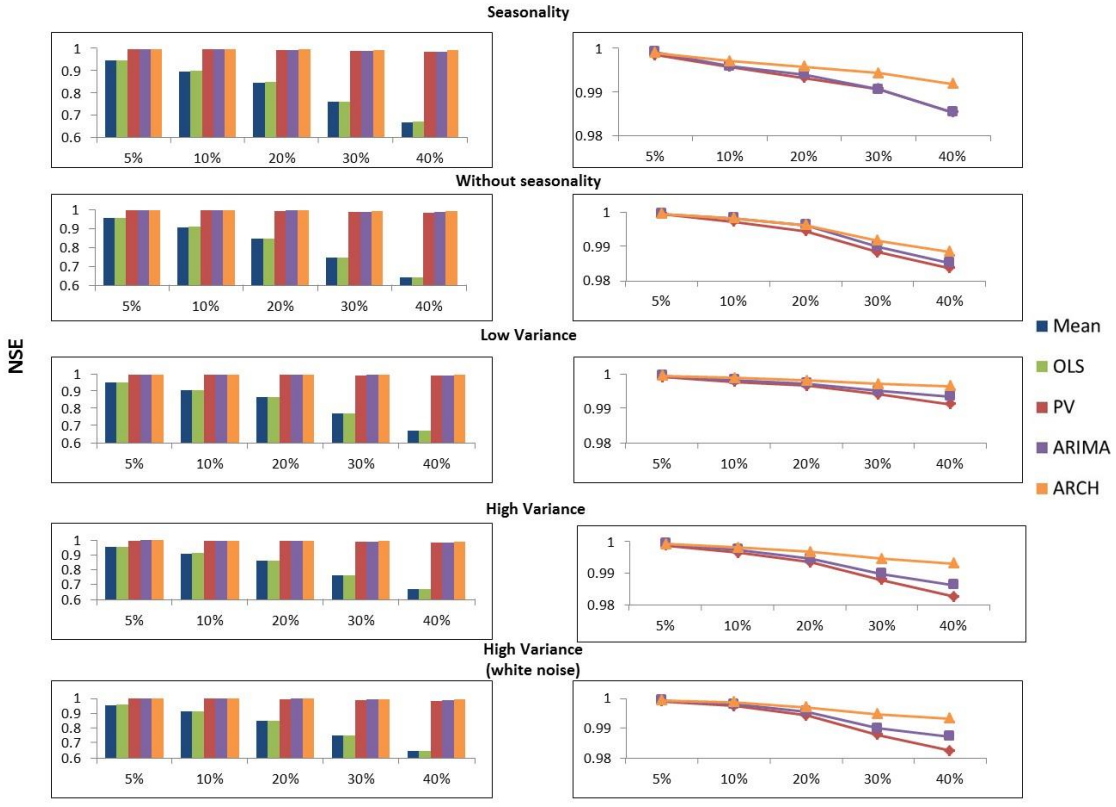

Figure 3. 12: Nash-Sutcliffe efficiency of imputation methods for different scenarios

Table 3. 3: Results of Nash-Sutcliffe efficiency

Discharge time series seasonality

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	0.946	0.948	0.999	0.999	0.999
10%	0.897	0.900	0.996	0.996	0.997
20%	0.847	0.850	0.993	0.994	0.996
30%	0.760	0.763	0.991	0.991	0.994
40%	0.669	0.671	0.985	0.985	0.992

Discharge time series without seasonality

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	0.973	0.973	0.999	0.999	0.999
10%	0.909	0.909	0.997	0.998	0.998
20%	0.836	0.836	0.994	0.996	0.996
30%	0.733	0.732	0.988	0.990	0.992
40%	0.644	0.644	0.984	0.985	0.988

Discharge time series with low variance

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	0.953	0.953	0.999	1.000	1.000
10%	0.908	0.909	0.998	0.998	0.999
20%	0.867	0.868	0.997	0.997	0.998
30%	0.771	0.770	0.994	0.995	0.997
40%	0.672	0.669	0.991	0.994	0.997

Discharge time series with high variance

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	0.955	0.956	0.999	0.999	0.999
10%	0.910	0.913	0.997	0.998	0.998
20%	0.860	0.862	0.994	0.995	0.997
30%	0.765	0.765	0.988	0.990	0.995
40%	0.669	0.669	0.983	0.986	0.993

Discharge time series with high variance (white noise)

Percentage of missing data	Mean	OLS	PV	ARIMA	ARCH
5%	0.955	0.955	0.999	0.999	0.999
10%	0.909	0.909	0.997	0.998	0.999
20%	0.846	0.847	0.994	0.996	0.997
30%	0.748	0.749	0.988	0.990	0.995
40%	0.644	0.645	0.982	0.987	0.993

3.5 Application of ARIMA/ARCH models for groundwater time series

In Section 3.4, we found that ARIMA and ARCH models perform significantly better in imputing missing hydrological data than alternative and widely used methods that do not consider the characteristic of time series data. In this section, we additionally apply ARIMA and ARCH models to ground water time-series which observed from Lake Bötzsee. The region is about 20 km northeast of Berlin, also in Brandenburg in Northeast Germany in the time period from January 2012 to May 2014. We do so in order to validate the performance advantage of time series models in an additional context beyond the one described in Section 3.4. In this endeavor, we not only model the observed ground water time series (GWBR1) in order to impute missing values following the identical approach as in Section 3.4. In addition, we also model artificially smoothed versions of the observed ground water time series in order to analyses how different degrees of volatility in a time series affect the relative performance of ARIMA and ARCH models. We have three additional time series that have been smoothed by Moving Average (MA) processes by three different levels (MA101, MA501, MA1001). Figure 3.13 shows the four different groundwater time-series. For each of these time-series, we fit an ARIMA(0,1,2) model which is based on the differenced time-series (in order to induce stationarity) of Q_s^t and includes a first- and second order moving average component (MA(2)). We found this parametrization to fit the smoothened data best. We use the obtained parameter estimates to predict missing values in Q_s^t. The presented ARCH results are derived from the an ARIMA(0,1,2) model which has been extended by a first-order AR process of the squared error term to allow for heteroscedasticity.

The results from this exercise are relatively clear and can be summarized as follows: ARCH models consistently outperform ARIMA models in their imputation performance also in this setting. Additionally, the performance differences between ARIMA and ARCH models seem to be relatively unaffected by the applied smoothing.[3]

The detailed results for comparisons according to the Mean Squared Error can be found in Table 3.4 whereas comparisons according to the Nash Sutcliff Efficiency criterion can be found in Table 3.5.

[3] Note that we estimated ARIMA (0,1,2) models as well as ARCH (1) based on ARIMA(0,1,2) models as the best fit the ground water data observed.

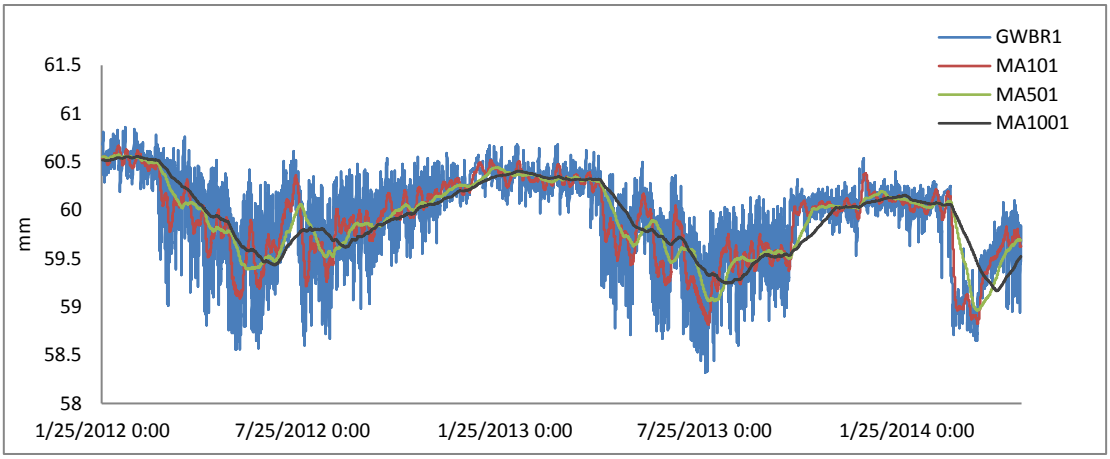

Figure 3. 13: Observed ground water time-series from Lake Bötzsee and three smoothed time-series

Table 3. 4: Mean Squared Error of imputation application for groundwater time-series

Missing data percentage	GWBR1			
	PV	MEAN	ARIMA	ARCH
5%	2.34E-04	1.17E-02	1.13E-04	9.07E-05
10%	5.58E-04	2.36E-02	2.85E-04	2.00E-04
20%	1.42E-03	4.56E-02	9.07E-04	4.85E-04
30%	2.82E-03	6.80E-02	1.99E-03	9.01E-04
40%	5.39E-03	9.01E-02	4.22E-03	1.50E-03
	MA101			
	PV	MEAN	ARIMA	ARCH
5%	8.00E-07	8.80E-03	2.00E-07	1.00E-07
10%	2.10E-06	1.73E-02	6.00E-07	3.00E-07
20%	5.70E-06	3.37E-02	2.70E-06	7.00E-07
30%	1.29E-05	5.00E-02	7.60E-06	1.50E-06
40%	2.51E-05	6.59E-02	1.76E-05	3.00E-06
	MA501			
	PV	MEAN	ARIMA	ARCH
5%	5.01E-08	7.50E-03	1.17E-08	6.40E-09
10%	1.11E-07	1.48E-02	3.27E-08	1.40E-08
20%	2.65E-07	2.93E-02	1.19E-07	3.34E-08
30%	5.76E-07	4.33E-02	3.26E-07	7.14E-08
40%	1.09E-06	5.69E-02	7.33E-07	1.34E-07
	MA1001			
	PV	MEAN	ARIMA	ARCH
5%	1.83E-08	0.0067645	3.69E-09	2.28E-09
10%	4.16E-08	0.0132053	1.19E-08	5.09E-09
20%	1.11E-07	0.0261488	5.27E-08	1.36E-08
30%	2.37E-07	0.0389997	1.38E-07	2.89E-08
40%	4.34E-07	0.0515033	2.87E-07	5.31E-08

Figure 3.14 clearly demonstrates that ARCH models are characterized by lower MSEs than ARIMA models. While the relative advantage of using ARCH models for imputation in the context of ground water data is relatively small for low shares of missing data, Figure 3.14 shows that for increasing share of missing data ARCH models outperform ARIMA models more significantly. This reflects the findings we presented in Section 3.4 in the content of simulated discharge data. The pattern of a bigger relative advantage of ARCH models can also be found in the smoothed time series (MA101, MA501 and MA1001). As before, higher shares of missing values are accompanied by a bigger relative advantage of ARCH models (see again Figure 3.14 and Table 3.4).

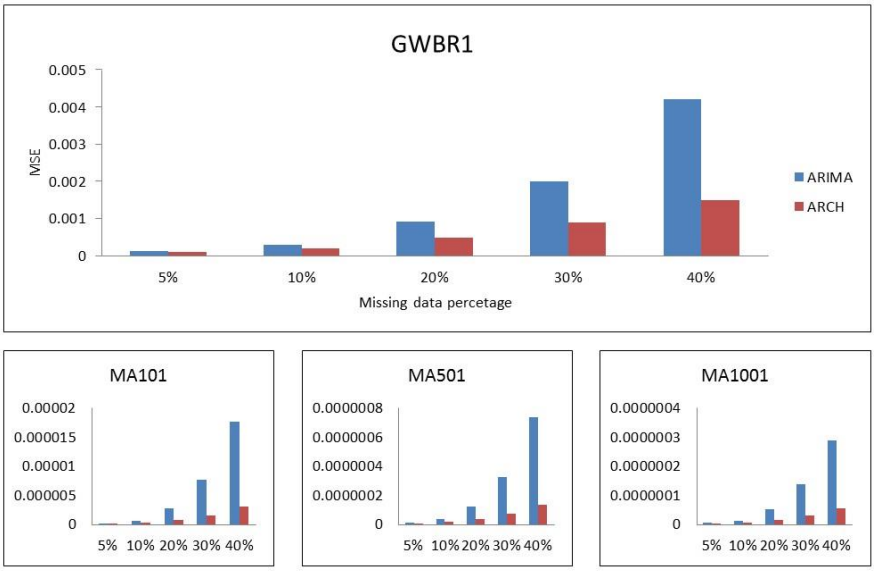

Figure 3. 14: Graphical results of Mean Squared Error of ARIMA/ARCH

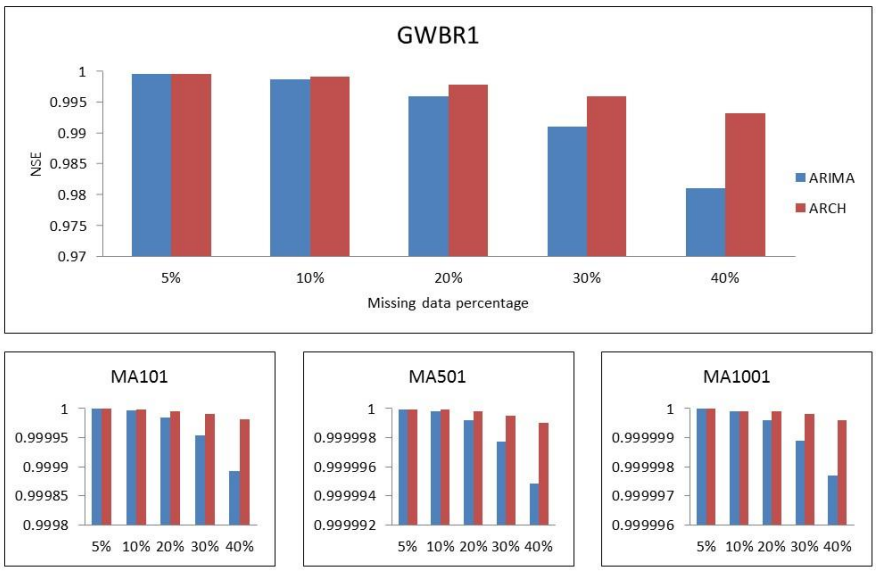

Figure 3. 15: Graphical results of Nash-Sutcliffe Efficiency of ARIMA/ARCH

Regarding the Nash Sutcliff Efficiency, we report similar results (see Table 3.5). For low shares of missing values the imputation performance of both ARCH and ARIMA is very similar. For increasing shares of missing data, however, ARCH models achieve significantly higher NSEs than comparable ARIMA models. Again, this pattern does not affect by the degree of smoothing applied to the time series as can easily be seen in Figure 3.15.

Table 3. 5: Nash-Sutcliffe Efficiency of imputation application for groundwater time-series

Missing data percentage	GWBR1			
	PV	MEAN	ARIMA	ARCH
5%	0.999	0.947	0.999	1.000
10%	0.997	0.893	0.999	0.999
20%	0.994	0.794	0.996	0.998
30%	0.987	0.693	0.991	0.996
40%	0.976	0.594	0.981	0.993
	MA101			
	PV	MEAN	ARIMA	ARCH
5%	1.000	0.946	1.000	1.000
10%	1.000	0.894	1.000	1.000
20%	1.000	0.793	1.000	1.000
30%	1.000	0.693	1.000	1.000
40%	1.000	0.595	1.000	1.000
	MA501			
	PV	MEAN	ARIMA	ARCH
5%	1.000	0.947	1.000	1.000
10%	1.000	0.896	1.000	1.000
20%	1.000	0.793	1.000	1.000
30%	1.000	0.694	1.000	1.000
40%	1.000	0.598	1.000	1.000
	MA1001			
	PV	MEAN	ARIMA	ARCH
5%	1.000	0.947	1.000	1.000
10%	1.000	0.897	1.000	1.000
20%	1.000	0.796	1.000	1.000
30%	1.000	0.695	1.000	1.000
40%	1.000	0.598	1.000	1.000

The performance of both ARIMA and ARCH models increases with higher levels of autocorrelation in the time-series data to be modeled. This is intuitive as an increase in autocorrelation makes the behavior of the time series more “predictable”: the value of *y* if period *t* has a stronger link to past values and can therefore be approximated with higher precision. In Table 3.6 and Figure 3.16 we report detailed findings comparing the

performance of not only ARIMA and ARCH models but also relatively simple methods in the case of 40% of the observations are missing for different levels of autocorrelation. Note, that the original time series GWBR1 is characterized by modest levels of autocorrelation while the smoothed time series MA101, MA501 and MA1001 are characterized by increasing levels of autocorrelation. It can be clearly seen, that the performance of these methods increases with increasing levels of autocorrelation and is highest for MA1001 – which is the time series with the highest levels of autocorrelation.

Table 3. 6: MSE and NSE of imputation application when data have 40% missing

MSE	40% Missing data percentage			
	MEAN	PV	ARIMA	ARCH
GWBR1	9.01E-02	5.39E-03	4.22E-03	1.50E-03
MA101	6.59E-02	2.51E-05	1.76E-05	3.00E-06
MA501	5.69E-02	1.09E-06	7.33E-07	1.34E-07
MA1001	5.15E-02	4.34E-07	2.87E-07	5.31E-08

NSE	40% Missing data percentage			
	MEAN	PV	ARIMA	ARCH
GWBR1	0.594	0.976	0.981	0.993
MA101	0.595	1.000	1.000	1.000
MA501	0.598	1.000	1.000	1.000
MA1001	0.598	1.000	1.000	1.000

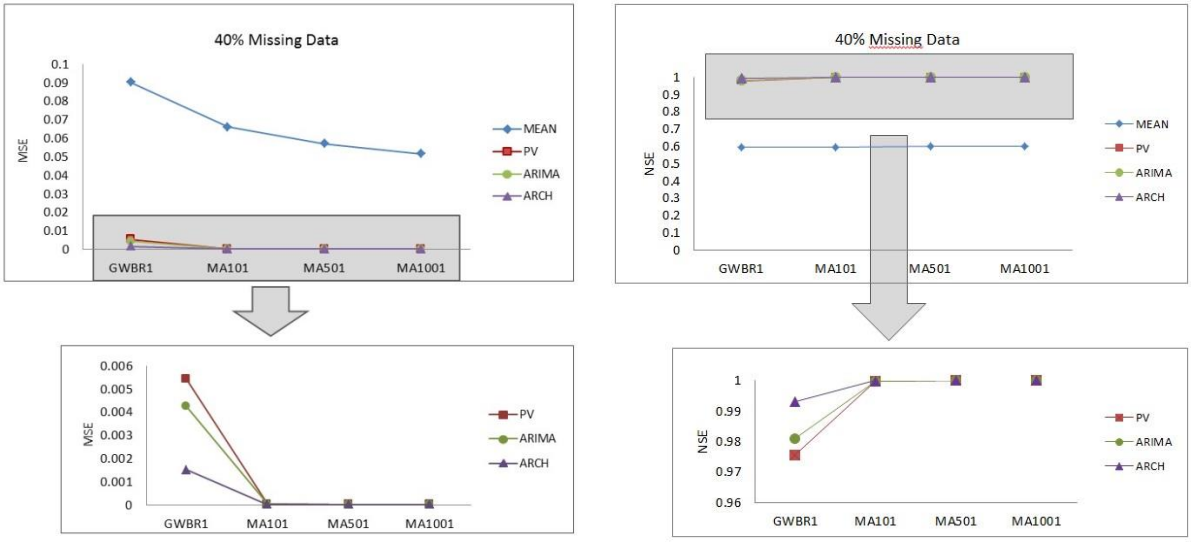

Figure 3. 16: MSE and NSE of imputation application when data have 40% missing

3.6 Conclusion

Complete time series data are a necessary precondition for most statistical approaches in hydrology, including determination of the flow duration curve, autocorrelation function, spectrum analysis, extreme value analysis based on the generalized extreme value distribution of annual blocks, principal component analysis, etc. In these cases, researchers need to resort to imputation methods in order to replace missing values with approximations as these statistical approaches require gap-free dataset. In this chapter, we evaluated the performance of five different imputation methods as follows. We created five time series of discharge data that exhibit different patterns of volatility using the HBV model. From these reference time series we randomly deleted a given share of observations to be imputed by the different approaches whose performance has been evaluated by the MSE and the NSE criteria. Our findings reveal that imputation methods that neglect the time series nature of the underlying reference data perform significantly worse than imputation methods that exploit this feature of the data. Moreover, advanced time series methods such as ARCH significantly outperform relatively simple time series method such as the preceding value imputation.

These findings are important for number of reasons: First, hydrological data is by its definition time series data that is typically characterized by typical feature such as autocorrelation and seasonality. In the presence of these features, the results obtained from commonly used imputation methods such as the wide-spread mean-value imputation can be improved significantly. As our study clearly reveals, even a relatively simple imputation algorithm that exploits the time series nature of the data – the preceding value approach – performs significantly better.

Second, we were also able to demonstrate that advanced regression-based time series imputation method such as ARIMA and ARCH models yield better results than the relatively simple preceding value imputation. Comparing the performance measures in Tables 3.2 to 3.5 shows that the ARIMA/ARCH models achieve significantly lower MSE values and significantly higher NSE values. While the latter is easy to implement and still performs much better than mean-value or OLS imputation techniques, imputation results can be optimized by relying on advanced econometric techniques. This is true in particular in situations where

a large fraction of observations is characterized by missing values. The larger the share of missing values the higher the performance advantage of advanced time series methods. The performance advantage of econometric time series methods is noteworthy as – as of now – their application in hydrological settings still is limited (see chapter 2).

Despite they overall encouraging findings there are, however, some caveats to be mentioned. On the conceptual level, our results have been obtained using data from only one catchment area (Brandenburg) and the results might differ for data obtained from other catchments. In order to ameliorate concerns regarding the broader applicability of our results, we varied the original data in order to obtain four additional time series that exhibit different volatility/seasonality characteristics. The results obtained are robust towards these variations. On the practical level, the implementation of the advanced econometric models (ARIMA and ARCH) requires statistical software packages such as R or STATA as these model typically are not implemented in standard hydrological software packages.

4 Multivariate time-series approaches for imputing hydrological data

4.1 Introduction

Due to the multiple reasons, hydrological data is often characterized by missing data (Elshorbagy et al., 2002; Johnston, 1999). A growing literature examines to what extent recent methodological advances in imputation approaches, i.e., techniques that approximate missing values using predictive models like Predictive Value Imputation (PVI) or unique-value imputation can be applied to hydrological data (Saar-Tsechansky & Provost, 2007). In this context, particular attention has been paid to the fact that hydrological data typically is time-series data: Hydrological variables such as precipitation or discharge are measured over time at fixed intervals constituting a time-series of sequential measurements (Berne, Delrieu, Creutin, & Obled, 2004). Often, these time-series of hydrological variables are characterized by internal time dependencies (such as autocorrelation, trend or seasonal variation) that can be exploited by appropriate models for the purpose of approximating missing values (Machiwal & Jha, 2012).

Statistical techniques that explicitly model such time dependencies (time-series models) therefore lend themselves for imputation in hydrology: Missing values in a given variable are approximated based on predictions derived from the underlying time-dependencies and the observed value(s) of a particular variable. In this context, chapter 2 applied advanced time-series models originally developed in financial econometrics to hydrological data with missing values. Autoregressive Integrated Moving Average (ARIMA) models and Autoregressive Conditional Heteroscedasticity (ARCH) models have been found to significantly outperform commonly used imputation techniques such as mean imputation or Ordinary Least Squares (OLS) based imputations which are widespread but ignore the time-series nature of the data (Hamilton, 1994).

Both ARIMA and ARCH exclusively exploit time-dependencies in the time-series of a given variable while ignoring other available information that might improve the quality of imputation (Degiannakis & Xekalaki, 2004; Stergiou et al., 1997). For instance, discharge in a given area will be correlated to (i) additional hydrological measurements taken in the same

area such as the amount of precipitation and also (ii) the discharge measures from neighboring areas. Including these additional information in a statistical model of the dependent variable might therefore also affect the quality of imputation performance.

In this chapter, we analyze to what extent extensions of the time-series models can be used that incorporate additional exogenous regressors. For this approach, we used some specific model types of the ARIMA and ARCH models - ARIMAX and ARCHX models - that exclusively exploit the time-series properties of the dependent variable. In particular, we compare imputations for various shares of missing values in a time-series of daily discharge derived from alternative time-series model: In a first step, we impute missing values from ARIMA and ARCH models that exclusively rely on the observed time-series of discharge. Second, we approximate missing values using extended ARIMAX and ARCHX models that include additional exogenous regressors such as precipitation, potential evapotranspiration or discharge measured from neighboring catchment areas. Finally, we compare the results from the different imputations in order to determine which approach yields the best results. The comparisons are based on the commonly used Mean Squared Error (MSE) and Nash Sutcliff Efficiency (NSE) criteria (Harville & Jeske, 1992; Nash & Sutcliffe, 1970).

We are using data from various small catchments in Brandenburg in the north of Germany over the period of ten years from November 1989 to October 1998. Our data from one gauge covers daily measurements of discharge, precipitation and (an estimation of) potential evapotranspiration. In addition, we got information on discharge measures for two neighboring gauges from the same catchment for the same time periods. The results from imputations based on differently specified time-series models indicate that both the performance derived from ARIMA and ARCH models can be significantly improved when exogenous regressors are included compared to the results that rely only on historical values of the specific dependent variable. Our results also indicate that including discharge measure from neighboring gauges as exogenous regressors has a better effect on the improvement rates than the inclusions of additional hydrological variables from the same gauge. These findings have important implications for practitioners that are confronted with missing data in hydrological time-series.

4.2 Study region and data

4.2.1 Overview

The study site is located in the federal state of Brandenburg located in Northeast Germany (Figure 4.1). Time series of potential evapotranspiration, observed precipitation and discharge from some of the gauges in research area were chosen. The whole area of Brandenburg is 29,479 km^2. Water is discharged either to the Northern Sea by the river Elbe or to the Baltic Sea by the river Oder. The surface river networks in Brandenburg have been anthropogenically altered for centuries (Merz & Pekdeger, 2011; Nützmann, Wolter, Venohr, & Pusch, 2011). In this region, sediments that mainly consist of glaciofluvial sands and tills are relatively young. Luvisols, albeluvisols and cambisols are the main soil types (Lieberoth, 1982). Forest area contributes to 35% of the whole area. Agricultural land is another main land use type with 34% cropland and 9% pasture (Uwe Schindler, Mueller, Eulenstein, & Dannowski, 2008). With an annual long-term precipitation sum of around between 600 to 650 mm and an annual long-term potential evapotranspiration sum of 510 mm between 1961 and 1991 (Hydrologischer Atlas von Deutschland, 2003), it is one of the areas in with the lowest climatic water balance in Germany. Due to high climatic water demand, the potential evapotranspiration only leaving 100 mm per year as runoff (Lischeid & Nathkin, 2011).

For the last centuries, the hydrological system in this area has been altered by humans' behavior. For more detailed description and overview on human impacts and hydrological changes within this landscape we refer to Merz and Pekdeger (2011) and Germer et al. (2011). Some of the artificial ditches and streams have existed for more than ten decades. More than 100 discharge gauges are maintained by the ministry of the Environment, Health and Consumer Protection of the Federal State of Brandenburg. Hydrological data such as precipitation, discharge or temperature is typically is collected over time at given intervals. In this chapter, we are using data from gauge Boblitz (reference Nr. 5838700) over a decade from November 1989 to October 1998. Our data from this gauge covers daily measurements of discharge (Q), precipitation (P). Potential evapotranspiration (E) which calculated by measured data is also used. In addition, we get information on discharge measures for two neighboring gauges – Schönfeld (reference Nr. 5838800) and Vetschau (reference Nr.

5836900) from the same catchment for the same time periods (the location of three main gauges shows in Figure 4.1).

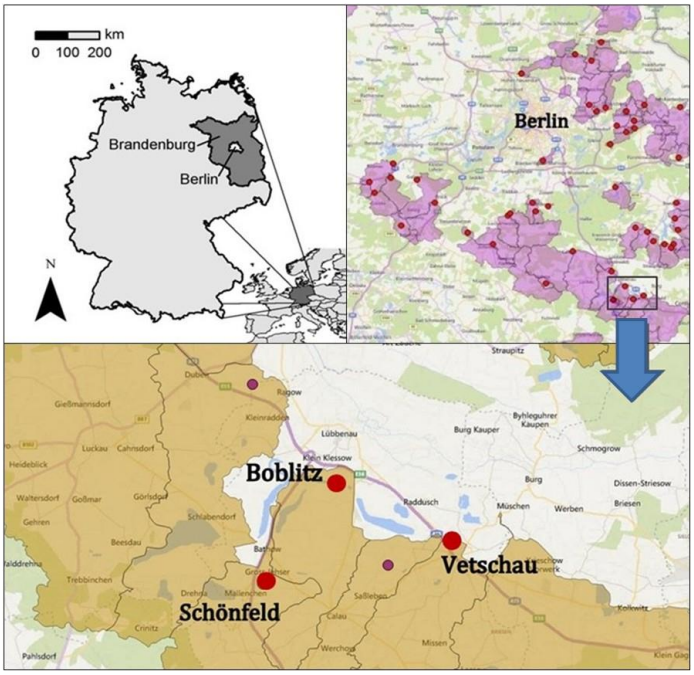

Figure 4. 1: The study area (Federal State of Brandenburg, Germany) and the location of three main gauges. Note: the violet colored areas in the upper right panel indicate basins of observation.

4.2.2 Data description

Before we model the time series of discharge from Boblitz (5838700) in Section 4.4, we briefly present key descriptive statistics on the dependent variable and additional variables that we will use as exogenous regressors in the subsequent analyses. In particular, these are information regarding precipitation and potential evapotranspiration from Boblitz (P_5838700 and E_5838700) as well as discharge information from the neighboring gauges Schönfeld (Q_5838800) and Vetschau (Q_5836900). Table 4.1 provides first summary statistics and pairwise correlation between the five variables in our sample. It is worth noting that discharge from Boblitz (Q_5838700) is highly correlated with discharge from Schönfeld (Q_5838800, coefficient of correlation 0.5789) but much less and negatively with discharge from Vetschau (Q_5836900, coefficient of correlation -0.0091). Potential evapotranspiration and precipitation from the same gauge also show only moderate correlations with the discharge measure. However, it needs to be taken into account that the coefficients of correlation do not consider any dynamic effects resulting from a delay in the effect one variable might have on another (Box et al., 2015). For instance, potential evapotranspiration

in a given period might explain discharge in a subsequent period (DeMeo, Laczniak, Boyd, Smith, & Nylund, 2003). This would explain why we don't observe higher correlations.

		Mean	S.D.	(1)	(2)	(3)	(4)	(5)
(1)	Discharge (5838700)	0.554	0.222	1				
(2)	Precipitation (5838700)	1.637	3.991	-0.0211	1			
(3)	Evaporation (5838700)	1.737	1.627	-0.0485	-0.0769	1		
(4)	Discharge (5838800)	0.404	0.331	0.5789	-0.0355	-0.1944	1	
(5)	Discharge (5836900)	0.495	0.137	-0.0091	0.0273	-0.3254	-0.1035	1

Table 4. 1: Summary statistics and pairwise correlations coefficients used in the study (N=3,287).
S.D. stands for standard deviation.

Figure 4.2 plots the various time series. Potential evapotranspiration and precipitation follow the typical stable seasonal patterns that can be expected for these variables. The discharge time series are more volatile which is common. It is noteworthy that the discharge time series of the gauges Boblitz (Q_5838700) and Schönfeld (Q_5838800) follow a similar drop in average discharge around 1992 whereas the discharge time series of the gauge Vetschau (Q_5836900) is not characterized by a similar drop. The time series of discharge from Boblitz (Q_5838700) is characterized by strong partial autocorrelation with its first lag (0.9864 mm/day) which quickly drops down to 0.0328 mm/day with its second lag. This lag structure suggest an ARIMA (1,1,1) model as an appropriate time series model.

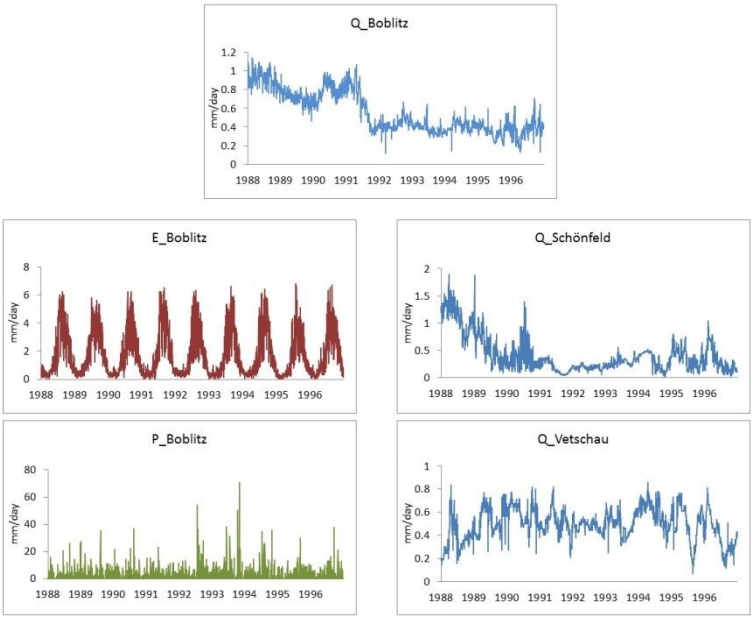

Figure 4. 2: Plot of the time series of the dependent variable (discharge in Boblitz) and additional variable which will be used as exogenous regressors.

4.3 Working steps

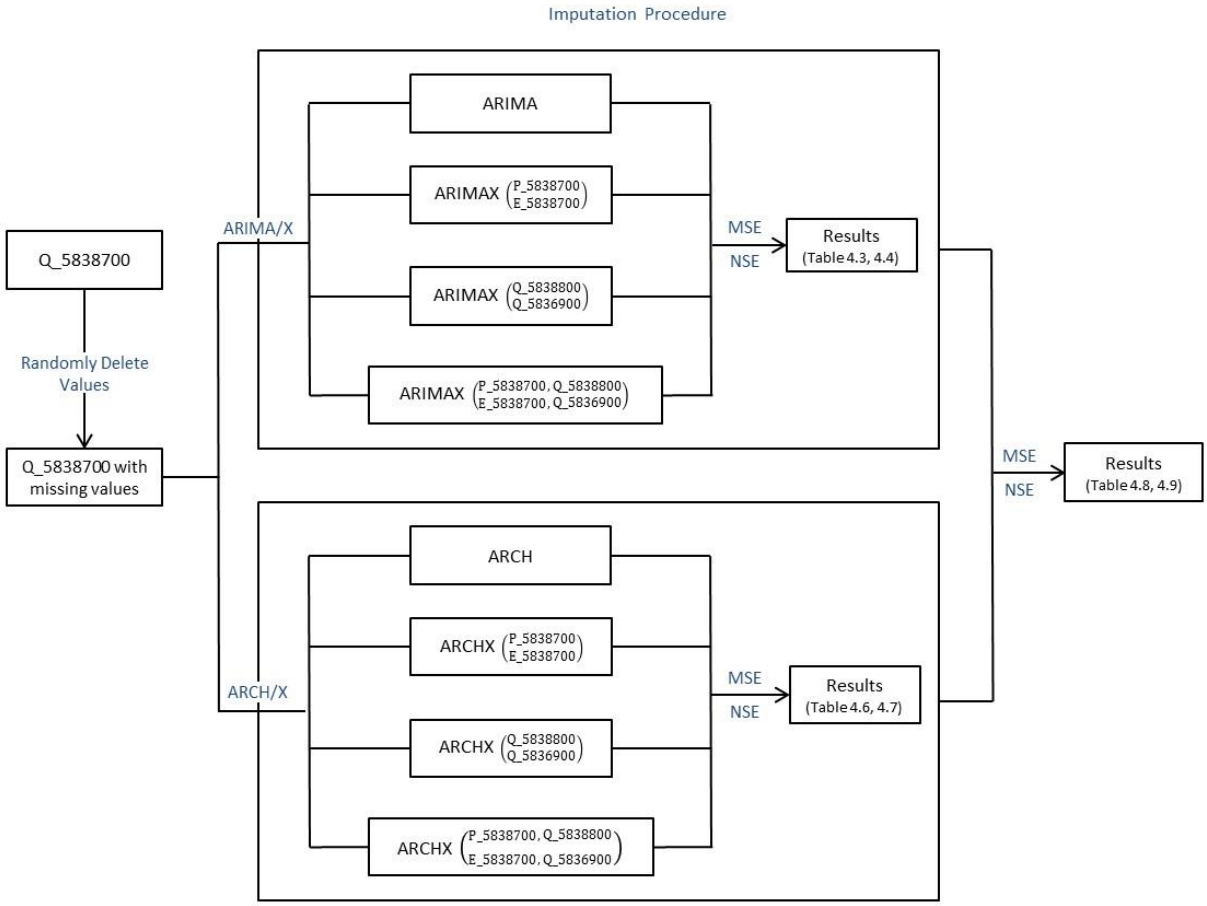

Figure 4. 3: Graphical process of the approach

In the context of imputation, it needs to be noted that univariate time series and respective models such are ARIMA and ARCH are special cases. Instead of covariates like in multivariate datasets, time dependencies (autocorrelations of the dependent variable) have to be exploited to perform an effective imputation since no additional information from exogenous variables is available. In chapter 2, we give an overview of alternative approaches to impute missing values in univariate time series and evaluate their performance in an empirical application. Our findings show that ARIMA and ARCH perform well compared to a number of alternative imputation approaches. In this section, we compare how the inclusion of exogenous variables can further increase the imputation performance of these time series models. In particular, we will compare the performance of univariate ARIMA/ARCH models with those of ARIMAX/ARCHX models that include additional information from the same gauge as the dependent variable (potential evapotranspiration and precipitation) or/and observations of discharge from neighboring gauges. As the values of these additional

variables are also observed for given time periods in which the dependent variable contains missing values, we expect their inclusion to considerably improve imputation performance. The steps of our research are summarized in Figure 4.2.

In order to evaluate imputation performance of these models we rely on the data described in Section 4.2. First, we randomly delete a given fraction of observations in the dependent variable from the reference data set – discharge from gauge Boblitz (Q_5838700). In particular, we delete 5%, 10%, 20%, 30% and 40% of the reference data sequentially. This will yield additional insights in how performance differs between different models depending on the share of data missing. Second, we fit alternative specifications of ARIMA/X and ARCH/X models to the resulting data sets, i.e., we estimate these models based only on the observations which are not missing.[4] In the final step, we use these fitted models to impute missing values of Q_5838700. Again, it needs to be highlighted that the imputations in the univariate models (ARIMA/ARCH) exploit only time dependencies of Q_5838700 while the multivariate models (ARIMAX/ARCHX) also exploit the value of the exogenous regressors in the same time periods where parameters in Q_5838700 were missing.

We measure the quality of the different imputations by comparing the imputed values of Q_5838700 with the values of the original reference time series. For this purpose, we compute commonly used measures of model fit: the Mean Squared Error (*MSE*) and the Nash Sutcliff Efficiency (*NSE*) criterion. The MSE measures the average of the squares of the difference between imputed and observed values (Schunn & Wallach, 2005) and can be compared across different models in order to assess which performs better. Formally, let Q_o^t be the observed discharge time-series (reference data Q_5838700) and Q_i^t be the time-series of discharge including imputed values from one of the imputation methods for the periods $t = 1, \cdots, T$. According to Harville and Jeske (1992), the *MSE* is then simply defined as

$$MSE = \frac{1}{T}\sum_{i=1}^{T}(Q_i^t - Q_o^t)^2. \quad (4.1)$$

It is a measure of the quality of the imputation and is always non-negative with values closer to zero being preferable. The *NSE* is a normalization of the *MSE* (Nash & Sutcliffe, 1970). It

[4] Note, that we nevertheless present results from fitting the models to the complete dataset below in order to report coefficient estimates and model fit.

can be obtained by dividing the *MSE* by the variance of the observations and subtracting that ratio from 1. Hence, the *NSE* is defined as

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_o^t - Q_i^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})^2} = 1 - \frac{MSE}{\sigma_{Q_o}^2}. \quad (4.2)$$

The range of *NSE* lies between 1.0 (perfect fit) and $-\infty$ and values closer to 1 are preferable.

4.4 Time-series based imputation approaches including exogenous regressors

4.4.1 Overview

Statistical models can be used to impute missing values in existing datasets (D. B. Rubin & Little, 2002). They formally summarize patterns in the data and express statistical relationships between observed values of the variable. Then the models are used to project the patterns in the data into the missing values (Schafer & Olsen, 1998). In other words, the statistical models approximate missing values based on observed values.

There are several types of statistical models in general that can be used for imputation purposes (Schafer & Olsen, 1998). Time series models account for the fact that data points taken over time may have an internal structure or time dependency (such as autocorrelation, trend or seasonal variation) that can be exploited in deriving a model of the underlying stochastic process. Generally speaking, time series models employ the statistical properties of the historical observations of the variable of interest in order to specify a formal model and estimate the unknown model parameters (Montgomery, Jennings, & Kulahci, 2015). Once the parameters of such models have been estimated based on observed values of a given variable, its missing values can be approximated. Cross-sectional regression models, on the other hand, make use of relationships between the variable of interest and one or more related predictor (or independent) variables to describe the forces that cause or drive the observed values of the variable of interest (Barros & Hirakata, 2003). Once these relations have been quantified – typically by using regression methods – missing values of the variable of interest can be approximated conditional on the independent variables.

In addition to time-series and cross-sectional models, there are statistical models that combine the properties of these two model categories. These models can be seen as

regression models with a serially dependent response variable, one or more independent variables and a stochastic error term. They are generally termed as Transfer Function Models (TFM) (Box & Jenkins, 1976) and there is broad range of different TFMs available. The identification and the estimation of many of these models, however, can be challenging. For a broader discussion on TFMs please refer to Box et al. (2015) and Montgomery et al. (2015) as a more comprehensive discussion would be beyond the scope of this chapter.

In this chapter, we restrict ourselves to a discussion of how statistical time-series models that have been previously used for the imputation of missing values in hydrological settings (see for instance chapter 2). The models can be extended to incorporate the effect of additional independent variables. In particular, we describe used extensions of the ARMA/ARIMA models and ARCH models – ARIMAX and ARCHX models - that not only capture dynamic behaviour of a given variable of interest but also allowed to model how this dynamic behaviour is affected by additional exogenous regressors (Note that these models can be considered as special cases of the more general transfer function models). After the introduction of these models we briefly discuss the availability of exogenous regressors that can potentially be included in models of a hydrological outcome variable of interest.

4.4.2 Models set-up

ARMA/X and ARIMA/X

In the following we denote observed values of a variable of interest *y* in a given period *t* with y_t. Often, observed values of y_t and y_{t-j} are correlated over time, i.e., autocorrelation between different measures of *y* exists. A commonly used specification of a random process that generates autocorrelation among different observations of *y* is the so-called autoregressive process of *pth* order AR (*p*) which is defined as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t, \quad (4.3)$$

In (4.3) epsilon is a random error term that follows a standard normal distribution and is independent over time with $E(\varepsilon_t \varepsilon_{t-i})=0$ *for all* $i \neq t$.

p here denotes the number of lagged values of y_t that enter the process. The random disturbance in the AR(p) model ε_t is an identically distributed (iid) error term with zero mean and constant variance. Alternatively, a stochastic process that generates autocorrelation in a

time-series can be created by specifying the contemporary value of y_t as a function of its mean μ and a sequence of past random shocks with

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots\ldots + \theta_p\varepsilon_{t-q}\,. \tag{4.4}$$

AR(p) and MA(q) processes are the building block of combined models which model y_t both as a function of prior values in the time series (AR terms) and the errors made in previous periods (MA terms). These combined models are known as ARMA(p,q) models (Box et al., 2015) and are typically written as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q}\,. \tag{4.5}$$

ARMA(*p,q*) models contain a *pth*-order autoregressive component in the observable time series, y_t, and a *qth*-order moving average component of the unobservable random shocks ε_t. A general assumption is that ε_t follows a random process with zero mean $E(\varepsilon_t)$ and constant variance $E(\varepsilon_t^2)=\sigma^2$. It is important to highlight that ARMA models can be fitted to data only if the underlying time-series y_t is weakly stationary. If y_t is not stationary, stationarity can often be achieved by differencing the time-series one or more times (Box et al., 2015). In this case, the ARMA (*p,q*) model (Autoregressive Moving Average) becomes an ARIMA (*p,d,q*) model (Autoregressive Integrated Moving Average) where *d* denotes the order of differencing, i.e., the number of time y_t is differenced to achieve stationarity.

It needs to be noticed, that ARMA and ARIMA models only exploit the information contained in the observed time-series. In many applications, however, the researcher seeks to add additional explanatory variables *x* that affect the outcome y_t in addition to its own history (Yang, Huang, & Huang, 1995). For instance, the current amount of discharge does not only depend on its own history but is also correlated to other variables such as temperature or the amount of discharge in adjacent geographical areas. In order to model the effect of so-called independent or exogenous variables, ARMA and ARIMA can be easily extended by including an independent variable x_t (or a vector thereof) on the right hand side of the equation with

$$y_t = \beta x_t\, \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q}\,. \tag{4.6}$$

where x_t is a covariate at time t and β is its coefficient. ARMA and ARIMA models that incorporate exogenous variables x_t are typically called ARMAX and ARIMAX models.

This straight-forward approach, however, has the disadvantage that the covariate coefficient β is hard to interpret. The value of β is not the marginal effect (the effect on y_t when x_t is increased by one unit) as it is in linear ordinary least squares (OLS) regressions. In fact, the presence of lagged values of the response variable y on the right hand side of the equation implies that β can only be interpreted conditional on the value of previous values of y, which does not have a meaningful interpretation. If these models are used for out-of-sample predictions or imputation purposes, however, a clear interpretation of the estimated coefficients is of less importance. Finally, Maximum Likelihood methods can be used to estimate the parameters of ARMAX/ARIMAX models and are implemented in available statistical software packages such as STATA.

ARCH/X

As we described above, both ARMA and ARIMA models are based on the relatively strict assumption that variance of the error terms is constant over time with $E(\varepsilon_t^2)=\sigma^2$ over time. In many applications including typical hydrological settings, however, this assumption might be too restrictive to represent the real data generating process. For instance, local climates might be characterized by a period of stable conditions followed by change in weather that drastically alters relevant outcomes (Hughes et al., 2011) which renders the assumption of constant autocorrelation too narrow. More realistic would be an assumption of changing variance and hence changing autocorrelation of the observed outcomes over time (heteroscedasticity).

Auto Regressive Conditional Heteroscedasticity (ARCH) models are generalizations of the ARMA/ARIMA models that in addition to past values of y_t also captures time-varying volatility. While ARCH models are holding the unconditional variance of ε_t constant with $E(\varepsilon_t^2) = \sigma^2$ they allow its conditional variance to follow an AR process of its own with

$$\varepsilon^2 = \zeta + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + \upsilon_t \,, \tag{4.7}$$

where υ_t is a new white noise process. Similar to ARMA and ARIMA models, the ARCH model can easily be extended to incorporate the effect of exogenous variables by including an

independent variable x_t (or a vector of independent variables) on the right hand side of the equation (Note that despite the inclusion of exogenous variables these models typically are called ARCH models. Taken together the ARCHX model is defined by the following two equations:

$$y_t = \beta x_t + \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots\ldots + \alpha_p y_{t-p} + \varepsilon_t, \quad (4.8)$$

and

$$\varepsilon^2 = \zeta + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + \upsilon_t. \quad (4.9)$$

The term ARCHX can be found but is not frequently used. Again, the coefficient β is not the marginal effect that a change in *x* has on *y*. Since our major purpose is using ARCH models for prediction purposes we are not interested, however, in an interpretation of the found coefficients. ARCH models can be fitted to observed data using simple OLS estimator which are implemented in standard statistical software packages. In our case, we fit an ARCH model that extends an ARIMA (*p,d,q*) model by a first-order autoregressive process for the variance of the error term ε_t^2 using STATA 13's arch command.

4.4.3 Exogenous regressors in hydrological settings

The application of time series models with exogenous regressors naturally requires the existence of at least one additional exogenous variable that (i) has been observed for the same time period and measured in the same intervals as the dependent variable and that (ii) is correlated to the dependent variable to be modeled. In hydrological settings, there are several natural candidates for variables that fulfill these requirements.

First, a number of variables are typically collected at the same time by hydrological observation stations in a given research area. Most frequently, information with regard to discharge (Q) and precipitation (P) are collected at the same time and are therefore available for equal time periods and identical measurement intervals. Moreover, potential evapotranspiration (E_{pot}) which can be calculated by measured data, such as, global radiation (R_{glob}), air humidity (*Rf*), day length (*DJ*), daily average air temperature (T_m) and daily average wind speed (*wi*), can be also considered as an exogenous regressor. There are many different ways to calculate E_{pot}, such as Thornthwaite equation, Penman–Monteith equation

(Edwards & Warwick, 1984; Thornthwaite, 1948). According to Wendling (Wendling, Schellin, & Thomä, 1991), the formula is:

$$E_{pot} = 2.4 \frac{T_m+22}{T_m+123}\left(\frac{R_{glob}}{410} + (0.5 + 0.54wi) + \frac{(100-Rf)DI}{905}\right) \quad (4.10)$$

Where, T_m is daily average air temperature in °C, R_{glob} is global radiation in Jcm^{-2}, *wi* is wind speed ms^{-1}, *Rf* is air humidity in %, *DJ* is day length in h.

Typically, the majority of collected hydrological variables can be expected to correlate one with each other. In fact, most hydrological models are based on functional relations between these variables (Devia, Ganasri, & Dwarakish, 2015). If a time series of one these variables have to be modeled, the other variables can then be used as additional exogenous regressors that might help to improve the statistical models. Second, it is reasonable to assume that hydrological variables are characterized by spatial correlations. In particular in neighboring research areas hydrological variables can be expected to be determined by the same regional weather conditions with variations limited by the local micro-climate. For instance, it cannot be expected that the amount of precipitation in neighboring regions is highly correlated. As discharge is a function of precipitation, also measures of discharge should be spatially correlated to a certain extent. For this reason, observations of the dependent variables in neighboring regions can be used as exogenous regressors in time-series models of hydrological variables.

In this chapter, we exemplarily model the time series of discharge from a given gauge in order to examine how additional exogenous variables can improve the model fit. More importantly, we will also examine whether models including exogenous regressors achieve better imputation performance than univariate time series models. In this application, we will extend the univariate model of discharge by (i) precipitation and potential evapotranspiration as additional measures obtained from the same gauge and (ii) discharge data from two neighboring gauges.

4.5 Evaluation of time-series based imputation approaches accounting for exogenous regressors

4.5.1 ARIMA and ARIMAX models

We present estimation results from fitting alternative ARIMA and ARIMAX models to the complete data set described in section 4.2 before we then discuss the performance of these models when used to impute missing data. Table 4.2 presents the estimation results from four different specifications. Since coefficients of ARIMA models don't have an intuitive interpretation we focus this brief discussion of the results to the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) which are also reported in Table 4.2. AIC and BIC are commonly used criteria for model selection as they allow comparing the quality of the different models regression models that have been estimated on the same data (Akaike, 2011; Bhat & Kumar, 2010). Note that lower values of AIC and BIC are preferable as they indicate better model fit.

Column (1) of Table 4.2 presents the results from fitting a parsimoniously specified ARIMA (1,1,1) model to the time series Q_583700.[5] Both the AR(1) as well as the MA(1) term are highly significant reflecting the serial correlation of the dependent variable. We gradually add exogenous regressors to the ARIMA model in Columns (2) to (4). First, we add precipitation (P) and potential evapotranspiration (E) from Boblitz (5838700) as regressors to the ARIMA model. Their coefficients are statistically highly significant and the inclusion of these variables lowers both AIC and BIC indicating an improvement in model fit. Next, we add the discharges (Q) observed in the neighboring gauges to the ARIMA model, see Column (3). Their coefficients are highly significant. Interestingly, the inclusion of Q from neighboring gauges improves AIC and BIC measure more than the inclusion of precipitation and potential evapotranspiration measures from the same gauge. We therefore conclude that information on Q from neighboring geographic areas has higher predictive power then precipitation and potential evapotranspiration from the same area. Finally, Column (4) presents the results from a comprehensive model that contains both information on precipitation and potential evapotranspiration from Boblitz as well as information on Q from neighboring gauges. In this

[5] In unreported robustness tests we estimated alternative specifications with more complex lag structures. The results, however, didn't improve significantly and hence we prefer this specification.

model, the coefficients of all exogenous regressors are highly significant. Moreover, AIC and BIC are minimized when compared to the respective values of the three preceding models.

		ARIMA Regressions			
		(1)	(2)	(3)	(4)
Variables	D.Q_5838800			0.05844**	0.05764**
				(0.00664)	(0.00670)
	D.Q_5836990			0.1112**	0.10931**
				(0.01201)	(0.01197)
	D.P_5838700		-0.00044**		-0.00043**
			(0.00010)		(0.00011)
	D.E_5838700		-0.00349**		-0.00311**
			(0.00089)		(0.00089)
	Constant	-0.00016	-0.00016	-0.00015	-0.00015
		(0.00015)	(0.00015)	(0.00015)	(0.00016)
ARMA	AR(1)	0.83054**	0.82999**	0.82247**	0.82166**
		(0.01047)	(0.01049)	(0.01084)	(0.01099)
	MA(1)	-0.96026**	-0.96024**	-1.04466**	-0.95633**
		(0.00704)	(0.00706)	(0.00793)	(0.00748)
Variance	Constant	0.0354**	0.03526**	0.03333**	0.03470**
		(0.00021)	(0.00021)	(0.00032)	(0.00021)
N		3,286	3,286	3,286	3,286
AIC		-12,573.05	-12,594.13	-12,676.70	-12,696.08
BIC		-12,548.66	-12,557.55	-12,640.11	-12,647.30

Table 4. 2: Results from different ARIMA/ARIMAX models of Q_583700. Note: Standard errors in parentheses. * denotes coefficients significantly different from 0 on the 5% level. ** denotes coefficients significantly different from zero on the 1% level.

As described in Section 4.3, we ran several simulations in order to evaluate the performance of these four ARIMA/ARIMAX models in an imputation context. Table 4.3 summarizes the results for this exercise. Note Table 4.3 reports the percentage changes in MSE resulting from the inclusion of different exogenous regressors relative to the basic ARIMA model of Column (1) in Table 4.2 for different shares of missing values. For instance, the entry in line 4 and Column (3) of Table 4.3 represents the relative change in MSE if information on Q from the two neighboring gauges are included into the basic ARIMA model and 30% of the data set has been imputed. Table 4.4 has to be read in a similar way, but reports the relative change in NSE rather than MSE.

Share of missing values	Relative change of MSE			
	(1)	(2)	(3)	(4)
5%	-	-0.98%	-12.71%	-13.45%
10%	-	0.67%	-8.80%	-8.51%
20%	-	-0.10%	-4.90%	-4.90%
30%	-	-0.34%	-5.25%	-5.52%
40%	-	-0.06%	-6.08%	-5.75%

Table 4. 3: Relative changes of MSE by inclusion of exogeonous regressors relative to a baseline ARIMA model of Q_583700 without exogenous regressors.

Table 4.3 indicates that the inclusion of exogenous regressors can improve the imputation performance of ARIMA models. First, the inclusion of hydrological information from the same gauge as the dependent variable such as potential evapotranspiration or precipitation does not lead to a large reduction of the MSE relative to a univariate ARIMA model. The relative changes of the MSE criterion all remain below 1% in Column (2) of Table 4.3. The inclusion of information on Q from neighboring gauges, however, improves the imputation performance dramatically. (Note that lower values of MSEs are preferable) Imputation results based on both models including Q_5836990 and Q_5838800 (Columns (3) and (4) of Table 4.3) are characterized by significantly lower MSE measures when compared to the results from a pure ARIMA model. In particular, for lower shares of missing data the inclusion of these exogenous regressors can reduce the MSE measures by 8% to 13% which is a significant improvement of imputation performance.

Share of missing values	Relative change of NSE			
	(1)	(2)	(3)	(4)
5%	-	0.09%	1.06%	1.11%
10%	-	-0.15%	1.88%	1.80%
20%	-	0.06%	3.02%	3.01%
30%	-	0.29%	4.81%	5.05%
40%	-	0.07%	8.82%	8.35%

Table 4. 4: Relative changes of NSE by inclusion of exogeonous regressors relative to a baseline ARIMA model of Q_583700 without exogenous regressors.

When applying the NSE measure to evaluate the performance of the different models in an imputation context, we find similar results (Note that higher NSEs are preferable.) The inclusion of information from the same gauge does not noticeable improve NSE values. However, and similar to the findings regarding MSEs, the inclusion of Q values from

neighboring gauges significantly improves the imputation performance of ARIMA based models. Interestingly, according to the NSE values the inclusion of exogenous regressors outperforms univariate models in particular for a higher share of missing values. This is not surprising, as univariate time-series models have to rely solely on time-dependencies in Q which becomes more challenging as the length of gaps to be filled increases in the share of missing values. ARIMAX models, on the other hand, can exploit the information of the exogenous regressors also in periods where Q is missing.

4.5.2 ARCH and ARCHX models

		ARCH Regressions			
		(1)	(2)	(3)	(4)
Variables	D.Q_5838800			0.06011**	0.05987**
				(0.00520)	(0.00519)
	D.Q_5836990			0.09191**	0.08703**
				(0.01213)	(0.01202)
	D.P_5838700		-0.00044**		-0.00043**
			(0.00008)		(0.00009)
	D.E_5838700		-0.0042**		-0.0040**
			(0.00069)		(0.00068)
	Constant	-0.00045**	-0.00046**	-0.00045**	-0.00047**
		(0.00014)	(0.00013)	(0.00013)	(0.00013)
ARIMA	AR(1)	0.79817**	0.79075**	0.78475**	0.77782**
		(0.01094)	(0.01118)	(0.01141)	(0.01160)
	MA(1)	-0.94777**	-0.94469**	-0.9432**	-0.93979**
		(0.00728)	(0.00750)	(0.00755)	(0.00775)
ARCH	AR(1)	0.39426**	0.41825**	0.39626**	0.42252**
		(0.02448)	(0.02552)	(0.02436)	(0.02565)
Variance	Constant	0.00082**	0.00079**	0.00079**	0.00076**
		(0.00001)	(0.00001)	(0.00001)	(0.00001)
N		3,286	3,286	3,286	3,286
AIC		-13,024.74	-13,067.53	-13,134.86	-13,174.95
BIC		-12,994.26	-13,024.85	-13,092.18	-13,120.07

Table 4. 5: Results from different ARCH/ARCHX models of Q_583700. Note: Standard errors in parentheses. * denotes coefficients significantly different from 0 on the 5% level. ** denotes coefficients significantly different from zero on the 1% level.

Similar to Section 5.1, we present the results from fitting ARCH and ARCHX models to the complete data set before discussing the performance of these models in the context of imputation. Accordingly, Table 4.5 contains the estimation results from different ARCH

models with differing sets of exogenous variables along with AIC and BIC measure for the respective models.

Table 4.6 reports the relative change in the observed MSE values for the different ARCHX models when compared to the univariate ARCH model without exogenous regressors (Column (1) of Table 4.6) for different shares of missing values. The results largely resemble the findings reported for the ARIMA models in Table 4.3. The inclusion of exogenous regressors generally reduces MSE values but most of the reduction is due to the inclusion of information on Q from neighboring gauges. Moreover, MSEs increases are more pronounced for lower shares of missing values when compared to situations with a high share of missing values.

Share of	Relative change of MSE			
missing values	(1)	(2)	(3)	(4)
5%	-	0.00%	-12.65%	-12.41%
10%	-	-0.48%	-9.63%	-8.77%
20%	-	0.16%	-5.16%	-4.90%
30%	-	-0.43%	-5.23%	-5.70%
40%	-	-0.27%	-5.75%	-6.12%

Table 4. 6: Relative changes of MSE by inclusion of exogeonous regressors relative to a baseline ARCH model of Q_583700 without exogenous regressors.

Finally, Table 4.7 reports the relative improvements of NSE measures when the different models are applied to imputation. Again, the results reflect the previous findings with regard to ARIMA models presented in Table 4.4: The inclusion of exogenous regressors improves the NSE measures in particular in settings where a high share of observations of the dependent variable is missing.

Share of	Relative change of NSE			
missing values	(1)	(2)	(3)	(4)
5%	-	0.01%	1.08%	1.06%
10%	-	0.11%	2.06%	1.88%
20%	-	-0.12%	3.21%	3.05%
30%	-	0.39%	4.78%	5.23%
40%	-	0.39%	8.33%	8.85%

Table 4. 7: Relative changes of NSE by inclusion of exogeonous regressors relative to a baseline ARCH model of Q_583700 without exogenous regressors.

4.5.3 Comparison of ARIMA/X and ARCH/X models

The results presented in Section 4.5.2 indicate that the predictive power of both ARIMA and ARCH models can be significantly improved by the inclusion additional exogenous regressors. Comparing AIC and BIC values for identically specified ARIMA/X and ARCH/X models in Tables 4.2 and 4.5 reveals that ARCH/X models are generally better fitting the data as their AIC and BIC values are lower for specifications containing the same set of variables. As this comparison is based on fitting the models to the full data set, it does not necessarily allow us to draw conclusions regarding their predictive power in the context of the imputation of missing values. For this purpose, we compare the MSE and NSE measures obtained from ARIMA/X with those from ARCH/X models in Tables 4.8 and 4.9.

Share of	MSE ARIMA/X vs. ARCH/X			
missing values	(1)	(2)	(3)	(4)
5%	-2.44%	-3.46%	-2.52%	-3.67%
10%	-0.29%	0.85%	0.63%	0.00%
20%	-1.32%	-1.59%	-1.04%	-1.32%
30%	0.07%	0.16%	0.05%	0.26%
40%	0.24%	0.45%	-0.11%	0.63%

Table 4. 8: Relative changes of MSE between different ARIMA/X and ARCH/X models (corresponding to Tables 4.2 and 4.5) and shares of missing values.

Table 4.8 reports the reports the relative difference as percentage value between ARIMA/X and ARCH/X models under similar conditions. For instance, the entry in Column (3) and line 4 (0.05%) indicates, that imputations based on the ARIMAX model of Column (3) in Table 4.2 lead to a MSE measure that is 0.05% higher than the MSE obtained from imputations based on the ARCHX model of Column (3) in Table 4.5 when 30% of the data is missing. It becomes apparent, that the difference in the imputation performance between ARIMA/X and ARCH/X models is negligible when applying the MSE criterion. The maximum difference is a 3.67% lower MSE achieved by ARIMAX model (4) compared to ARCHX model (4) when only 5% of the data is missing. For higher shares of missing values the differences become even smaller.

Measuring predictive power using the NSE criterion yields a similar picture. The difference between imputations obtained from ARIMA/X and from ARCH/X models under comparable conditions is negligible. The differences of the observed NSE values are within -1% and 1% and do not suggest the superiority of one method over the other.

Share of missing values	NSE ARIMA/X vs. ARCH/X			
	(1)	(2)	(3)	(4)
5%	0.20%	0.28%	0.18%	0.25%
10%	0.07%	-0.19%	-0.11%	-0.01%
20%	0.80%	0.99%	0.61%	0.77%
30%	-0.05%	-0.15%	-0.03%	-0.23%
40%	-0.35%	-0.66%	0.13%	-0.85%

Table 4. 9: Relative changes of NSE by inclusion of exogeonous regressors relative to a baseline ARCH model of Q_583700 without exogenous regressors. Note: Percentage values in Table reflect percentage change of observed NSE.

4.6 Conclusion

Missing data is common phenomenon in hydrological data (Elshorbagy et al., 2002). As a consequence, imputation methods receive an increasing attention. Due to the time-series nature of hydrological data, alternative time-series models have been applied to the imputation of missing values (Fung, 2006). Commonly used time-series methods typically are purely based on time-dependencies, such as autocorrelations, of a single (dependent) variable and do not further information that might be available. In this chapter, we examined how the inclusion of information beyond the time-series of the variable of interest itself can improve imputation results. We have done this for a particular class of time-series models that has been shown to perform well in hydrological settings before: ARIMA and ARCH models. Extensions of these models to incorporate additional exogenous regressors are readily available with ARIMAX and ARCHX models and can easily be implemented even for large datasets as they are included in advanced statistical software packages such as STATA or R.

Our simulation-based evaluation of the ARIMAX and ARCHX models is based on discharge data that spans a period of 10 years. Our findings clearly indicate that an inclusion of additional exogenous regressors such as precipitation, potential evapotranspiration and discharge measures from neighboring research areas considerable –ARIMAX and ARCHX - improves the quality of imputation when compared to simple ARIMA and ARCH models. It is noteworthy that, the inclusion of measurements of the dependent variable (discharge) from neighboring catchments has a bigger effect on imputation quality when compared to (i) the inclusion of additional (other) regressors taken from the same catchment area and (ii) to

simple ARIMA and ARCH models. This finding implies that spatial correlations between discharge measures from adjacent catchments are more valuable in imputing missing data than additional information from the area under investigation. When comparing the performance of ARIMA/X to this of ARCH/X models in these settings, differences between these two model types, on the other hand, are negligible. This suggests a minor importance of heteroscedasticity in the error terms for imputation purposes in time-series models which include additional regressors.

These results bear important implications for both scholars as well as practitioners alike that seek to approximate missing values by statistical imputation methods. While statistical time-series models have been shown to perform comparably well in existing studies we have shown that their performance can be further enhanced by the inclusion of exogenous regressors. In particular including measures of the dependent variable from other catchments improved the results significantly. As the additional variables we used are typically readily available this approach can be implemented easily in many circumstances. Moreover, as our results indicate that the choice between ARIMA/X and ARCH/X is less important than the choice of additional regressors. Our study has limitations, however. In particular, the results derived here have been derived using data from a single research area. More comprehensive validation of our results using data from different settings therefore seems to be warranted. In particular, the observed pattern of improvements in imputation precision depending on the type of additional regressors warrants further research: Knowing which additional variables can boost imputation precision in a more general context will prove invaluable as a rule of thumb for practical work. The presented results can only be a first step towards such a rule of thumb.

5 Synthesis

Complete time series data are a necessary precondition for most statistical approaches in hydrology, including the determination of the flow duration curve, autocorrelation function, spectrum analysis, extreme value analysis based on the generalized extreme value distribution of annual blocks, principal component analysis, etc (McKnight et al., 2007). However, missing data is a common problem in hydrological data (precipitation, discharge, head fluctuation, etc.) and poses a serious problem for many statistical approaches require complete data sources in hydrology such as missing data is often harmful beyond reducing statistical power (Elshorbagy et al., 2002; McKnight et al., 2007). For reasons of convenience, researchers often resort to simple solutions to deal with missing data such as simply discarding observations characterized by missing data or by replacing missing data with a 'naïve' guess (such as the mean of all other observations). Despite their convenience, these solutions have severe statistical shortcomings.

In chapter 2, various imputation methods available to the hydrological researchers have been reviewed, including arithmetic mean imputation, Principal Component Analysis (PCA), regression-based methods and multiple imputation methods. Principal component analysis (PCA) - based as well as regression-based imputation methods can improve the accuracy of missing value imputation and reduce statistical problems induced by naïve imputation approaches (I. Jolliffe, 2002). Nevertheless, the discussion argues that these methods neglect the time-series nature of hydrological data that often requires more flexible non-linear models. We therefore put an emphasis on time-series regressions approaches that exploit the time series nature of hydrological data. Auto Regressive Conditional Heteroscedasticity (ARCH) models which originate from finance and econometrics and Autoregressive Integrated Moving Average (ARIMA) models are discussed regarding the applicability to hydrological contexts here. We focused our attention on discussing econometric time-series methods as they explicitly model the particular statistical properties of hydrological time-series (autocorrelation and heteroscedasticity) which are mostly neglected in algorithmic machine learning approaches.

It needs to be stressed that there have been few studies concerning imputation of missing data in time series context in hydrology in general. Despite its focus on particular focus on

selected methods, our survey clearly shows that there are methodological advances driven by other fields of research that bear relevance for hydrology as well. According to our knowledge, the hydrological community paid little attention to the imputation ability of neither time-series models in general and ARCH models in particular nor other advanced imputation approaches developed for the analysis of economic and financial time series data.

In chapter 3, the performances of imputation techniques which are widespread and easy to use but ignore the time series nature of hydrological data and imputation techniques exploiting their time series nature are compared. We created five time series of discharge data that exhibit different patterns of volatility by using the Hydrologiska Byråns Vattenbalansavdelning (HBV) model. From these reference time series we randomly deleted a given share of observations to be imputed by the different approaches whose performance has been evaluated by Mean Squared Error (MSE) and Nash Sutcliff efficiency (NSE) as performance measures. We find that econometric time series models such as Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Conditional Heteroscedasticity (ARCH) model outperform alternative imputation approaches such as mean imputation or Ordinary Least Squares (OLS) based regression methods. These findings hold across different scenarios in which we vary both the share of missing values in the data as well as crucial characteristics of the time series such as seasonality or volatility. Our findings in this chapter reveal that imputation methods that neglect the time series nature of the underlying reference data perform significantly worse than imputation methods that exploit this feature of the data. Moreover, advanced time series methods such as ARCH significantly outperform relatively simple time series method such as the preceding value imputation.

These findings have important implications for scholars and practitioners who work with datasets characterized by missing values for number of reasons: First, hydrological data is by its definition time series data that is characterized by typical feature such as autocorrelation and seasonality . In the presence of these features, the results obtained from commonly used imputation methods such as the wide-spread mean-value imputation can be improved significantly. As our study clearly reveals, even a relatively simple imputation algorithm that exploits the time series nature of the data – the preceding value approach – performs significantly better.

Second, we were also able to demonstrate that advanced regression-based time series imputation method such as ARIMA and ARCH models yield better results than the relatively simple preceding value imputation. While the latter is easy to implement and still performs much better than mean-value or OLS imputation techniques, imputation results can be optimized by relying on advanced econometric techniques. This is true in particular in situations where a large fraction of observations is characterized by missing values. The larger the share of missing values the higher the performance advantage of advanced time series methods. The performance advantage of econometric time series methods is noteworthy as – as of now – their application in hydrological settings still is limited.

According to chapter 3, univariate ARIMA and ARCH models have successfully been used for the imputation of missing value in hydrological data. Due to the time-series nature of hydrological data, alternative time-series models have been applied to the imputation of missing values (Fung, 2006). Commonly used time-series methods typically are purely based on time-dependencies, such as autocorrelations, of a single (dependent) variable and do not further information that might be available. So in chapter 4, we examined how the inclusion of information beyond the time-series of the variable of interest itself can improve imputation results. Extensions of these models to incorporate additional exogenous regressors are readily available with ARIMAX and ARCHX models. Using discharge data from Brandenburg in the northeast of Germany, we compare the imputation performance of univariate ARIMA and ARCH models which have been shown well in hydrological settings before with the performance of extended models – ARIMAX and ARCHX models.

First, we impute missing values from ARIMA and ARCH models that exclusively rely on the observed time-series of discharge from Brandenburg that spans a period of 10 years. Second, we approximate missing values using extended ARIMAX and ARCHX models that include additional exogenous regressors such as precipitation, potential evapotranspiration or discharge measured from neighboring catchment areas. Finally, we compare the results from the different imputations in order to determine which approach yields the best results. Relying on Mean Squared Error (MSE) and Nash Sutcliff Efficiency (NSE) as performance measures, our findings clearly indicate that an inclusion of additional exogenous regressors – ARIMAX and ARCHX models- improves the quality of imputation when compared to simple ARIMA and ARCH models. In particular, the inclusion of discharge measures of neighboring

catchments has a bigger effect on imputation quality when compared to simple ARIMA and ARCH models. When comparing the performance of ARIMA/X to this of ARCH/X models in these settings, differences are negligible.

These results from chapter 4 bear important implications for both scholars as well as practitioners alike that seek to approximate missing values by statistical imputation methods. While statistical time-series models have been shown to perform comparably well in existing studies it was shown that their performance can be further enhanced by the inclusion of exogenous regressors. The additional variables we used are typically readily available and therefore this approach can be implemented easily. Moreover, as our results indicate, the choice between ARIMA/X and ARCH/X is less important than the choice of additional regressors.

Despite they overall encouraging findings there are, however, some caveats to be mentioned. On the conceptual level, our results have been obtained using data from only one catchment area (Brandenburg) and the results might differ for data obtained from other catchments. More comprehensive validation of our results using data from different settings therefore seems to be warranted.

We do not address the question of performance advantages of either of these advanced methods in an applied setting. However, we hope that our survey stimulates additional research into these methods and their applicability in hydrology. Whether and to what extent the advanced imputation methods presented here lead to more precise hydrological analyses itself (rather than the improvement of imputation quality) in the presence of incomplete datasets ultimately remains an empirical question. Future research can easily address this question in the context of simulation-based comparisons of different imputation methods within a well-defined hydrological application.

Moreover, in this dissertation we focused on imputation methods that have a strong statistical foundation. Other fields of research such as computer and data sciences provide alternative methods with different methodological underpinnings that can be applied to data imputation. Most prominently, neural networks (or more precisely artificial neural networks) have been motivated by the recognition that the human brain processes information in a way that is fundamentally different from the typical digital computer is

frequently applied. As of today, there is little evidence, to what extent artificial neural networks or similar methods in their spirit can be applied to the imputation of missing data in hydrological settings. Our study is mute on this topic, and we defer the task of evaluating the performance of artificial neural networks in hydrological settings to future research.

Reference

Adhikari, R., & Agrawal, R. (2013). An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*.

Akaike, H. (2011). Akaike's Information Criterion *International Encyclopedia of Statistical Science* (pp. 25-25): Springer.

Allison, P. D. (1999). Multiple imputation for missing data: A cautionary tale: Philadelphia.

Allison, P. D. (2001). *Missing data* (Vol. 136): Sage publications.

Allison, P. D. (2012). *Handling missing data by maximum likelihood.* Paper presented at the SAS global forum.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37.

Barros, A. J., & Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC medical research methodology, 3*(1), 1.

Baur, D. G., & Lucey, B. M. (2009). Flights and contagion—An empirical analysis of stock–bond correlations. *Journal of Financial stability, 5*(4), 339-352.

Bergström, S. (1992). *The HBV model: Its structure and applications*: Swedish Meteorological and Hydrological Institute.

Bergström, S., & Forsman, A. (1973). Development of a conceptual deterministic rainfall-runoff model. *Hydrology Research, 4*(3), 147-170.

Bergström, S., & Singh, V. (1995). The HBV model. *Computer models of watershed hydrology.*, 443-476.

Berne, A., Delrieu, G., Creutin, J.-D., & Obled, C. (2004). Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology, 299*(3), 166-179.

Bhat, H., & Kumar, N. (2010). On the derivation of the Bayesian Information Criterion. *School of Natural Sciences, University of California*.

Box, G. E., & Jenkins, G. M. (1976). Time series analysis, control, and forecasting. *San Francisco, CA: Holden Day, 3226*(3228), 10.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*: John Wiley & Sons.

Bradbury, K., Baker, V., Barros, A., Campana, M., Gray, K., Haan, C., . . . Shabman, L. (1999). Hydrologic Hazards Science at the US, Geological Survey Commission on Geosciences, Environment and Resources: National Research Council: National Academy Press.

Cool, A. L. (2000). A Review of Methods for Dealing with Missing Data.

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research. *New directions for institutional research, 2005*(127), 33-49.

de Leeuw, J. (1986). *Multidimensional Data Analysis: Proceedings of a Workshop, Pembroke College, Cambridge University, England, June 30-July 2, 1985* (Vol. 7): DSWO Press.

Degiannakis, S., & Xekalaki, E. (2004). Autoregressive conditional heteroscedasticity (ARCH) models: A review. *Quality Technology & Quantitative Management, 1*(2), 271-324.

DeMeo, G. A., Laczniak, R. J., Boyd, R. A., Smith, J., & Nylund, W. E. (2003). Estimated ground-water discharge by evapotranspiration from Death Valley, California, 1997–2001. *US Geological Survey Water-Resources Investigations Report*, 03-4254.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Devia, G. K., Ganasri, B., & Dwarakish, G. (2015). A review on hydrological models. *Aquatic Procedia, 4*, 1001-1007.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology, 59*(10), 1087-1091.

Edwards, W., & Warwick, N. (1984). Transpiration from a kiwifruit vine as estimated by the heat pulse technique and the Penman-Monteith equation. *New Zealand Journal of Agricultural Research, 27*(4), 537-543.

Elshorbagy, A., Simonovic, S., & Panu, U. (2002). Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology, 255*(1), 123-133.

Enders, C. K. (2010). *Applied missing data analysis*: Guilford Press.

Eom, K. S., Hahn, S. B., & Joo, S. (2004). Partial price adjustment and autocorrelation in foreign exchange markets. *Preprint, University of California at Berkeley*.

Fama, E. F., & French, K. R. (1988). Permanent and temporary components of stock prices. *The Journal of Political Economy*, 246-273.

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition, 41*(12), 3692-3705.

Feinberg, E. A., & Genethliou, D. (2005). Load forecasting *Applied mathematics for restructured electric power systems* (pp. 269-285): Springer.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*: Cambridge University Press.

Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika, 41*(3), 409-415.

Fung, D. S. C. (2006). *Methods for the estimation of missing values in time series.* Edith Cowan University Perth.

Germer, S., Kaiser, K., Bens, O., & Hüttl, R. F. (2011). Water balance changes and responses of ecosystems and society in the Berlin-Brandenburg region–a review. *DIE ERDE–Journal of the Geographical Society of Berlin, 142*(1-2), 65-95.

Gill, M. K., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Water resources research, 43*(7).

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology, 60*, 549-576.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology, 142*(12), 1255-1264.

Guzman, J. A., Moriasi, D., Chu, M., Starks, P., Steiner, J., & Gowda, P. (2013). A tool for mapping and spatio-temporal analysis of hydrological data. *Environmental modelling & software, 48*, 163-170.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2): Princeton university press Princeton.

Harrington, D. (2008). *Confirmatory factor analysis*: Oxford University Press, USA.

Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association, 87*(419), 724-731.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison.

Hawkins, M., & Merriam, V. (1991). An overmodeled world. *Direct Marketing*, 21-24.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods, 2*(1), 64.

Hughes, C. E., Cendón, D. I., Johansen, M. P., & Meredith, K. T. (2011). Climate change and groundwater *Sustaining Groundwater Resources* (pp. 97-117): Springer.

Hydrologischer Atlas von Deutschland (2003). Bundesministerium für Umwelt. *Naturschutz und Reaktorsicherheit*.

Johnston, C. A. (1999). *Development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data.* Virginia Tech.

Jolliffe, I. (2002). *Principal component analysis*: Wiley Online Library.

Jolliffe, I. T. (1993). Principal component analysis: a beginner's guide—II. Pitfalls, myths and extensions. *Weather, 48*(8), 246-253.

Kiers, H. A. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika, 62*(2), 251-266.

Killingtveit, Å., & Sand, K. (1990). *On areal distribution of snowcover in a mountainous area.* Paper presented at the Proceedings of Northern Hydrology Symposium.

Kim, J.-O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research, 6*(2), 215-240.

King, G., Honaker, J., Joseph, A., & Scheve, K. (1998). *List-wise deletion is evil: what to do about missing data in political science.* Paper presented at the Annual Meeting of the American Political Science Association, Boston.

Kondrashov, D., & Ghil, M. (2006). Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics, 13*(2), 151-159.

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American journal of epidemiology, 171*(5), 624-632.

Lieberoth, I. (1982). *Bodenkunde: Aufbau, Entstehung, Kennzeichnung und Eigenschaften der landwirtschaftlich genutzten Böden der DDR*: VEB Deutscher Landwirtschaftsverlag.

Lindström, G., & Bergström, S. (1992). Improving the HBV and PULSE-models by use of temperature anomalies. *Vannet i Norden, 1*, 16-23.

Lischeid, G., & Nathkin, M. (2011). The Potential of Land-Use Change to Mitigate Water Scarcity in Northeast Germany–a Review. *DIE ERDE–Journal of the Geographical Society of Berlin, 142*(1-2), 97-113.

Little, R., & Rubin, D. (1987). Analysis with missing data: John Wiley & Sons, New York.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287-296.

Machiwal, D., & Jha, M. K. (2012). Methods for time series analysis *Hydrologic Time Series Analysis: Theory and Practice* (pp. 51-84): Springer.

Malhotra, N. K. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research*, 74-84.

Marques, C., Ferreira, J., Rocha, A., Castanheira, J., Melo-Goncalves, P., Vaz, N., & Dias, J. (2006). Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth, Parts A/B/C, 31*(18), 1172-1179.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(1), 22-36.

Mcdonald, R. A., Thurston, P. W., & Nelson, M. R. (2000). A Monte Carlo study of missing item methods. *Organizational Research Methods, 3*(1), 71-92.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*: Guilford Press.

Merz, C., & Pekdeger, A. (2011). Anthropogenic changes in the landscape hydrology of the Berlin-Brandenburg region. *DIE ERDE–Journal of the Geographical Society of Berlin, 142*(1-2), 21-39.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*: John Wiley & Sons.

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.

Natkhin, M., Steidl, J., Dietrich, O., Dannowski, R., & Lischeid, G. (2012). Differentiating between climate effects and forest growth dynamics effects on decreasing groundwater recharge in a lowland region in Northeast Germany. *Journal of Hydrology, 448*, 245-254.

Nützmann, G., Wolter, C., Venohr, M., & Pusch, M. (2011). Historical patterns of anthropogenic impacts on freshwaters in the Berlin-Brandenburg region. *DIE ERDE–Journal of the Geographical Society of Berlin, 142*(1-2), 41-64.

Pandey, P. K., Singh, Y., & Tripathi, S. (2011). Image processing using principle component analysis. *International Journal of Computer Applications (0975–8887) Volume*.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research, 74*(4), 525-556.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation, 7*(4), 353-383.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What to Do when Data Are Missing in Group Randomized Controlled Trials. NCEE 2009-0049. *National Center for Education Evaluation and Regional Assistance*.

Raaijmakers, Q. A. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59*(5), 725-748.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annu. Rev. Public Health, 25*, 99-117.

Renner, C. B., & Braun, L. (1990). *Die Anwendung des Niederschlag-Abfluss Modells HBV3-ETH (V 3.0) auf verschiedene Einzugsgebiete in der Schweiz*: Geographisches Institut ETH Zürich.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel psychology, 47*(3), 537-560.

Roth, P. L., Switzer, F. S., & Switzer, D. M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods, 2*(3), 211-232.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81): John Wiley & Sons.

Rubin, D. B., & Little, R. J. (2002). Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*.

Rubin, L. H., Witkiewitz, K., St Andre, J., & Reilly, S. (2007). Methods for Handling Missing Data in the Behavioral Neurosciences: Don t Throw the Baby Rat out with the Bath Water. *Journal of Undergraduate Neuroscience Education, 5*(2), 71-77.

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research, 8*(Jul), 1623-1657.

Saunders, J. A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E. K., & Pescarino, R. (2006). Imputing missing data: A comparison of methods for social work researchers. *Social work research, 30*(1), 19-31.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2), 147.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research, 33*(4), 545-571.

Schindler, U., Mueller, L., Eulenstein, F., & Dannowski, R. (2008). A long-term hydrological soil study on the effects of soil and land use on deep seepage dynamics in northeast Germany. *Archives of Agronomy and Soil Science, 54*(5), 451-463.

Schindler, U., & Müller, L. (2010). Data of hydraulic properties of North East and North Central German soils. *Earth System Science Data, 2*(2), 189-194.

Schunn, C. D., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. *Psychologie der Kognition: Reden and vorträge anlässlich der emeritierung von Werner Tack*, 115-154.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 367-377.

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis: online] Technical Report.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., & Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing, 70*(16), 2861-2869.

Stergiou, K., Christou, E., & Petrakis, G. (1997). Modelling and forecasting monthly fisheries catches: comparison of regression, univariate and multivariate time series methods. *Fisheries Research, 29*(1), 55-95.

Stock, J. H., Watson, M. W., & Addison-Wesley, P. (2007). Introduction to econometrics.

Tannenbaum, C. E. (2009). The empirical nature and statistical treatment of missing Data.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical review, 38*(1), 55-94.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management, 24*(1), 53-62.

van der Heijden, G. J., Donders, A. R. T., Stijnen, T., & Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in

multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology, 59*(10), 1102-1109.

Vehvilainen, B. (1986). Modelling and forecasting snowmelt floods for operational forecasting in Finland. *Modelling Snowmelt Induced Processes*, 245-256.

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis *A practical approach to microarray data analysis* (pp. 91-109): Springer.

Wendling, U., Schellin, H.-G., & Thomä, M. (1991). Bereitstellung von täglichen Informationen zum Wasserhaushalt des Bodens für die Zwecke der agrarmeteorologischen Beratung. *Zeitschrift für Meteorologie, 41*(6), 468-475.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.). Modeling longitudinal and multiple group data: Practical issues: Applied approaches and specific examples (pp. 219-240). Mahwah, NJ: Lawrence Erlbaum.

Yang, H.-T., Huang, C.-M., & Huang, C.-L. (1995). *Identification of ARMAX model for short term load forecasting: an evolutionary programming approach.* Paper presented at the Power Industry Computer Application Conference, 1995. Conference Proceedings., 1995 IEEE.

Yanik, B., & Avci, I. (2004). Determination of regional flow duration curves. *Itue Dergisi/d, 4*(5), 19-30.

Zhang, Q., Wang, B.-D., He, B., Peng, Y., & Ren, M.-L. (2011). Singular spectrum analysis and ARIMA hybrid model for annual runoff forecasting. *Water resources management, 25*(11), 2683-2703.

Zhu, F., & Wang, D. (2008). Local estimation in AR models with nonparametric ARCH errors. *Communications in Statistics—Theory and Methods, 37*(10), 1591-1609.

Appendix I - List of publications

Y. Gao, C. Merz, G. Lischeid and M. Schneider (2016). Dealing with missing data in hydrological data. Environmental Earth Sciences, under review.

Y. Gao, C. Merz, G. Lischeid and M. Wegehenkel (2017). Alternative imputation approaches and their performance differences, submitted soon.

Appendix II – Curriculum Vitae

For reasons of data protection, the Curriculum vitae is not published in the online version.