Performance of GRADE in simulating flood wave characteristics in the Rhine basin

Ing. H. Trul
May 2016, Enschede

UNIVERSITEIT TWENTE.

Deltares
Enabling Delta Life

Performance of GRADE in simulating flood wave characteristics in the Rhine basin

Document type:	Master thesis	
Status:	Final	
Date:	17-05-2016	
Place:	Enschede	
University:	University of Twente	
	Faculty of Engineering Technology	
	Department of Water Engineering and Management	
External institute	Deltares	
	Department of Hydrology	
Author:	Ing. H. Trul	
	hizkiatrul1@gmail.com	
Supervisors:	Prof. dr. J.C.J. Kwadijk	University of Twente and Deltares
	Dr. ir. M.J. Booij	University of Twente
	Dr. F.C. Sperna Weiland	Deltares
	Ir. M. Hegnauer	Deltares

Photograph on cover: 1995 flood in Waal, source: http://www.heemkundeverenigingleeuwen.nl/pdf_watersn/Hoogwater1995-2015.pdf

Preface

In this master thesis my findings regarding the characteristics of Rhine flood waves simulated with GRADE are presented. This study is inter alia done to graduate from the master Water Engineering and Management at the University of Twente. I spent approximately 8 months on this study, which was generally a fun thing to do. I learned a lot about how to do scientific research in the hydrological field. Mainly the calculations and interpretation of the calculation results was interesting work in my opinion.

First of all I would like to thank Jaap Kwadijk for giving me the opportunity to work on this interesting project. His feedback helped me to understand how scientific research should be done and to improve this thesis. Also Martijn Booij was of great support during the whole study for which I am grateful to him. I want to thank both Frederiek Sperna Weiland and Mark Hegnauer for all their help and good advice. My classmates and colleagues from the University of Twente and Deltares helped me with difficulties I experienced during the process. They also positively distracted me from the work sometimes, which I very much appreciated, because this made me come to other better ideas and thoughts in some cases. Finally I would like to thank my friends and family who supported me during the whole graduation process. Especially I would like to thank Gwen Kamphuis for being there for me to vent my ideas and frustrations and for giving me the support I needed.

Hizkia Trul

Deventer, 17-05-2016

Summary

Hydraulic boundary conditions, with a low occurrence probability, are needed to carry out quality assessments of flood protection measures constructed in and around the Dutch Rhine. The physically based method, called Generator Rainfall and Discharge Extremes (GRADE), is used to determine these hydraulic boundary conditions. Within GRADE synthetic weather, generated by resampling of 56 years of historical precipitation and temperature data, is fed into the hydrological model HBV to simulate continuous daily discharge series. Extreme flood waves, selected from the continuous discharge series, will be used as the hydraulic boundary condition to assess the required stability of for example the dikes. Disapproved dike stretches should be reinforced, which might have large financial implications and can lead to public resistance. It is therefore important that the physical characteristics of flood waves simulated with GRADE are in accordance with reality. The objective of this research is to assess the performance of the hydrological model HBV and the combined performance of the weather generator (WG) with HBV, used within GRADE, in simulating the flood wave characteristics (peak discharge, peak timing, volume, duration and number of flood waves per year) and the contributions to flood waves at Lobith of 7 major sub-basins in the Rhine basin.

The flood waves have been selected from the continuous observed and simulated discharge series by the use of a threshold value and a time window. Observed and simulated flood waves from the period 1951-2006 have been compared to each other, to assess the performance of the HBV model in simulating the flood wave characteristics. For each characteristic the ratio between observed and simulated is calculated to detect structural over- or underestimations, the mean absolute relative error from the maximum simulated or observed characteristic (MAREM) is calculated to quantify the difference and the coefficient of determination (R^2) is calculated to assess the linear relation between the observed and simulated flood wave characteristic. The performance of HBV, HBV-WG and WG in simulating the flood wave characteristics has been evaluated by comparing the statistics of the flood characteristics, obtained from observed and simulated flood waves from the period 1951-2006 and simulated flood waves from 10.000 year synthetic discharge series. A t-test has been used to assess the equality of the means, a F-test has been used to assess the equality of the variances and a cumulative distribution function (CDF) has been used to visualize the differences between the observed and simulated flood wave characteristics.

The results showed that the performance of HBV and HBV-WG in simulating the volumetric contributions of the 7 major Rhine sub-basins to flood waves detected at Lobith is good. Contributing discharges from the Moselle are a little underestimated by errors in the synthetic weather series.

The results showed that the performance of HBV, HBV-WG and WG in simulating all flood wave characteristics of the whole Rhine basin at Lobith and Andernach is good. Also the characteristics of flood waves from the Main are simulated well. The simulated flood wave characteristics from the Moselle and Neckar differ slightly from the observed ones. For the Neckar this is mainly due to the HBV model. The errors detected in flood waves from the Moselle can be attributed to both the HBV model and WG. The flood wave characteristics from both Alpine sub-basins are poorly simulated due to the HBV model.

The largest errors are found in flood waves from the East Alpine Rhine basin. The peak discharges and volumes of the winter flood waves are overestimated, whereas peak discharges of all flood waves from this basin are less overestimated. The volumes calculated from all flood waves are underestimated.

Flood wave durations from winter flood waves are less underestimated than the durations of all simulated flood waves. Flood wave peaks of waves that contribute to Lobith waves are often simulated earlier than the observed ones, whereas assessing all flood waves reveals that flood wave peaks are simulated too late. The performance of HBV in simulating snow storage is probably responsible for the main errors. Presumably too little water is allocated to the snow storage, so that in winter there will be too much discharge, whereas in early summer there will be too little. All detected errors in this basin can be explained by this possible reason.

Overall the peak discharge is the best simulated flood wave characteristic. Flood wave volumes, durations and number of flood waves are generally underestimated. The HBV model is in most cases responsible for the largest errors. Often the performance of HBV-WG is slightly worse than the performance of HBV only. The skill of the WG in reproducing comparable weather is good, however errors in flood wave characteristics simulated with HBV are often slightly increased due to extra uncertainty incorporated in the WG.

It is recommended to do an in depth validation of HBV models for the Alpine region to assess the skill of these models in simulating the underlying hydrological processes that drive the discharge. It is furthermore recommended to be reserved in using GRADE in its current form for river flood applications in the Netherlands. Flood protection measures which will be disapproved, due to hydraulic boundary condition obtained from GRADE, should be ameliorated. People negatively affected by projects concerning the improvement of flood protection measures might use the large differences found for the Alpine region flood waves as argument against using GRADE. Because the Alpine region is responsible for 29% of the total Lobith wave discharge, those people have a point. It is therefore recommended to improve the HBV models for the Alpine region. Assessing possible GRADE extensions to simulate for example flooding in the Netherlands due to dike failure might be an interesting next step in GRADE's development.

Contents

List of Symbols, Abbreviations and Acronyms

$alpha$	Parameter for the non-linear behaviour in the response function used in the HBV models of German sub-basins
$AMSL$	Above mean sea level
$BAFU$	Bundesanstalt für umwelt (in English: (Swiss) federal office for the environment (FOEN))
$beta$	HBV parameter to control the increase in soil moisture for every mm of precipitation
BfG	Bundesanstalt für gewässerkunde (in English: (German) federal institute of hydrology)
CDF	Cumulative distribution function
$cfmax$	HBV parameter to control snowmelt rate in the Alpine region basins (mm/day)
CHR	International commission for the hydrology of the Rhine basin
DWD	Deutch wetterdienst (in English: German weather service)
$E - OBS$	European land-only, high-resolution gridded observational dataset
f	Probability density function
$F(x)$	Probability of non-exceedance
fc	Parameter of the maximum value of the soil moisture storage, used in HBV (mm)
$F - test$	Statistical test, used to assess the equality of the variances, named after Sir R.A. Fisher
GEV	Generalized extreme value
$GLUE$	Generalized likelihood uncertainty estimation
$GRADE$	Generator rainfall and discharge extremes
$GRDC$	Global runoff data centre
$H0$	Null hypothesis of the statistical tests
$H1$	Alternative hypothesis of the statistical tests
HBV	Hydrologiska byråns vattenbalansavdelning (in English: the water balance department of the hydrological bureau (in Sweden))
$HYMOG$	Hydrologische modellierungsgrundlagen im rheingebiet (in English: Hydrological modelling basis in the Rhine basin)
$HYRAS$	Hydrologische rasterdaten (in English: hydrological gridded data)
i	Index number of the flood event
k	Lag used for autocorrelation (days)
khq	Recession parameter at high flow used in HBV (1/day)
$KNMI$	Koninklijk Nederlands meteorologisch instituut (in English: royal Dutch meteorological institute)
$Lobith\ waves$	Only the threshold waves from upstream Rhine sub-basins that contribute to flood waves at Lobith
lp	Limit for potential evaporation, HBV parameter used for German basins
MAE	Mean absolute error
$MAPTE$	Mean absolute peak time error
$MAREM$	Mean absolute error from maximum
N	Total number of flood events
n_o	Number of the analysed observed flood wave characteristic
n_s	Number of the analysed simulated flood wave characteristic
NSE	Nash and Sutcliffe efficiency
O	Observed

obs	Observed
$perc$	Percolation parameter used in HBV (mm/day)
P_o	Time step of observed peak discharge
P_s	Time step of simulated peak discharge
$p - value$	Probability value used for statistical tests
Q	Discharge (m^3/s)
$Q5$	Discharge that is exceeded in 5% of the time (m^3/s)
r	Autocorrelation
R^2	Coefficient of determination
$REVE$	Relative extreme value error
$RIZA$	Rijksinstituut voor integraal zoetwaterbeheer en afvalwaterbehandeling (in English: institute for inland water management and waste water treatment)
RVE	Relative volume error
s	Pooled standard deviation
S	Simulated
$S(T)$	Simulated extreme discharge for a return period T (m^3/s)
sim	Simulated
T	Return period (years)
$T5$	The discharge corresponding to a return period of 5 years, obtained from the Gumbel and GEV distributions (m^3/s)
$T20$	The discharge corresponding to a return period of 20 years, obtained from the Gumbel and GEV distributions (m^3/s)
$threshold\ waves$	Flood waves from upstream Rhine sub-basins selected by the use of a Q5 threshold and time window specific for that sub-basin
tt	Threshold temperature above which snowmelt occurs used for the HBV models of the Alpine region basins (°C)
$t - test$	Statistical test, used to assess the equality of the means, also known as student-test
WG	Weather generator
WTI	Wettelijk toets instrumentarium (in English: legal assessment instrument)
$ZWE\ areas$	Zwischeneinzugsgebieten (in English: intermediate basins)
λ	Temporal correlation length (days)
λ_Q	Correlation length of continuous observed discharge series (days)
λ_{Q_5}	Estimated correlation length for observed discharges exceeding Q5 (days)
μ_{Q_o}	Mean of the observed flood wave characteristic
μ_{Q_s}	Mean of the simulated flood wave characteristic
$\rho(\tau)$	Autocorrelation for a specific lag
σ_{Q_5}	Standard deviation of observed discharge exceeding Q5
$\sigma_{Q_o}^2$	Variance of the observed flood wave characteristic
$\sigma_{Q_s}^2$	Variance of the simulated flood wave characteristic
σ_Q	Standard deviation of the observed discharge series
τ	Lag used for autocorrelation (days)

1 Introduction

The conducted research is introduced in this chapter. The background, state of the art knowledge, research gap, research objective and questions and report outline are discussed respectively in paragraph 1.1, 1.2, 1.3, 1.4 and 1.5.

1.1 Background

International setting

The impact of river flooding on the security of people, material losses and economic damages is substantial (Kundzewicz et al., 2010). To reduce the river flood risk, hard and soft flood protection measures are applied in and around rivers all over the world. Soft flood protection measures primarily focus on reducing the impact of a flood, rather than preventing from one. Hard flood protection measures are constructed to avoid flooding during periods when river discharges are high due to for example extreme precipitation and/or snowmelt upstream in the basin. Assessing the quality of flood protection measures requires information about extreme discharges. One method to obtain low probability extreme discharges is by selecting them from long synthetic discharges series, which are generated by the use of generated weather series fed into a hydrological model. Several studies describe methods that couple weather generators to hydrological models to simulate continuous discharge series from which low probability extreme discharges are obtained (Blazkova & Beven, 2004; Haberlandt et al., 2008; Kuchment & Gelfan, 2011; Hegnauer et al., 2014; Falter et al., 2015). The goal of all these studies is to provide extreme discharges for assessments concerning river flood risks.

Dutch setting

Hard flood protection measures, applied in the Netherlands to protect the land from flooding's of the main rivers Rhine and Meuse, are for example dikes and room for the river projects. In order to assess the quality of the dikes, the Dutch Water Act oblige dike managers to do a dike stability assessment every 6 years, the intention is however to reduce this frequency to once every 12 years (*Bestuursakkoord Water*, 2011). The assessment will be done as described in the legal assessment instrument (WTI). The failure mechanisms, the mechanisms that can lead to dike failure, are assessed based on the dike's reaction to the loaded hydraulic boundary condition. Much research is being done in order to gain knowledge about protection from river floods. The ongoing research results in alterations in flood protection standards and in new methods to assess the quality of the flood defence infrastructure. Therefore the WTI is updated before every dike assessment round. The upcoming, fourth assessment round will start in 2017 (EURECO, 2015). One major alteration, which will be implemented in the new WTI, is that the inundation probability of a dike stretch will be assessed rather than the exceedance probability of high water levels (Ministerie van Infrastructuur en Milieu & Ministerie van Economische Zaken, 2014).

Until now the exceedance probability for the primary flood defences around the main rivers is set at once every 1250 years (Ministerie van Verkeer en Waterstaat, 2007). A discharge with this probability of occurrence is obtained from extrapolating historical peak discharge series. The obtained discharge and hence the water level is the hydraulic boundary condition that the levees have to withstand. If the water level becomes higher than the crest height minus the minimum freeboard, the levee will be disapproved. In fact only the failure mechanism overtopping is assessed. The probability that a dike or part of it fails and the area behind it inundates will be assessed in the near future. This inundation probability is based on consequences of dike failure (Ministerie van Infrastructuur en Milieu & Ministerie van Economische Zaken, 2014). A dike that protects many people will for example get a smaller allowed inundation probability than one that does not protect a single person. The allowed inundation probability of a specific dike can differ between once every 300 years up to once every 100.000 years. Furthermore not only dike overtopping will be assessed, but all failure mechanisms that can lead to dike failure.

Because of the strict norms formulated to protect from extreme river floods, information about discharges with a small probability of occurrence up to once every 100.000 years might be necessary to assess the quality of the dikes. Because the until now used time series of peak discharge based on observations is relatively short, approximately 100 annual discharge peaks, a lot of uncertainty is incorporated by extrapolating to a design discharge in the order of 1/100.000 year. Furthermore physical behaviour like for example upstream flooding is not explicitly incorporated by the extrapolation of extreme discharge peaks (Hegnauer et al., 2014). Therefore a more physically based method called, Generator Rainfall and Discharge Extremes (GRADE), will be used in the upcoming dike assessment round to determine the hydraulic boundary conditions (Knoeff & Steffess, 2014).

In GRADE a weather generator (WG), a hydrological model and a hydraulic model are coupled in order to generate low probability discharges. The WG uses a technique called nearest neighbour resampling to simulate long synthetic daily precipitation and temperature series for the Rhine basin (Schmeits et al., 2014a). The hydrologic response to this simulated weather data is calculated with the conceptual semi distributed HBV model (Hegnauer et al., 2014). By a simplified Muskingum approach the hydrological river routing is applied. The Rhine from Maxau to Lobith, along with the downstream sections of the tributaries Neckar, Main, Nahe, Lahn, Moselle, Sieg, Ruhr and Lippe are hydraulically routed with the use of the hydraulic model called SOBEK (Hegnauer et al., 2014).

1.2 State of the art knowledge

Other studies describe models that calculate the occurrence of low probability floods in a comparable manner as has been done in GRADE (Blazkova & Beven, 2004; Haberlandt et al., 2008; Kuchment & Gelfan, 2011; Falter et al., 2015). The validation done to assess the performance of these models in simulating discharges, focuses primarily on historical peak discharges calculated with the applied hydrological model. Blazkova and Beven (2004) calibrated their model with the use of the GLUE method, just like has been done in GRADE. Haberlandt et al. (2008) validated the skill of the used hydrological model in simulating the discharge by comparing them with historical observations by the use of the Nash and Sutcliffe efficiency criterion. Falter et al. (2015) checked the performance of their model by validation of the different model components. The skill of the hydrological model has been assessed by comparing the simulated discharge with historical observed discharge series calculated with the Nash and Sutcliffe efficiency criterion. Kuchment and Gelfan (2011) assessed the performance of their model by validation of the observed and simulated hydrographs for the period from 1960 to 1980. They did the validation based on flood volumes and peak discharges and calculated the performance by the Nash and Sutcliffe efficiency criterion. Haberlandt et al. (2008) and Falter et al. (2015) evaluated the performance of the combination between weather generator and hydrological model by a visual comparison of the flood frequency curves calculated from the simulations and observations.

Hegnauer et al. (2014) validated the Rhine part of the HBV model combined with the SOBEK model for historical flood events at Lobith and 5 other upstream gauging stations. They assessed the skill of the model based on comparing historical flood peak discharges with simulated flood peak discharges.

Furthermore the observed and simulated discharge series of the period between 1993 and 1995, containing two of the three highest measured discharges, have been compared using the Nash and Sutcliffe efficiency criterion. Only the skill of the applied hydrological and hydraulic model in simulating peak discharges has been validated with these validation steps. The combined performance of the hydrological model and weather generator used in GRADE has only been assessed by comparing frequency discharge curves, calculated from simulated and observed annual peak discharges at Lobith. From all executed validation steps it can be concluded that the performance of the model components of GRADE in simulating flood peaks at Lobith is good (Hegnauer et al., 2014).

Evaluation of the discussed methods focuses primarily on the whole discharge series and peak discharges simulated with the hydrological model. The performance in simulating flood wave characteristics like peak timing, volume, duration and number of flood waves per year is generally not evaluated. The performance of the combination between hydrological model and weather generator is assessed for the most downstream gauge in the basin by the use of only the annual maximum discharges by Haberlandt et al. (2008), Hegnauer et al. (2014) and Falter et al. (2015).

1.3 Research Gap

Scientific significance

From the state of the art knowledge it can be concluded that the evaluation of the performance of models, in which a weather generator is combined with a hydrological model, in simulating low probability discharges is primarily focused on peak discharges. All discussed models are however designed for river flood studies. Focussing on the peak discharge solely, says only little about the performance of these models in simulating low probability flood waves used in these river flood assessments. A more comprehensive evaluation of the performance in simulating other flood wave characteristics like peak timing, volume, duration and number of waves per year will show if the model is appropriate to simulated low probability river floods. The importance of the peak timing is that it shows if peak discharges are simulated at the correct moment in time. The volume and duration are important, because they give an indication about how well the discharges during flood waves are simulated overall. Assessing the number of flood waves shows the performance of the model in simulating the annual or long term appearance of flood waves. It is furthermore of interest to assess the performance of the model in simulating the flood wave characteristics of upstream basins and the volumetric contributions of these basins to downstream flood waves. This analysis namely shows if the model is applicable to simulate flood waves from upstream sub-basins and if the model simulates downstream flood waves in a physically correct manner.

Dutch public significance

All dike failure mechanisms are important when assessing the inundation probability of dikes in the Netherlands. It is therefore important that the performance of GRADE in simulating the characteristics of the flood waves is assessed. To assess for example piping and macro instability not only the flood wave peak discharge, but also the wave duration, volume and number of waves per year are important. The discharge corresponding to the peak of the flood waves is of interest, because this value is the maximum discharge that a levee has to withstand during a flood wave.

Being the tool used to determine the boundary conditions for assessing the hydraulic loading on the flood protection structures along the river Rhine, knowledge about the performance of GRADE in simulating the physical characteristics of flood waves in the Rhine is necessary. The model is designed to simulate physically reasonable low probability flood events, by the use of calculations based on simplifications of reality. It is therefore important that calculated values are not only comparable with observed values, basically reproducing the data, but are also physically plausible. For example a good simulated discharge at Lobith can be based on a too high contribution of sub-basin *A* due to the weather generator and a too low contribution of sub-basin *B* due to the HBV model. Because of this the performance of GRADE in producing flood events at Lobith based on truthful contributions from the main tributaries and model components is of interest as well. To do so it is necessary to assess the performance of GRADE in simulating the flood wave characteristics at Lobith and upstream sub-basins. The performance assessment of GRADE in simulating flood wave characteristics focuses on the hydrological model and the WG. The hydraulic modelling is not incorporated in this analysis, because the simulation time to obtain the required discharge series is long.

1.4 Research objective and questions

To use GRADE for calculating the hydraulic boundary conditions in the Rhine, it is important that the instrument simulates the flood wave characteristics in a physically correct manner. Therefore research have been done to achieve the next objective.

<u>The objective is to assess the performance of the hydrological model HBV and the combined performance of the weather generator and HBV, used within GRADE, in simulating flood wave characteristics (peak discharge, peak timing, volume, duration and number of flood waves per year) and the contributions of 7 major Rhine sub-basins to flood waves at Lobith.</u>

The next two research questions are formulated to guide the research. First the performance of HBV in simulating the flood wave characteristics from the 7 major sub-basins is assessed. The performance is assessed in two ways, 1. with the focus on sub-basin flood waves that contribute to flood waves at Lobith and 2. with the focus on all flood waves from the sub-basins. Secondly the combined performance of the HBV model and the weather generator is assessed by statistically comparing the flood wave characteristics obtained from measurements, HBV simulated discharges and synthetic discharge series simulated with HBV fed with generated weather series.

1. *How well are the flood wave characteristics peak discharge, peak timing, volume, duration and number of waves per year from the Rhine at Lobith and upstream sub-basins simulated with HBV when comparing the flood wave characteristics obtained from discharge observations and simulations?*
2. *What is the performance of the combination of HBV and WG in simulating the flood wave characteristics peak discharge, volume, duration and number of flood waves per year of flood waves from the Rhine at Lobith and upstream sub-basin when comparing the characteristics of flood waves obtained from the observed and simulated discharge series?*

1.5 Report outline

The report is structured as follows:

Chapter 2 GRADE, Study Area and Data. A description is given of the GRADE model at first in this chapter. Secondly the division of the Rhine basin is discussed and a description of the characteristics of the defined sub-basins is given. Finally the data used in the research is summarized.

Chapter 3 Methods. This chapter describes the methods used to assess the performance of HBV and HBV-WG in simulating the flood wave characteristics. First the selection of the flood waves from the continuous discharge series is given. Secondly the flood wave characteristics are discussed. Thereafter the procedure to evaluate the performance of the HBV model in simulating the flood wave characteristics is discussed. Finally the procedure to evaluate the combined performance of HBV and WG in simulating the flood wave characteristics is discussed.

Chapter 4 Results of HBV evaluation. In this chapter the results to answer research question 1 are discussed. For each flood wave characteristic separately the performance of HBV is described.

Chapter 5 Results of HBV, HBV-WG and WG evaluation. The results to answer research question 2 are presented in chapter 5. For each flood wave characteristic separately the results are discussed.

Chapter 6 Discussion. This chapter describes the limitations of choices made within this research. The applicability of the GRADE outcomes for Dutch river flood assessments is discussed as well. Furthermore the international applicability of the research outcomes is discussed.

Chapter 7 Conclusions and Recommendations. In this chapter the conclusions of the research are discussed per research question. Recommendations followed from the findings are discussed as well.

2 GRADE, Study Area and Data

This chapter is about the GRADE model, the division of the Rhine basin used to assess the performance of the model in simulating flood wave characteristics and the used data. Paragraph 2.1 is about GRADE and focusses on the weather generator and hydrological model. The calibration of the hydrological model is described extensively in paragraph 2.1.3, because this calibration gives crucial information needed to understand the model behaviour. The division of the Rhine basin and the characteristics of the sub-basins are discussed in paragraph 2.2. Finally the data used for the evaluation is described in sub-chapter 2.3.

2.1 Generator of Rainfall And Discharge Extremes (GRADE)

The description of GRADE is based on the work done by Hegnauer et al. (2014). GRADE is designed to provide a more physically based method, than the extrapolation method formerly used, for the estimation of extreme discharge probabilities for the river basins of the Rhine and Meuse. To reduce the uncertainty in estimating the design discharge by extrapolation of historical annual maximum discharges, Parmet et al. (1999) published the first attempt of a more physically based approach for the Rhine basin. This development was coordinated by the former institute of inland water management and waste water treatment (RIZA) and executed in association with the royal Dutch meteorological institute (KNMI) and the German federal institute of hydrology (BfG). The researchers developed a methodology consisting of a stochastic weather generator and the hydrological model HBV. The stochastic weather generator generates, through nearest neighbour resampling, low probability weather conditions on the basis of historical input data. The generated weather is input for a hydrological model, which computes the corresponding runoff for each sub-basin. The runoff from all basins is input for a hydraulic model, called SOBEK, which calculates discharges and water levels for the main downstream channels of both river basins. First calculations with this new GRADE method showed promising results for the main tributaries of the river Rhine (Eberle et al., 2002). Figure 1 shows the different components of GRADE and their relation.

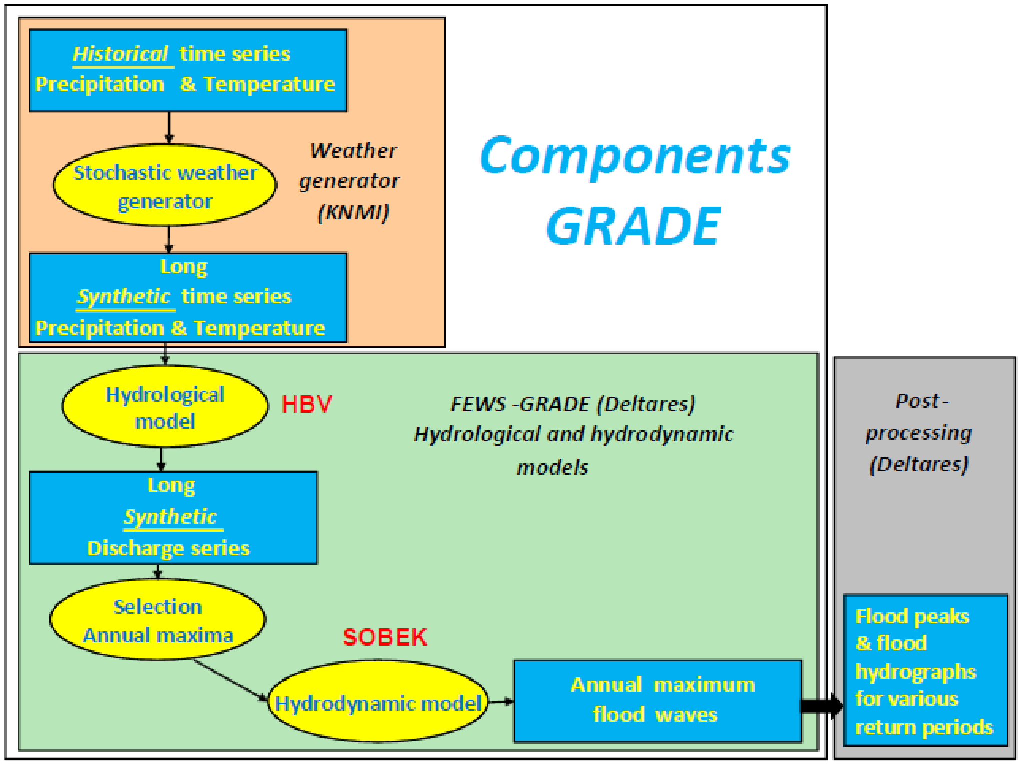
Historical time series
Precipitation & Temperature
Stochastic weather generator
Weather generator (KNMI)
Components GRADE
Long Synthetic time series
Precipitation & Temperature
Hydrological model
HBV
FEWS -GRADE (Deltares)
Hydrological and hydrodynamic models
Post-processing (Deltares)
Long Synthetic Discharge series
Selection Annual maxima
SOBEK
Hydrodynamic model
Annual maximum flood waves
Flood peaks & flood hydrographs for various return periods

Figure 1 Components of GRADE (Hegnauer et al., 2014)

2.1.1 The stochastic weather generator

The type of WG is chosen to fit the requirements needed for the goal of the instrument, namely flood probability assessment. Extreme historical floods in the Dutch part of the Rhine are mainly influenced by multi-day precipitation amounts, rather than by single-day precipitation (Schmeits et al., 2014a). The 1995 Rhine flood can for example be related to extreme 10-day precipitation amounts (Ulbrich & Fink, 1995). Therefore single day extreme rainfall events are not that important for the genesis of Rhine floods. The resampling technique known as nearest neighbour resampling is chosen to generate weather series, because this technique enables the creation of more extreme multi-day precipitation amounts with the use of only the observed daily precipitation. The synthetic weather series are based on resampling of the observed data. The stochastic WG produces daily synthetic precipitation and temperature series for the 134 sub-basins in the Rhine basin. The weather data grids, HYRAS (only precipitation) and E-OBS (both precipitation and temperature) (Schmeits et al., 2014a), are used as input for the WG. The spatial resolutions of the grids are respectively 5 km * 5 km and 25 km *25 km. The data sets consist of daily values with information for the 56-year period between 1951 and 2006. They are constructed based on the information from different rainfall and temperature stations in the basin. The station data are interpolated to construct the grids (Schmeits et al., 2014a). The stochastic WG works as follows. To model the amount of precipitation or the temperature at time step *n*, first 61 values are picked from a moving windows centred around the day of interest within each year. This is done to deal with seasonal variability. The historical data set consist of 56 years, so 56*61=3416 values are selected. For incorporating spatial dependencies and autocorrelation a feature vector is used to select values with similar characteristics as the one at the previous time step, the so called nearest neighbours. The WG for the Rhine basin uses a feature vector of three elements to find the nearest neighbours in the historical data. It uses the standardized daily temperature, averaged over the 134 sub-basins, the standardized daily precipitation, averaged over the 134 sub-basins and the fraction of sub-basins with daily rainfall larger than 0,3 mm (Schmeits et al., 2014a). The standardized values are the deviations of the long-term calendar day average values. Standardization is done to reduce the effect of the annual cycle on the selection of the nearest neighbours (Schmeits et al., 2014b). The fraction of sub-basins with daily rainfall exceeding 0,3 mm helps to distinguish between large-scale and convective precipitation (Eberle et al., 2002). With the feature vector 10 nearest neighbours are selected. The number of nearest neighbours was set to 10, because larger values generally worsen the reproduction of the autocorrelation coefficient. Values with characteristics nearest to the previous day will be selected to be the nearest neighbours. For both the precipitation and temperature one of the 10 nearest neighbour values is picked randomly to be the value for time step *n*. The spatial correlation is preserved, because all grid cells get the value corresponding to the sampled day from the observed series. In the random selection, a decreasing kernel is used to give more weight to the closest neighbours. The randomly chosen value, *n*, is used to select the next value, *n*+1. This procedure is followed until the desired time series length is obtained. See figure 2 for a schematization of this resampling procedure.

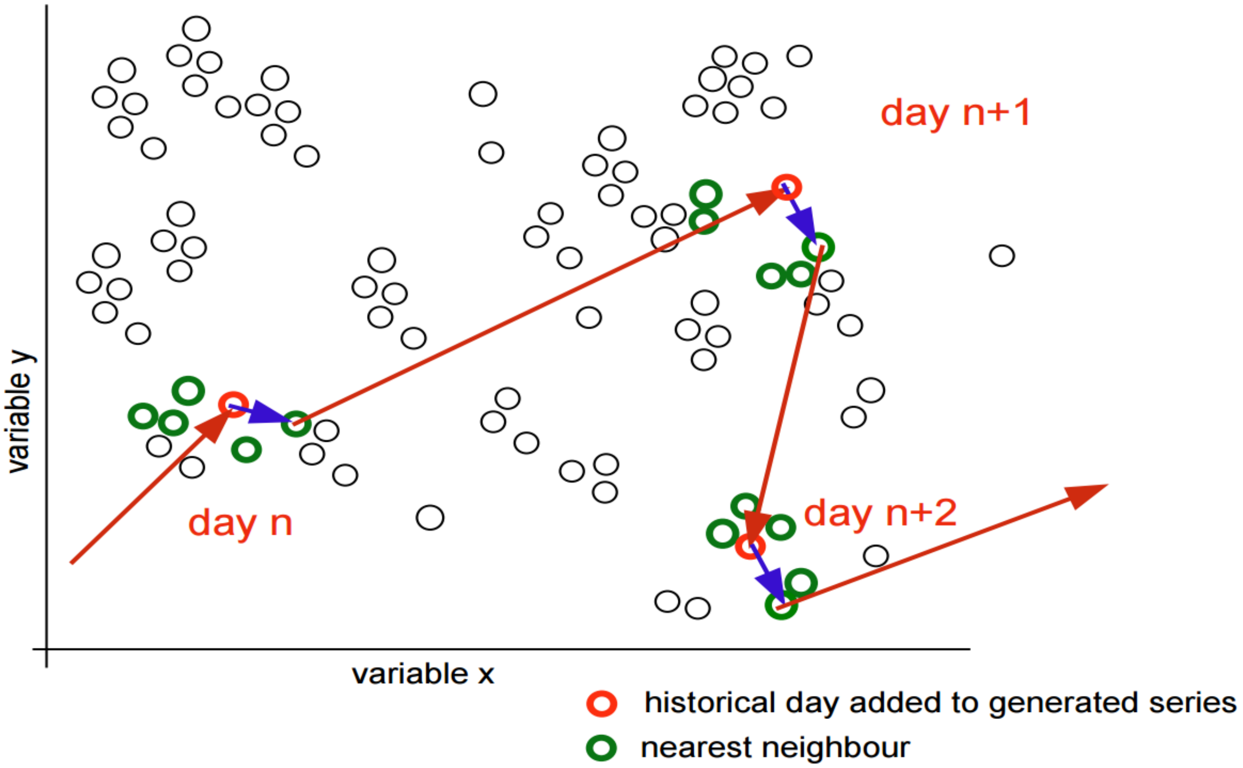

Figure 2 Schematization of the nearest neighbour resampling technique for two variables. (Leander & Buishand, 2004)

2.1.2 Hydrological model

The generated weather series are input for the HBV hydrological rainfall-runoff model that calculates the daily discharge for all of the 148 sub-basins. The schematization of 148 sub-basins comprises 130 of the 134 sub-basins often used for hydrological modelling and the other 4 divided into 18, because of the lakes in Switzerland (figure 4). Time series of daily sub-basin averaged values for the temperature and precipitation are used to calculate the discharge on a daily basis. The HBV96 version is used within GRADE, figure 3 shows the schematization of the interaction between the components of this model. The choice for this model is based on the evaluation of different hydrological models to apply in the Meuse and Rhine basins (Passchier, 1996).

HBV is a conceptual semi-distributed rainfall runoff model. It uses temperature and precipitation to calculate snow cover, evaporation, soil moisture storage and runoff (Lindström et al., 1997). HBV describes the most important runoff generating processes in a simple and robust manner. First in the snow routine, the accumulation or melt of snow is calculated based on the temperature and precipitation. Secondly within the soil routine, precipitation and melt water is allocated to runoff and/or evaporation and/or soil moisture. Thirdly within the runoff generation routine a fast runoff flow and a base flow is calculated. In the transformation function the actual discharge of a sub-basins is calculated using the MAXBAS parameter, which is a routing parameter that simulates the lag and attenuation occurring throughout the basin (Winsemius et al., 2013). Finally with a simplified Muskingum approach the hydrological river routing between sub-basins is simulated. This Muskingum method is based on the mass balance equation. It calculates the outflow from a basin by the inflow, along with a time parameter for travel time between in- and outflow point, plus the change in storage in the basin (Shaw et al., 2011). The discharge series simulated by the use of the Muskingum routing is used to find the maximum annual discharge at Lobith. Only for a time window of 30 days before until 20 days after this maximum discharge hydraulic routing using SOBEK is applied.

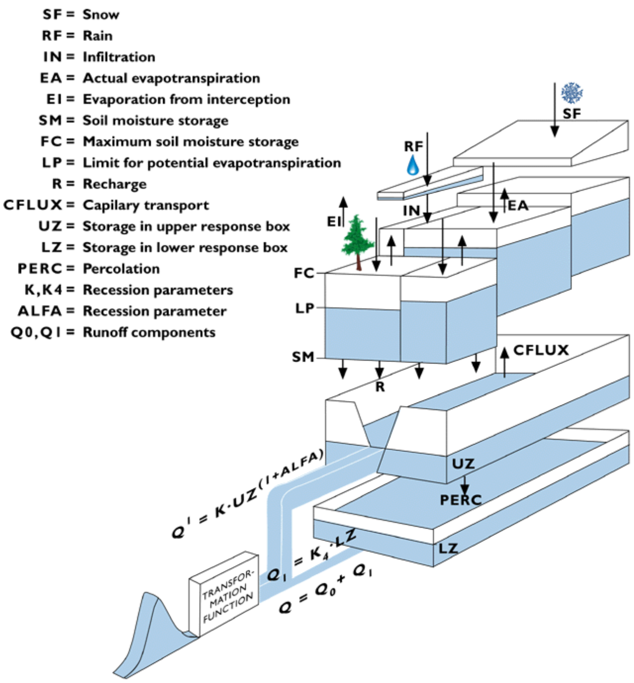
SF = Snow
RF = Rain
IN = Infiltration
EA = Actual evapotranspiration
EI = Evaporation from interception
SM = Soil moisture storage
FC = Maximum soil moisture storage
LP = Limit for potential evapotranspiration
R = Recharge
CFLUX = Capilary transport
UZ = Storage in upper response box
LZ = Storage in lower response box
PERC = Percolation
K,K4 = Recession parameters
ALFA = Recession parameter
Q0,Q1 = Runoff components
SF
RF
IN
EA
EI
FC
LP
SM
R
CFLUX
UZ
PERC
LZ
Q' = K·UZ^(1+ALFA)
Q1 = K4·LZ
Q = Q0 + Q1
TRANSFORMATION FUNCTION

Figure 3 Schematization of the HBV model ("The HBV model," 2015)

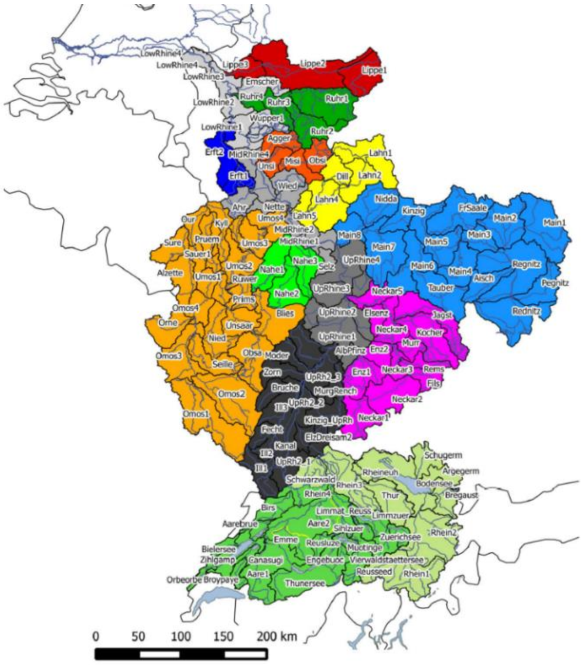

Figure 4 Sub-basins of the Rhine HBV model (Hegnauer et al., 2014)

2.1.3 Calibration HBV

Hegnauer and Verseveld (2013) and Winsemius et al. (2013) calibrated the HBV models of the sub-basins for which reliable data was available. They grouped all 148 sub-basins into 15 major sub-basins, see figure 4. The 15 major sub-basins have been calibrated independent from each other. Therefore no inflow from other major sub-basins have been used, this has been done by excluding the sub-basins in which the Rhine channel is located from this calibration. The sub-basins within these 15 major sub-basins have been calibrated from upstream towards downstream. The outflow of calibrated upstream basins has been used as inflow to more downstream sub-basins. The HYRAS 2.0 precipitation dataset and the E-OBS v4 temperature dataset have been used as input, which are the same as have been used to generate the long synthetic weather series. They calibrated the model for the period 1989-2006, by the use of HYMOG discharge data (Steinrücke et al., 2012), which was sometimes completed with GRDC data (Hegnauer et al., 2014). The calibration has been done by optimizing the parameter values in order to obtain the best correspondence between simulated and observed discharges. It is however possible that multiple parameter sets give approximately the same results. Because of this reason the Generalized Likelihood Uncertainty Estimation (GLUE) method has been applied. It allows for multiple parameter sets to be applicable for describing the hydrology in a basin, hereby representing the uncertainty in the hydrological model parameterization.

First they conducted a Monte-Carlo analysis for the most upstream sub-basins in each of the 15 major sub-basins. In this analysis the model has been run 5000 times with different parameter values randomly picked from a pre-defined uniform distribution of each parameter. This is a reasonable number of runs, because Shrestha et al. (2009) found that the statistics for testing convergence were stable after 5000-10.000 simulations. A division has been made between the parameters used for the basins in the Alpine and the other sub-basins in the Rhine basin, because the influence of snow on discharges from the Alpine region is large, whereas this is not the case in the other sub-basins. See table 1 for the parameters used.

Table 1 Parameters used to calibrate HBV for the Rhine basin (Hegnauer & Verseveld, 2013; Winsemius et al., 2013)

Parameter	**Non Alpine region sub-basins (German part)**	**Alpine region sub-basins (Swiss part)**
	unit	*unit*
fc = Maximum value of the soil moisture storage	mm	mm
lp = Limit for potential evaporation	-	Not used
perc = Percolation	mm/day	mm/day
beta = Control for the increase in soil moisture for every mm of precipitation	-	-
alpha = Parameter for the non-linear behaviour in the response function	-	Not used
khq = Recession parameter at high flow	1/day	1/day
tt = Threshold temperature above which snowmelt occurs	Not used	°C
cfmax = Snowmelt rate	Not used	mm/day

The performance of each parameter set has been assessed with performance criteria and only the ones that meet the constraints of the criteria have been selected as the so called behavioural parameter sets. In table 2 the used performance criteria along with the constraints can be seen.

Table 2 Performance criteria used to calibrate HBV for the Rhine basin, along with the constraints used to select the behavioural parameter sets (T5 is the discharge corresponding to the 5 year return period, obtained from Gumbel and GEV distributions, T20 is the discharge corresponding to the 20 year return period, obtained from the Gumbel and GEV distributions)

Performance measure	**Constraints**

Nash and Sutcliffe efficiency	Should belong to the 10% highest NSE values obtained from the Monte-Carlo analysis.
Relative Volume Error	<0,1
Relative Extreme Value Error (T5 and T20)	<0,1

- Nash and Sutcliffe efficiency (see equation (1)) is a measure to assess the overall performance in simulating the discharge series. It is however biased towards errors in high flows and therefore high discharges will get more weight than lower ones (Legates & McCabe, 1999). This results in higher influence of peak discharges on the value of this criterion and lower influence of low flow conditions on the outcome.

  $$NSE = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(O_i - O_{mean})^2} \qquad (1)$$

  In which N is the total number of data points, S is the simulated discharge, O is the observed discharge and *i* is the index number of the data point.
- The relative volume error evaluates the long-term volumetric error. It is calculated by the summed difference between the observed and simulated discharges divided by the summed observed discharge, see equation (2).

  $$RVE = \frac{\sum_{i=1}^{N}(O_i - S_i)}{\sum_{i=1}^{N} O_i} \qquad (2)$$

  In which$N$,S,O and *i* are the same as used in equation (1).
- The relative extreme value error measures the deviation of the observed and simulated extreme values. It is calculated by subtracting the observed extreme discharge from the simulated one and dividing the result by the observed extreme discharge, see equation (3). The once in 5 year (T5) and once in 20 year (T20) extreme discharge are used in this calibration. The extreme values are obtained from both Gumbel and GEV distributions fitted through the observed and simulated discharge series.

  $$REVE = \frac{S(T) - O(T)}{O(T)} \qquad (3)$$

  In which $S(T)$ is the simulated extreme discharge for a return period T and $O(T)$ is the observed extreme discharge for a return period T.

Only single sub-basins have been calibrated for which appropriate discharge measurements were available. If measurements were not present multiple sub-basins were calibrated as a whole by the use of the available measurement data. The calibration process started with the most upstream sub-basins in each of the 15 major sub-basin. The selection of behavioural parameter sets for the most upstream sub-basins has only been done based on the constraints presented in table 2. The downstream neighbour sub-basins have been calibrated by combining the discharge calculated with a random selected parameter set for the sub-basin to calibrate with the input discharge from the already calibrated upstream sub-basin. This procedure has been repeated until the most downstream point each of the 15 major sub-basin was reached. For each sub-basin between the 10 and 100 behavioural parameter sets were defined. Only the sub-basins through which the Rhine flows, the ZWE areas (Zwischeneinzugsgebieten) are not calibrated with this procedure, because the contribution of those areas was expected to be small (Winsemius et al., 2013). The parameters for these ZWE areas are copied from calibrated sub-basins with a comparable average slope, because the slope is related to the hydrological processes that play an important role in the sub-basin. Figure 5 shows the highest obtained NSE values for the different sub-basins obtained during the GLUE analysis, also the uncalibrated sub-basins can be seen.

Winsemius et al. (2013) selected the parameter set used for the GRADE calculation as follows. Because GRADE has been designed to assess floods in the Rhine, parameter sets have been selected based on the calculation of the maximum 1/10 year discharge of each sub-basin. For each sub-basin all behavioural parameter sets have been used to calculate the 1/10 year discharge. The median 1/10 discharge has been selected from all simulations. For each sub-basin the parameter set used to calculate the median 1/10 year discharge has been selected as the parameter set to do the GRADE calculations. Validation of the derived parameter sets has not been done yet.

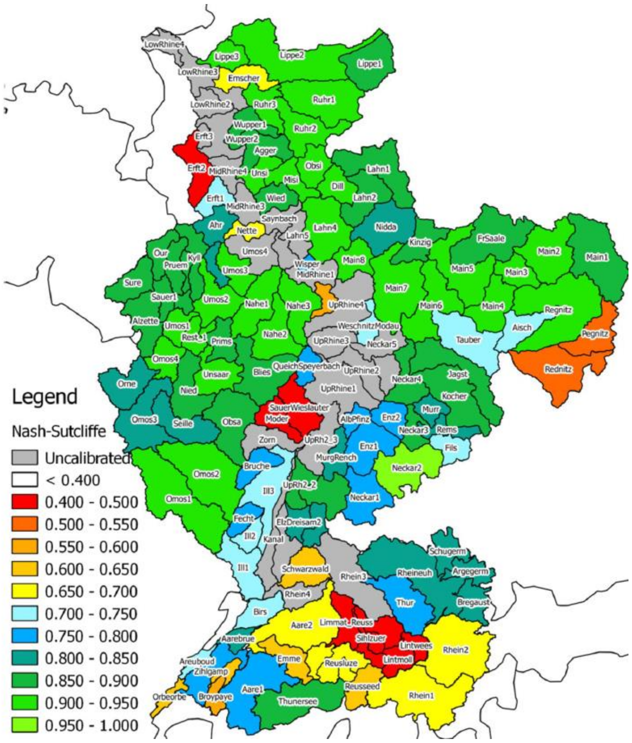

Figure 5 Highest NSE values for the different sub-basin HBV models obtained from the GLUE analysis (Hegnauer et al., 2014)

2.2 Division of the Rhine basin

2.2.1 Sub-basin division

The evaluation of the model is done by assessing the simulated flood waves selected from the discharges of 7 large upstream sub-basins, see figure 6. This division is used because other hydrological studies that focus on the Rhine basin used this same sub-division (Demirel et al., 2013). The next outlet stations are used, because these stations are located at the downstream sides of the different sub-basins and for these gauge locations the longest discharge series are available. Lobith for the Lower Rhine (LR), Andernach for the Middle Rhine (MR), Cochem for the Moselle, Frankfurt for the Main, Rockenau for the Neckar, Rekingen for the East Alpine Rhine (EA) and Untersiggenthal for the West Alpine Rhine (WA). The discharge at Lobith and Andernach consist of runoff from the sub-basin in where the station is located and the inflow from upstream sub-basins.

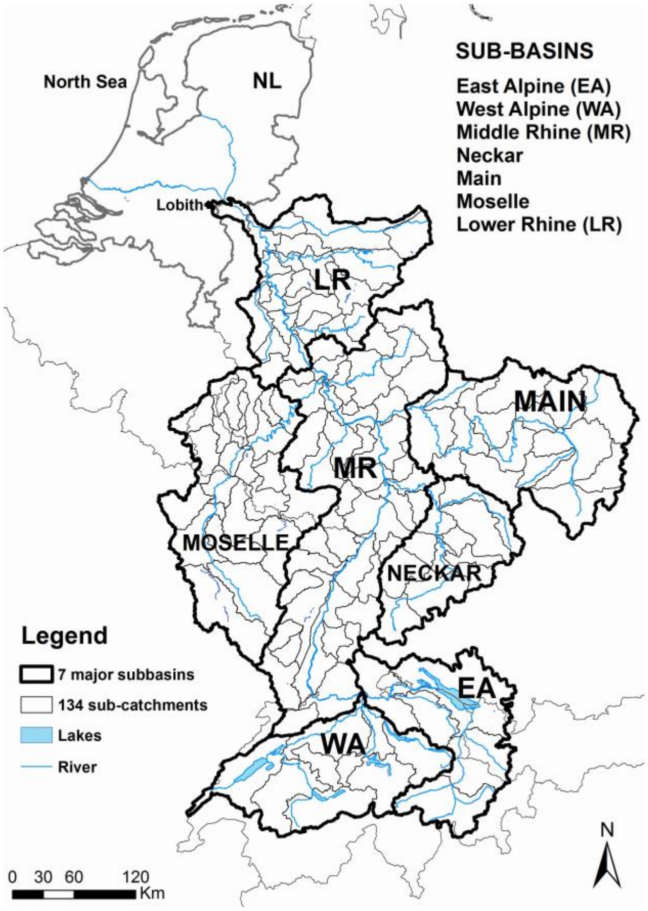

Figure 6 Seven major sub-basins of the Rhine upstream of Lobith (Demirel et al., 2013)

2.2.2 Sub-basin description

The surface area of the whole Rhine basin is approximately 185.000 km^2, from which 25.000 km^2 is located in the Netherlands. About 50% of the basin area is used for agriculture, 31,7% of the area is forest and 8,8% of the basin is classified as urban area (Tockner et al., 2009). The length of the river flowing from the Alps to the North Sea is about 1320km. In the Rhine basin two discharge regimes can be distinguished, the nival regime, which is dominated by snowfall and snowmelt, with low discharge in winter and high discharges in early summer, and the pluvial regime, which is dominated by net precipitation, with high discharges in winter and low discharge in summer (Belz et al., 2007). The average discharge of the river at Lobith is 2300m^3/s, the maximum discharge ever observed, in the year 1926, is 12.600m^3/s (Nienhuis, 2008). During summer more than 70% of the discharge at Lobith originates in the Alpine region, whereas in winter this is only 30% (Middelkoop & Haselen, 1999). The next descriptions of the different sub-basins are based on Tockner et al. (2009) and Tongal et al. (2013).

Lower Rhine

The surface area of the Lower Rhine is 23.738 km^2, the range in altitude is 5-779 meters above mean sea level (AMSL). The land use in the area is dominated by agriculture (38,4%), forest (27,9%) and urban

(18,3%). The length of the main river stretch is approximately 230 km. The discharge from this basin is composed of the inflowing water from the Rhine at Andernach and the runoff from this basin resulting from the net precipitation, which is approximately 273 mm per year.

Middle Rhine

The surface area of the Middle Rhine is 37.908 km^2, the range in altitude is 67-1340 meters AMSL. The land use in the area is dominated by forest (38,5%), agriculture (36%), pasture (16,4%) and urban (6,8%). The length of the main river stretch is approximately 500 km. The discharge from this basin is composed of the inflowing water from the Moselle, Main, Neckar, Alpine region and the runoff from this basin resulting from the net precipitation, which is approximately 344 mm per year.

Moselle

The surface area of the Moselle basin is 27.262 km^2, the range in altitude is 59-1326 meters AMSL. The land use in the area is dominated by agriculture (54%) and forest (37%), 6,7% is urban area. The length of the main river stretch is approximately 544 km. The discharge from this region results from the runoff of the approximately 365 mm net precipitation per year. The Moselle River is adapted to be a waterway for large cargo vessels. The adaption required the construction of 28 weirs with locks to manage the water levels. These weirs influence the natural discharge from the basin mainly during dry periods in order to ensure enough water for navigation.

Main

The surface area of the Main basin is 24.833 km^2, the range in altitude is 83-939 meters AMSL. The land use in the area is dominated by agriculture (54%) and forest (38%), 6,9% is urban area. The length of the main river stretch is approximately 524 km. The pluvial discharge regime from this region is fed by 255 mm net precipitation per year. The river is characterized by winter floods caused by rainfall. A complex of 34 weirs is used to regulate the water levels in the river, which disturb the natural discharge from the region during mainly dry periods.

Neckar

The surface area of the Neckar basin is 12.616 km^2, the range in altitude 90-970 meters AMSL. The land use in the area is dominated by agriculture (53%) and forest (36%), 10,2% is urban area. The length of the main river stretch is approximately 367 km. The pluvial discharge regime from this region is fed by 337 mm net precipitation per year. The flow variation from this region is high. In the river 27 weirs are constructed to regulate the water levels for navigation and hydro electrical power production. These anthropological influences disturb the natural discharge from the Neckar.

East Alpine Rhine

The surface area of the East Alpine Rhine basin is 16.051 km^2, the range in altitude is 143-3270 meters AMSL. The land use in the area is dominated by nature, namely 37,4% natural grasslands, 26,3% sparsley vegetated area and 22,6% forests. Only 1,9% of the area is classified as urban area. The primarily snow melt runoff of some nival head waters, including the Alpine Rhine, flow into lake Constance. The damped flow from this large lake determines largely the discharge from this region. A net precipitation of 890 mm per year is discharged from this region.

West Alpine Rhine

The surface area of the West Alpine Rhine basin is 17.679 km^2, the range in altitude is 252-4274 meters AMSL. The main river in this region is the Aare, which has a nival discharge regime. The length of the river is 295 km. The land use in the area is dominated by agriculture (38%) and forest (28%), 2,1% of the area is covered with glaciers and 3,2% of the area is urban area. Three major lakes are located in the West Alpine Rhine, namely Lake Neuchâtel, Lake Lucerne and Lake Zürich. The flow in the Aare is furthermore highly influenced by hydropower production. Nine power plants and seven reservoirs are located in the headwaters of this river. A net precipitation of 1003 mm per year is discharged from the sub-basin.

2.3 Data

Table 3 gives an overview of the used data sets. For most discharge stations a combination between GRDC (Global Runoff Data Centre) data and HYMOG data is used to obtain the required series length. The GRDC provided BfG discharge and water level data for the German sub-basins and BAFU discharge and water level data for the Swiss sub-basins. The discharge measurements of the BfG data are collected by the German Federal Institute of Hydrology and the BAFU data is collected by the Swiss Federal Office for the Environment. The hourly HYMOG discharge data is produced by Steinrücke et al. (2012). Where there is overlap between GRDC and BfG the GRDC data is prioritized for the selection of discharge series for the period 1951-2006. GRADE simulates discharges on a daily basis, so daily measured series are required as well. The hourly HYMOG data is therefore converted to daily data by averaging the hourly discharges of each day. The E-OBS version 4.0 (Haylock et al., 2008) containing daily gridded temperature data has been used in the HBV simulations. The HYRAS 2.0 dataset (Rauthe et al., 2013) containing gridded daily precipitation was made available by the German Weather service (DWD) via the (BfG). The potential evaporation is calculated based on the air temperature and sunshine duration provided by the International Commission of the Hydrology of the River Rhine basin (CHR), the DWD, Météo France and MeteoSchweiz, using the Penman-Wending approach (Eberle et al., 2005).

Table 3 Overview of used data sets

sub-basin	station	Type	time step	Source	available years	Selected years from source
Lower Rhine + upstream basins	Lobith	Discharge	Day	GRDC, BfG	1901-2008	1951-2006 1901-2008
Middle Rhine + upstream basins	Andernach	Discharge	Day	GRDC, BfG	1931-2003	1951-2003
			Hour	HYMOG	1989-2007	2004-2006
Moselle	Cochem	Discharge	Day	GRDC, BfG	1951-2001 and 2003	1951-2001
			Hour	HYMOG	1989-2007	2002-2006
Main	Frankfurt	Discharge	Day	GRDC, BfG	1963-1996	1963
			Day	GRDC, BfG	1964-2004	1964-2004
			Hour	GRDC, BfG	1990-2007	2005-2006
Neckar	Rockenau	Discharge	Day	GRDC, BfG	1951-2003	1951-2003
			Hour	HYMOG	1989-2007	2004-2006
West Alpine	Untersiggenthal	Discharge	Day	GRDC, BAFU	1935-2003	1951-2003
			Hour	HYMOG	1989-2007	2004-2006

	Lake Neuchâtel, Lake Lucerne and Lake Zürich	Lake water level	Day	GRDC, BAFU	1978-2008	1978-2006
East Alpine	Rekingen	Discharge	Day	GRDC, BAFU	1920-2003	1951-2003
			Hour	HYMOG	1989-2007	2004-2006
	Lake Constance	Lake water level	Day	GRDC, BAFU	1978-2008	1978-2006
Whole Rhine basin upstream of Lobith	Gridded (0,25 degree resolution)	Temperature	day	KNMI, E-OBS version 4.0	1950-2006	1951-2006
	Gridded (0,25 degree resolution)	Rainfall	Day	BfG, DWD, HYRAS 2.0	1951-2006	1951-2006
	Stations spread over the Rhine basin	Air temperature and Sunshine duration	Day	CHR, DWD, Météo France and MeteoSchweiz	-	-

3 Methods

In this chapter the methods used to assess the performance of HBV and HBV combined with the WG in simulating flood waves in the river Rhine is discussed. The way the flood waves are selected is described in paragraph 3.1. Subsequently the determination of the flood wave characteristics is discussed in paragraph 3.2. Thereafter in paragraph 3.3 and 3.4 the evaluation is discussed, which is divided into two assessments. In the first assessment the measured flood waves of the period 1951-2006 are compared with simulated flood waves. This first assessment is about the performance of HBV in simulating flood waves in the river Rhine and is described in paragraph 3.3. The second evaluation focuses on the performance of the combined performance of HBV and the WG in simulating the flood waves in the Rhine and is described in paragraph 3.4. Statistics are used to compare the simulated flood wave characteristics with those obtained from the observations.

3.1 Selecting the flood waves

In order to assess the performance in simulating flood wave characteristics, first flood waves should be obtained from the continuous discharge series. These flood waves are periods with high discharges. They are selected from the continuous daily discharge series by the use of two boundary conditions. A threshold value is used to determine which discharges belong to flood waves. Discharges above the threshold are part of the flood wave. A time window is used to determine the start and end of the wave, a flood wave starts or ends if in a window around the first and last discharge above the threshold no discharges above the threshold are found. In figure 7 an example of a selected wave at Lobith is given. The discharge between the most central vertical red lines belongs to the flood wave. The quantification of the boundary conditions is explained in the next paragraphs.

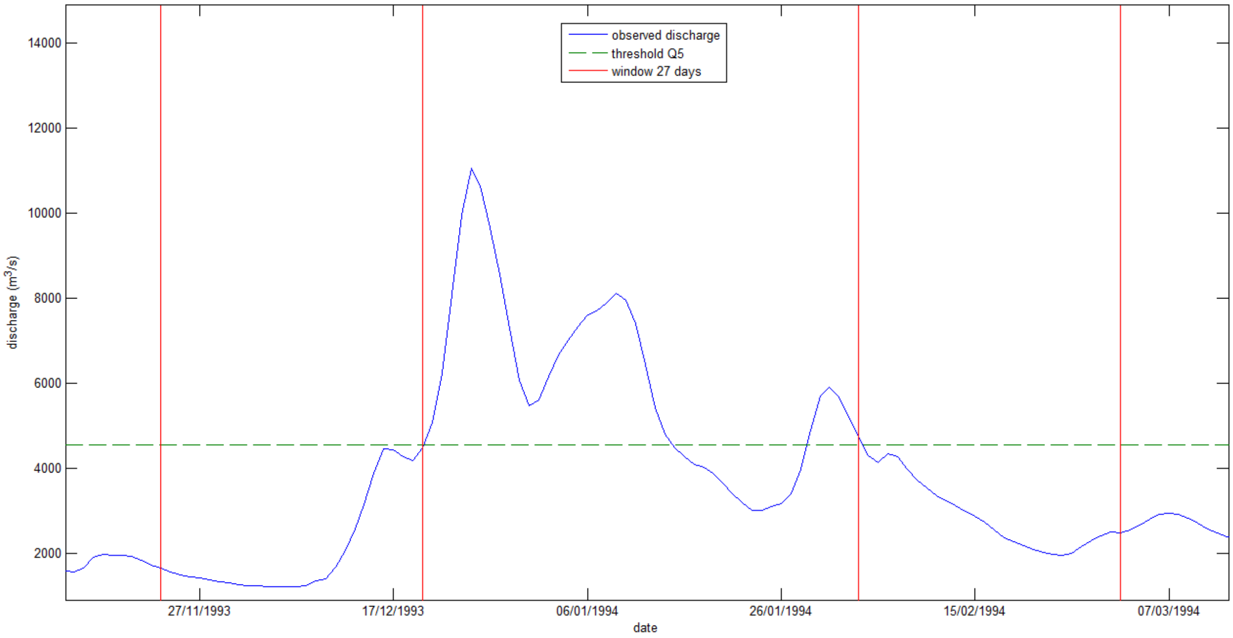

Figure 7 Example of the boundaries of an observed flood wave at Lobith, only the discharges above the green threshold line and between the two most central vertical red lines belong to the flood wave

3.1.1 Threshold

Observed and simulated flood waves are selected based on a threshold value. Discharges above the threshold will be allocated to a certain flood wave. Different threshold values are needed to select the waves of the 7 different outlet stations, because the discharge frequency distributions of these areas differ from each other. It is however important that the thresholds are consistent with each other in order to carry out a fair evaluation. It is therefore chosen that the threshold should have a clear statistical basis. In lowland rivers, where the river flows between floodplains, a physically logical threshold to select flood waves is the discharge corresponding to floodplain inundation levels, because water levels above this threshold will have an influence on the dikes. At Lobith this water level arises with discharges exceeding approximately 5000 m^3/s (Walker et al., 1993). However, when assessing the discharge of corresponding upstream sub-basins this measure cannot be applied, because not all upstream river stretches have flood plains. Therefore it is decided to use the discharge that is exceeded in 5% of time as statistical basis, the Q5 discharge, which is 4545 m^3/s at Lobith.

To assess the performance of the model the simulated flood waves are compared to measured waves. Therefore the Q5 threshold value is determined from the observed discharge series and used to identify the observed and simulated flood waves. See table 4 for an overview of the calculated threshold values for the different sub-basins.

Table 4 Q5 discharges calculated from the observed discharge series at the different outlet gauges

sub-basin	**station**	**Q5 observed (m^3/s)**
Lower Rhine + upstream basins	Lobith	4545
Middle Rhine + upstream basins	Andernach	4180
Moselle	Cochem	1010
Main	Frankfurt	502
Neckar	Rockenau	350
West Alpine Rhine	Untersiggenthal	1048
East Alpine Rhine	Rekingen	791

3.1.2 Window

All discharge values above the threshold are selected from the discharge series. Clusters of subsequent high discharges are attributed to one specific flood wave. Using a threshold results in allocating discharges to different flood waves, whereas they sometimes might logically be part of one single event. If a dike experiences multiple high waters in a relatively short period of time then there is for example not enough time for the water in the dike to flow out during periods with lower water levels. The successive high water levels will therefore cumulatively drive failure mechanisms like macro and micro instability. This is a reason to prevent ending a flood wave because the discharge is below the threshold for a few time steps. It is therefore chosen to allow for time steps with low discharge to be part of the flood wave in specific cases. This is done by the use of a window around the high discharges. So a flood wave ends when no discharges above threshold are found in a window behind the last discharge above the threshold from that specific flood wave.

The length of the window is estimated for each sub-basin separately. This is done by assessing the correlation between the discharges of subsequent time steps of the flood events. The assumption is that flood waves only exists of discharges that are correlated to each other. This implies that discharges below the threshold can only belong to the flood wave if they are temporarily correlated to the last previous discharge above the threshold. Autocorrelation, which can be calculated using equation (4), is used for this analysis.

$$r_k = \frac{\sum_{i=1}^{N-k}(Q_i-\bar{Q})(Q_{i+k}-\bar{Q})}{\sum_{i=1}^{N}(Q_i-\bar{Q})^2} \qquad (4)$$

r = autocorrelation

Q = discharge (m^3/s)

k = lag (days)

N = total number of discharge time steps

i = index number of the discharge time step

From the autocorrelation a correlation length can be calculated, which indicates from which lag on no significant correlation will be found. The correlation length estimated for discharge values above Q5 is of interest, however autocorrelation can only be calculated for continuous time series. Therefore the correlation length of the whole time series is used to estimate the correlation length for the discharges above Q5.

The way the correlation length can be calculated depends on the function of the autocorrelation. It seems that the autocorrelation, calculated for the continuous discharges of the different sub-basins, decays exponentially, see figure 8. An exponential temporal correlation function can therefore be assumed, because of this a comparable procedure as is used by Booij (2002) is followed. The assumed exponential temporal correlation function can be seen in equation (5).

$$\rho(\tau) = e^{(\frac{-\tau}{\lambda})} \qquad (5)$$

$\rho(\tau)$ = the autocorrelation value for a specific lag

τ = lag in days

λ = temporal correlation length in days

The temporal correlation length can be found by substituting equation (5) in equation (6) as is done by Whitehouse (2011).

$$\lambda = \int_0^\infty |\rho(\tau)| d\tau \qquad (6)$$

By doing so the temporal correlation length is found to be the length corresponding to an autocorrelation value of e^{-1}. Therefore the correlation length is the lag for which the autocorrelation becomes smaller than e^{-1}. In figure 8 the correlogram of the discharge at Lobith along with the line representing an autocorrelation of e^{-1} is shown as an example.

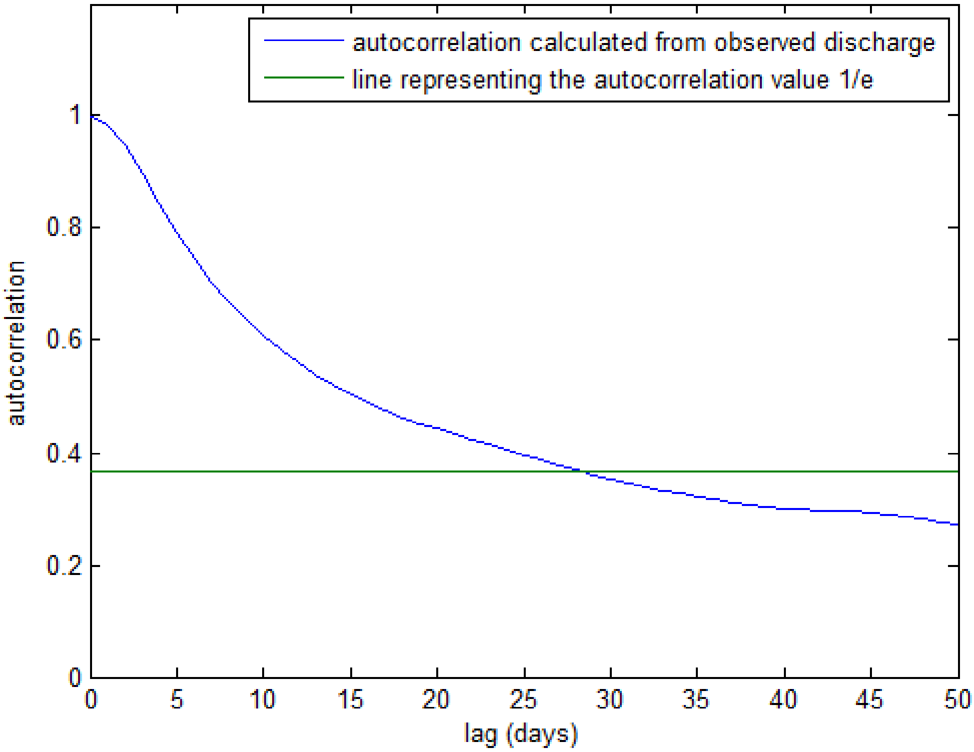

Figure 8 Correlogram of the discharge measured at Lobith

To estimate the correlation length belonging to the discharges above the threshold, it is assumed that the ratio in the standard deviation between the continuous time series and the discharge above Q5 is also the ratio in correlation length. The underlying assumption is that the larger the variability the smaller the autocorrelation. The physical assumption behind this is that more fluctuation in the flow leads to less correlation between discharges at subsequent time steps. This is however not always the case, if for example all discharges in a time series are multiplied by the same value, then the standard deviation increases or decreases, whereas the autocorrelation stays the same. Equation (7) shows how the correlation length and thus the window is estimated.

$$\lambda_{Q_5} \approx \frac{\sigma_Q}{\sigma_{Q_5}} * \lambda_Q \tag{7}$$

λ_{Q_5} = estimated correlation length for discharges exceeding Q5

σ_{Q_5} = standard deviation of discharges exceeding Q5

σ_Q = standard deviation of continuous discharge series

λ_Q = correlation length of continuous discharge series

The windows calculated for the different sub-basins are: 27 days at Lobith, 23 days at Andernach, 15 days at Cochem, 18 days at Frankfurt, 5 days at Rockenau, 71 days at Rekingen and 64 days at Untersiggenthal. Large windows are calculated for the discharge series at Rekingen and Untersiggenthal, because the flow from these two regions is less variable than from the other sub-basins. The discharges from the Alpine region are namely influenced by discharges from large lakes and snowmelt, which both dampen the variability of the precipitation. In the other sub-basins this damping effect is much smaller. The sensitivity analysis presented in appendix 3 shows that mainly the number of waves is affected by these large windows. The influence on the performance of simulating the flood wave characteristics is small, see appendix 3.

3.2 Definition of the flood wave characteristics

Observed and simulated flood wave characteristics are used to assess the performance of the GRADE components in simulating flood waves. The used characteristics are peak discharge, peak timing, volume, duration and number of flood waves per year, see figure 9.

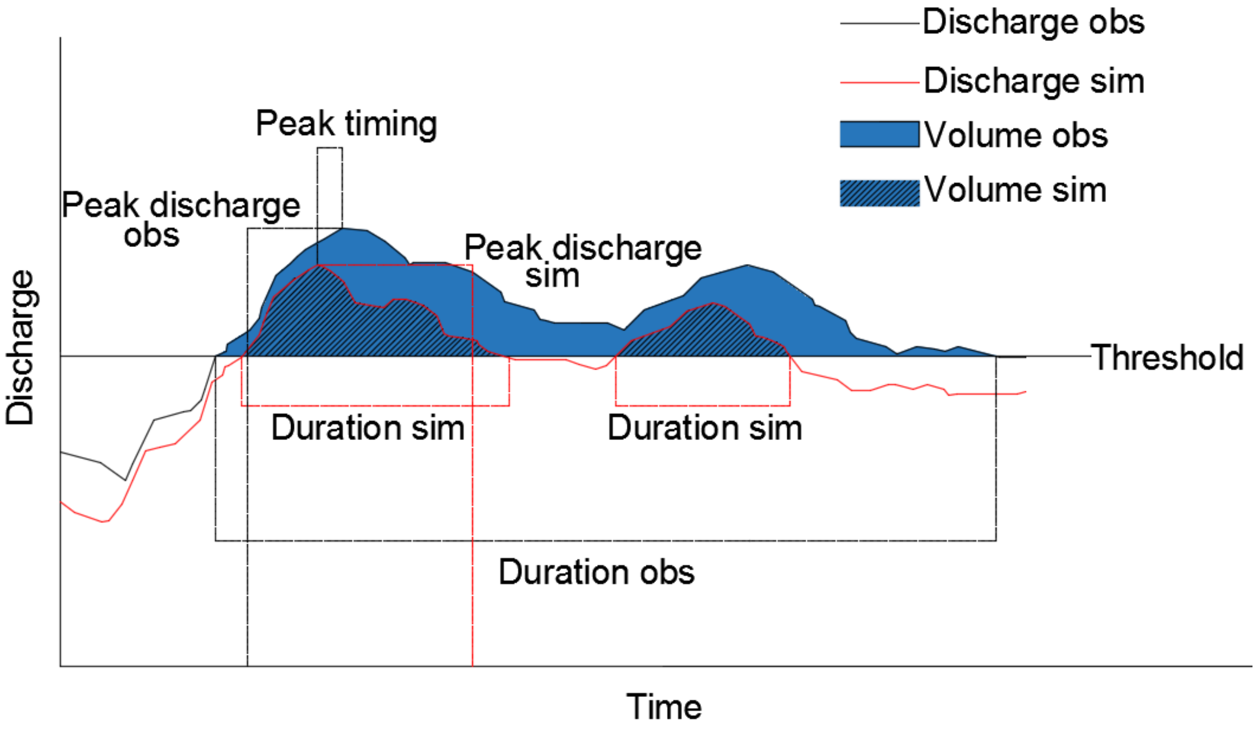

Figure 9 Definition of the observed and simulated flood wave characteristics

3.2.1 Peak discharge

The discharge corresponding to the peak discharge of the flood waves is important, because this value is the maximum discharge that a levee has to withstand during a flood wave. The peaks are selected from the identified observed and simulated flood waves by selecting the maximum discharge of the flood waves. Evaluating the performance of the HBV model in simulating the flood wave peaks is done by comparing the peak discharges from flood wave hits. In a flood wave there might be several peaks present, see for example figure 9. It is possible that in the observed flood wave the first peak is highest, whereas in the simulated flood wave the second peak is highest. To prevent that the peak discharges of these different peaks are used for the analysis, the highest simulated discharge is selected from a window of 5 days around the observed flood peak discharge for all sub-basins. In this way actual corresponding peaks are compared, rather than just the highest discharge of the flood waves.

3.2.2 Peak timing

Timing is the time difference in days between simulated and observed flood waves. The simulated waves can be earlier or later than the observed waves. The importance of the timing is that this characteristic reveals if the flood waves from sub-basins is simulated at the right moment in time, which is important for assessing the physically veracity of the simulated flood waves. To determine the timing a fixed point in the flood hydrograph will be used. This can for example be the start or the end of an event or the highest discharge of a flood wave. The timing determined based on start and end of a wave is highly influenced by the used threshold boundary. The timing should be zero if the discharge in a simulated flood wave is only structurally underestimated, however when using the difference between the first time step that the observed discharge is above the threshold and the first time step that the simulated discharge is above the threshold nonzero timings can be found, because the simulated wave has smaller discharges, which are later above threshold. To cope with this effect the timing in the peak is used, the time step of the peak is namely not influenced by structural over- or underestimation in the time series. The selected flood wave peaks, see section 3.2.1, used to determine the difference in peak discharge, are used to obtain the timing.

3.2.3 Flood wave volume

The flood wave volume is the total amount of water above the threshold during a flood. The importance of this characteristic is that it shows which volume of water is loaded to the flood defence structures.

3.2.4 Flood duration

The flood wave duration determines how long dikes are exposed to high water levels. The flood wave duration is defined as the time that the discharge is above the threshold. Not the difference between start and end of the flood wave is used as duration, because only discharges above the threshold are of interest when assessing the influence on the flood defence structures.

3.2.5 Number of flood waves per hydrological year

The number of flood waves per hydrological year give an indication about how often flood defences are exposed to high water levels. Most river flood waves appear around the start of the calendar year, see figure 10.

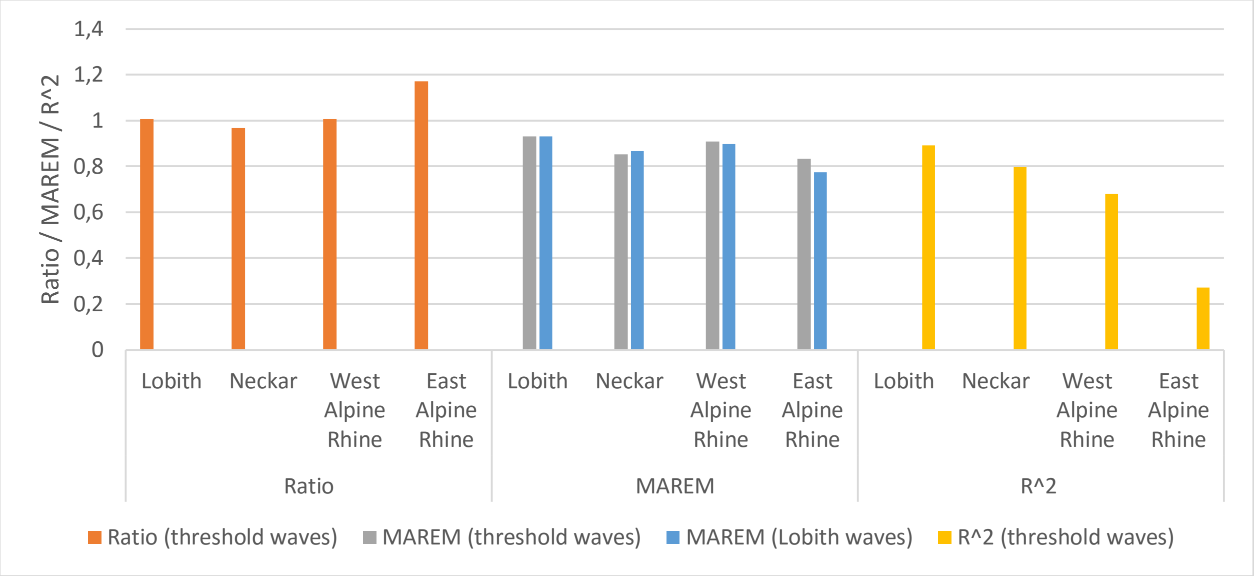

Figure 11 Performance of HBV in simulating the peak discharge of waves from the Rhine upstream of Lobith and three upstream sub-basins (values of 1 indicate good correspondence between simulated and observed waves, Ratio reveals if the model over- or underestimates, MAREM is a quantification of the absolute error and R^2 is the coefficient of determination)

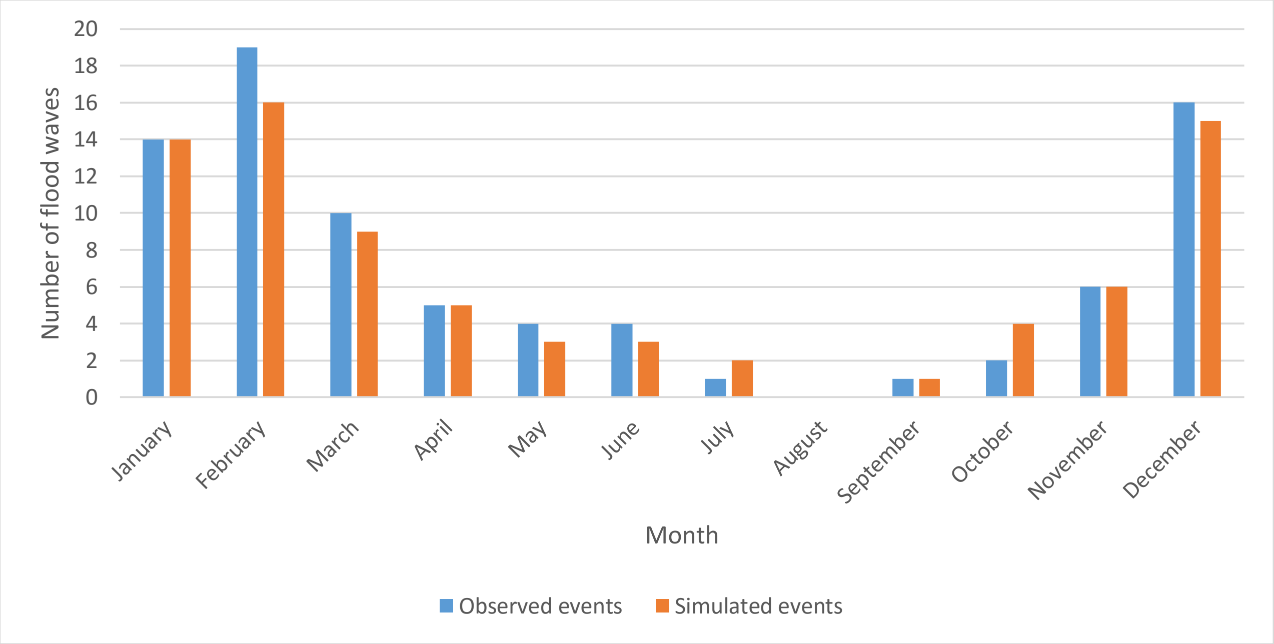

Figure 10 Number of observed and simulated flood events per month selected from the discharge series between 1951 and 2006 measured at Lobith

Because of this reason it is more logical to ascribe flood waves to the hydrological year, which starts the 1st of October and ends the 30th of September, instead of to the calendar year. The start date of the flood event determines to which hydrological year the event is attributed. Using the hydrological year instead of the calendar year ensures that most waves are allocated to the year they belong, because the number of waves that start at the end of the year and end at the beginning of the next year is smallest when using the hydrological year.

3.3 Evaluation of the performance of HBV in simulating flood waves for the period 1951-2006

The discussed flood wave characteristics are used to determine how well the performance of the HBV model is in simulating flood waves and discharge contributions to downstream flood waves of the 7 sub-basins for the period 1951-2006. Quantification of the performance in simulating the flood wave characteristics is done by the use of five different evaluation criteria, which are discussed in paragraph 3.3.1. The flood waves of the upstream sub-basins are selected in two ways, depending on the purpose of the analysis. First the contribution of the sub-basin discharges to measured flood waves at Lobith is assessed by selecting the upstream discharges by the dates on which flood waves at Lobith are observed, this is discussed in paragraph 3.3.2. Secondly the performance of the HBV model in simulating all flood waves from the 7 sub-basins is of interest, which is discussed in paragraph 3.3.3.

3.3.1 Evaluation criteria

In total five different evaluation criteria are used to assess the performance of the HBV model in simulating the flood wave characteristics. The mean absolute relative error from maximum (MAREM), ratio between the observed and simulated flood wave characteristics and coefficient of determination (R^2) are used to compare the observed and simulated flood wave characteristics, except peak timing. The mean absolute peak time error is used to quantify the time difference between observed and simulated peaks. A critical success index is used to show the difference between the number of observed and simulated events. The symbols *O*, *S, N* and *i* are used, where *O* means observed, *S* means simulated, *N* is the total number of flood wave events and *i* is the index number of the flood wave event.

3.3.1.1 Mean absolute relative error from maximum

A modified version of the mean absolute relative error is used to quantify the absolute relative difference between simulated and observed values. For each flood wave the value of the simulated characteristic is abstracted from the measured one and the absolute value of the outcome is divided by the highest value of the two. The average of all these absolute relative differences is calculated and subtracted from one, which is the MAREM, see equation (8).

$$\mathrm{MAREM} = 1 - \left(\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\frac{|(O_i - S_i)|}{\max(O_i, S_i)}\right) \tag{8}$$

The differences between the magnitudes of the flood wave characteristics of the various sub-basins are sometimes large. By the use of this measure a fair comparison can be made between those sub-basins, because the difference in absolute quantities does not disturb the outcomes of this formula. A MAREM value of 1 implies that the model perfectly reproduces the observed flood wave characteristics. Values near 0 indicate that differences are large.

3.3.1.2 Ratio between observed and simulated

The MAREM does not say anything about structural under- or overestimation of simulated flood wave characteristics. Therefore the ratio is calculated by dividing the sum of the simulated values with the sum of the observed values, see equation (9).

$$\mathrm{Ratio} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{S_i}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{O_i}} \tag{9}$$

A ratio of 1 suggests that there is no structural over- or underestimation. If the ratio is below one the model structurally underestimates the measurements. Ratios above one indicate that the model structurally overestimates the measurements.

3.3.1.3 Coefficient of determination

The combination of MAREM and ratio reveals how much the simulations differ from the observations and if there are structural over- or underestimations. Those measures do not indicate if the observed and simulated variability of the values from a specific flood wave characteristic correspond to each other. If for example the model simulates for each time step the average of the observed discharge, than the ratio will not indicate structural over- or underestimation and MAREM might be close to 1. One may conclude from those measures that the observed average is a good measure to use for simulation purposes. This is not always true, because the linear relation between observed and simulated might be wrong. To assess this linear relation the coefficient of determination (R^2) between observed and simulated is calculated. R^2 is an indicator of the extent to which the model explains the total variation in the observed data. It is calculated as:

$$R^2 = \left\{ \frac{\sum_{i=1}^{N} (O_i - O_{mean})(S_i - S_{mean})}{\left[\sum_{i=1}^{N} (O_i - O_{mean})^2\right]^{0.5} \left[\sum_{i=1}^{N} (S_i - S_{mean})^2\right]^{0.5}} \right\}^2 \tag{10}$$

R^2 is in fact the percentage of total variation in the simulation explained by the variation in the observations and is presented as a value between 0 and 1. A value of 1 means that the linear relation between observed and simulated is perfect, 100% of the simulated variation can be explained by the variation in the measurements. A value of zero means that there is no linear relation between observed and simulated values, this implies that one cannot say if the model will simulate a large or small value when a large value is measured.

3.3.1.4 Mean absolute peak time error

The peak timing cannot be assessed by the use of the previous explained evaluation criteria, because the timing is only one value, namely the difference between the observed and simulated peak time. Therefore the performance is alternatively assessed by the use of the mean absolute peak time error (MAPTE). It can be calculated using the next formula (Ehret & Zehe, 2011).

$$MAPTE = \frac{1}{N} * \sum_{n=1}^{N} \left| P_{o,n} - P_{s,n} \right| \tag{11}$$

N = total number of flood wave hits (see table 5 for the definition of a flood wave hit)
n = index number of the flood wave hit
P_o = time step of observed peak discharge (days)
P_s = time step of simulated peak discharge (days)

MAPTE gives the mean peak timing in days, because the discharge series are in days as well. The absolute difference is taken to avoid that positive and negative timings will average out. The value of MAPTE ranges between zero and infinity, in which zero means that there is no time difference between the observed and simulated flood wave peaks. Using this measure no emphasis is placed on too early or too late simulated waves, therefore also the mean peak time error is calculated to assess if the simulated flood waves are generally too early or too late. The mean peak time error differs from MAPTE because the actual difference rather than the absolute difference is used.

3.3.1.5 Critical success index

The previously explained performance measures are applied to assess the skill of the model in simulating the flood wave characteristics of the flood wave hits. The outcomes do not say anything about the skill of the model in simulating the occurrence of flood waves. Therefore the critical success index is used to indicate the correspondence between the number of simulated and observed flood waves. To do so the flood wave hits, false alarms and misses are calculated, see table 5 for the used definitions.

Table 5 Contingency table used to determine hits, false alarms, misses and correct negatives

		Observed discharge	
		> threshold	≤ threshold
Simulated discharge	> threshold	hits	false alarms
	≤ threshold	misses	correct negatives

With the identified hits, false alarms and misses the critical success index is calculated using equation (12) (Donaldson et al., 1975).

$$Critical\ Succes\ Index = \frac{hits}{hits + misses + false\ alarms} \tag{12}$$

The critical success index ranges from zero to one, in which one indicates optimal reproduction of the number of flood wave events. A value of zero means that no observed waves are simulated.

3.3.2 Volumetric contribution of upstream sub-basins to flood waves at Lobith

In this analysis the skill of HBV in simulating the contribution of the upstream sub-basins to measured flood wave events at Lobith is calculated. The discharges in these flood events are a superposition of the contributions of upstream sub-basins. First the observed flood waves at Lobith are identified. The dates for which flood waves are identified and the travel time from upstream stations to Lobith are used to select upstream discharge series. Accurately determining the contribution by the use of the travel time is difficult. Therefore the contribution is approximated using the maximum travel time in the Rhine basin instead of the actual travel time towards Lobith from each sub-basin separately. The maximum travel time of the water flow from the gauges in the Rhine basin towards Lobith is approximately 5 days (Bolwidt et al., 2007). To be conservative a 6 day window before the selected observed wave dates is used to select upstream discharge series. Furthermore a time difference between observed and simulated discharges might be present. A simulated discharge, which is simulated too late, will in such a case not be incorporated in the contribution calculation. To cope with this possibility it is chosen to use also a window of 6 days after the observed wave dates to select contributing discharge series. So if a flood wave observed at Lobith starts the 10th of January and ends the 20th of January, then observed and simulated discharges from the 4th until 26th of January are selected for all outlet stations. To make a fair comparison also the 6 days before and after the observed flood wave at Lobith are assessed. The total volumetric contribution (m^3/day) to the flood wave will be used, so for a fair comparison all contributions should have an equal number of time steps.

3.3.3 Performance of HBV in simulating sub-basin flood waves

Besides the performance of HBV in simulating the volumetric contributions of sub-basins to floods at Lobith, also the performance of HBV in simulating flood waves from the upstream sub-basins is of interest. The flood waves are selected in two different ways. First all sub-basin flood waves are selected as is discussed in chapter 3.1, these flood waves are called **threshold waves**. Secondly only those flood waves that contribute to flood waves at Lobith are selected in a comparable manner as is discussed in paragraph 3.3.2. Only sub-basin flood waves that start in the period 6 days before until 6 days after an observed flood wave at Lobith are selected, these flood waves are called **Lobith waves**. The simulated flood wave characteristics along with the occurrence of flood waves are evaluated by the use of the evaluation criteria discussed in paragraph 3.3.1. Only the skill of HBV in simulating the flood wave characteristics of flood wave hits is analysed in this assessment. However to assess the performance of simulating the number of flood waves per year not only the hits are incorporated, but all observed and simulated waves. This is possible because years are compared, whereas in the other analyses actual waves are compared.

3.4 Evaluation of the performance of HBV, HBV-WG and WG based on statistics

Statistical evaluation steps are conducted to assess the performance of HBV, HBV-WG and WG in simulating flood wave characteristics. Only statistics are compared because the generated weather used as input for HBV-WG does obviously not have to result in simulated discharges that are one to one corresponding to measured discharges. However the statistics of both the simulated and observed flood wave characteristics should at least be comparable. The statistics of the peak discharge, volume, duration and number of events per hydrological year are assessed using statistical tests, to assess the equality of means and variances (discussed in paragraph 3.4.1) and by cumulative distribution functions (discussed in paragraph 3.4.2.) In paragraph 3.4.2 the three comparisons to assess the performance of HBV, HBV-WG and WG are discussed.

3.4.1 Statistical tests to assess equality of mean and variance

An assessment of the equality of the mean and variance of the flood wave characteristics obtained from the measured, HBV simulated and HBV-WG simulated discharge series is conducted to assess the skill of the WG and the hydrological model. The assumption underlying this analysis is that a well performing model should calculate flood wave characteristics with the same statistics as those obtained from measured discharge series.

Mean

A two-sample t-test is used to check the hypothesis that the means of the flood wave characteristics are equal. A two-tailed test is done to test the hypothesis $H_0 : \mu_{Q_o} = \mu_{Q_s}$ against $H_1 : \mu_{Q_o} \neq \mu_{Q_s}$. The next three assumptions are made in order to use this statistical test.

1. The two vectors containing the flood wave characteristic are independent.

2. The two vectors containing the flood wave characteristic are normally distributed.

3. The variance of both vectors is equal.

The calculations are done by the use of equations (13) and (14) (Davis, 2002). Depending on the degrees of freedom and the confidence level a t-value will be obtained from the t-table. The null hypothesis will be rejected if the calculated t-value is smaller than the one obtained from the t-table.

$$t = \frac{\mu_{Q_O} - \mu_{Q_S}}{\sqrt{\frac{\sigma^2_{Q_O}}{n_O} + \frac{\sigma^2_{Q_S}}{n_S}}} \qquad (13)$$

μ_{Q_o} and $\mu_{Q_s}$are the means of the observed and simulated flood wave characteristic

$\sigma^2_{Q_o}$ and $\sigma^2_{Q_s}$are the variances of the observed and simulated flood wave characteristic

n_o and n_s are the number of the analysed observed and simulated flood wave characteristic

The pooled standard deviation, which is calculated by equation (14), is used as standard deviation when equal variances cannot be assumed.

$$s = \sqrt{\frac{(n_o-1)\sigma^2_{Q_o}+(n_s-1)\sigma^2_{Q_s}}{n_s+n_o-2}} \tag{14}$$

By the use of the ttest2 function in Matlab the exact probability value (p-value) is calculated. The underlying calculations done with this Matlab function are the same as described by Davis (2002). A t-test applied to exactly equal means provides a p-value of 1, which means that for all significance levels the null hypothesis of equality cannot be rejected. A significance level of 0,05 is used. For all tests in which the variance cannot be assumed equal, because the null hypothesis is rejected by the F-test, the approximate t-test is used, which is the default used in the ttest2 function of Matlab.

Variance

A two-sample F-test is used to check the hypothesis that the variances of both simulated (HBV and HBV-WG) and observed flood wave characteristics are equal. A two-tailed test is done to test the hypothesis $H_0 : \sigma^2_{Q_o} = \sigma^2_{Q_s}$ against $H_1 : \sigma^2_{Q_o} \neq \sigma^2_{Q_s}$. *F* is calculated by dividing the smallest variance by the highest variance, see equation (15) and (16) (Davis, 2002).

$$F = \frac{\sigma^2_{Q_o}}{\sigma^2_{Q_s}}, \qquad \text{if } \sigma^2_{Q_s} < \sigma^2_{Q_o} \tag{15}$$

$$F = \frac{\sigma^2_{Q_s}}{\sigma^2_{Q_o}}, \qquad \text{if } \sigma^2_{Q_o} < \sigma^2_{Q_s} \tag{16}$$

$\sigma^2_{Q_o}$= variance of the observed flood wave characteristic

$\sigma^2_{Q_s}$ = variance of the simulated flood wave characteristic

By the use of the F-table the minimum and maximum F-value can be obtained by the use of the degrees of freedom and the significance level. The degrees of freedom are the number of values from each dataset subtracted by 1. The significance level is chosen to be 0,05. The function vartest2 from MATLAB is used to do this calculation, which is the same as is described in Davis (2002). The outcome will be either 1 or 0, with 0 meaning the null hypothesis cannot be rejected, so with a significance of 5% it cannot be stated that the variances differ from each other. By the used Matlab function also the corresponding p-value is calculated. This value indicates at which significance level the null hypothesis will be rejected. So if the p-value is smaller than the set significance level, the null hypothesis will be rejected.

3.4.2 Cumulative distribution function

The cumulative distribution function (CDF) is used to visualize the difference between observed and simulated flood wave characteristics. The probability of randomly selecting a certain quantity of the examined flood wave characteristic can be obtained from the CDF. The CDF $F(x)$ is defined by equation (17).

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{17}$$

From equation (17) it can be seen that $F(x)$ is in fact the probability that a randomly selected flood characteristic is smaller or equal to a certain value. The CDF is calculated as follows. First the peak discharge, volume, duration and number of flood waves per year are calculated from the selected flood waves. The magnitudes of the characteristics are then sorted and plotted against the cumulative frequency, which is the position number of the characteristic quantity counted from the smallest value divided by the total number of events.

3.4.3 Steps to assess the performance of HBV, HBV-WG and WG

The discussed t-test, F-test and CDF are used to do the next three comparisons. In these comparison steps the performance of HBV, HBV-WG and WG in simulating the flood wave characteristics of the 7 sub-basins is assessed. The model component (HBV, HBV-WG or WG) responsible for the deviation of the simulation from the measurements is detected using this approach.

1. Comparison of observations from the historical period (1951-2006 for all stations except Frankfurt and 1962-2006 for Frankfurt) with HBV simulations for the same period to assess the performance of the HBV model in simulating the flood wave characteristics.
2. Comparison of the statistics of the observations from the historical period (1951-2006 for all stations except Frankfurt and 1962-2006 for Frankfurt) with the statistics of the 10.000 year simulations, simulated with HBV-WG, to assess the performance of HBV-WG in simulating the flood wave characteristics.
3. Comparison of the statistics of the 10.000 year simulation, simulated with HBV-WG, with the statistics of the HBV simulations for the historical period (1951-2006 for all stations except Frankfurt and 1962-2006 for Frankfurt) to assess the skill of the WG.

4 Results of HBV evaluation

The performance of the hydrological model HBV in simulating the flood wave characteristics assessed by comparing the measurements with simulations for the period 1951-2006 for the sub-basins in the Rhine is discussed in this chapter. The largest differences are found in the Neckar, West Alpine Rhine and East Alpine Rhine. Therefore the focus in this chapter is on those sub-basins. Also the results at Lobith are incorporated in all analyses, because the discharge at Lobith is central to this research. The performance of HBV in simulating the threshold waves is discussed for all performance measures. The result in simulating the Lobith waves is only implemented if the performance in simulating the Lobith waves differs substantially from the performance in simulating the threshold waves. The results are discussed per flood wave characteristic. The paragraphs 4.1 until 4.5 are respectively about the performance of HBV in simulating the peak discharge, peak timing, wave volume, wave duration and number of waves per hydrological year. In paragraph 4.6 the main conclusion are discussed.

4.1 Peak discharge

In figure 11 the performance of HBV in simulating the peak discharges of different sub-basins can be seen. In most basins some minor differences between simulated and observed peaks are found. This is expected, because the HBV calibration used for the GRADE simulations focuses on high flows, because the Nash Sutcliffe and the relative extreme value error criterion are used to determine the parameter values (Hegnauer et al., 2014). Just some minor difference between Lobith waves and threshold waves are detected. In general the flood wave peaks are underestimated, only the peaks from the East Alpine Rhine are structurally overestimated. The Alpine region peaks selected from the threshold waves are a slightly better simulated than those selected from the Lobith waves. Most Lobith waves appear in winter, whereas most Alpine region waves appear in early summer, see appendix 2. So when assessing the threshold waves especially early summer waves are assessed, whereas assessing Lobith waves focuses mainly on the winter waves. This indicates that peaks discharges from Alpine region waves that mainly occur in summer are slightly better simulated than the peak discharges from waves that mainly occur in winter.

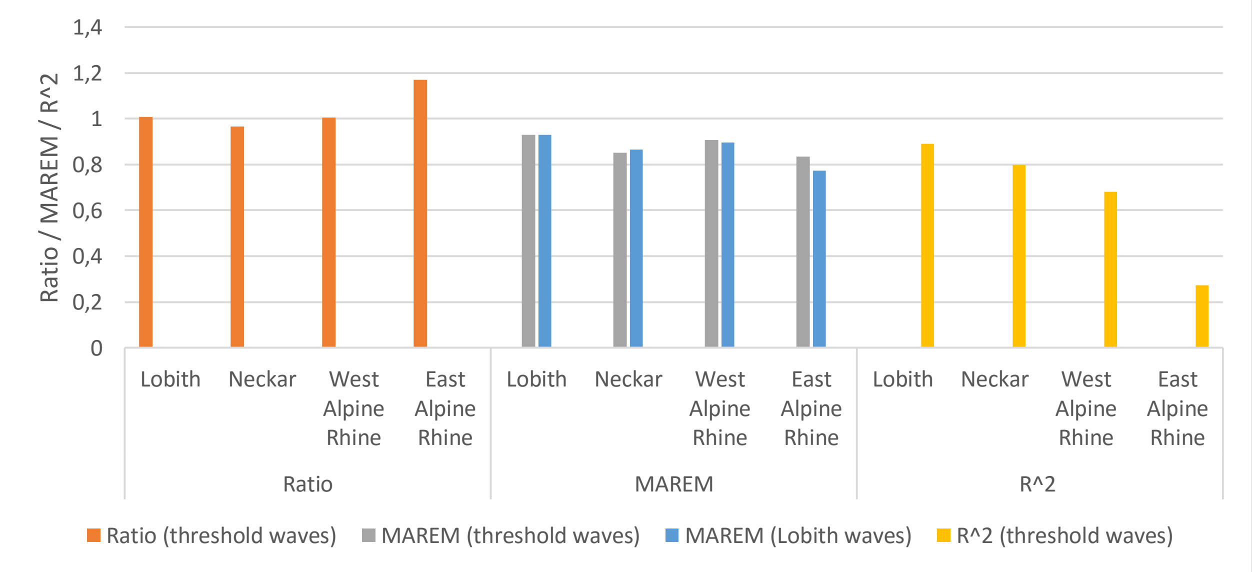

Figure 11 Performance of HBV in simulating the peak discharge of waves from the Rhine upstream of Lobith and three upstream sub-basins (values of 1 indicate good correspondence between simulated and observed waves, Ratio reveals if the model over- or underestimates, MAREM is a quantification of the absolute error and R^2 is the coefficient of determination)

The coefficient of determination shows large differences between the basins. A value of 0,89 is found at Lobith, whereas a value around 0,23 is found in the East Alpine Rhine. Figure 12 shows what this difference looks like. The points show the peak discharge from the wave hits. It can be seen that the peaks at Lobith follow the regression line much better than those from the East Alpine Rhine. So one can assume that if the observed peak discharge at Lobith is low also a low simulated peak will be found, whereas at Rekingen an observed low peak discharge might correspond to a small or high simulated peak discharge.

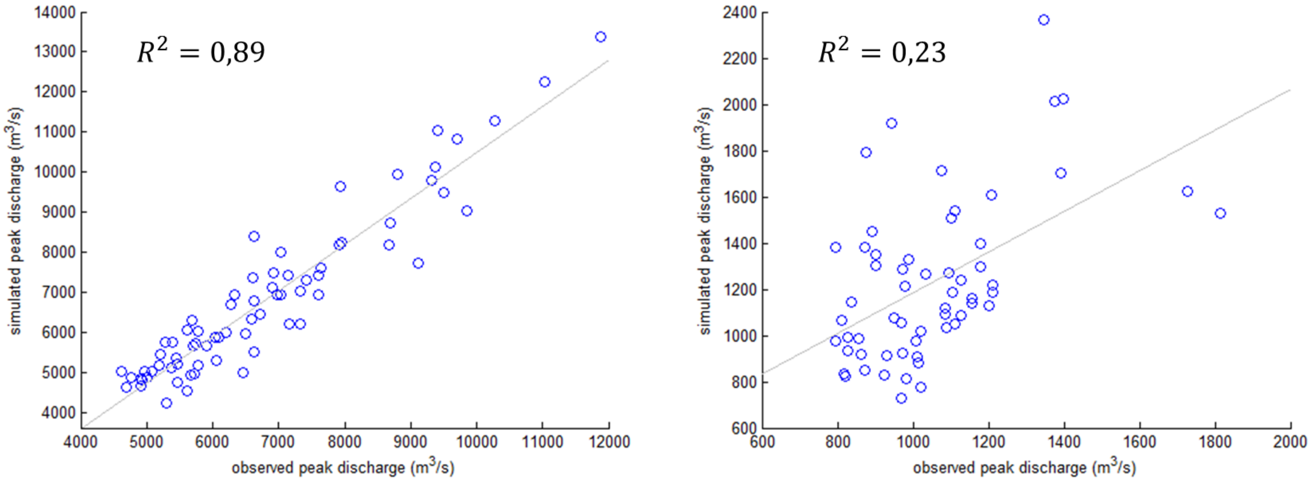

Figure 12 Difference in coefficient of determination between the threshold waves at Lobith (Left) and from the East Alpine Rhine at Rekingen (Right)

4.2 Peak timing

Figure 13 shows the timing differences between the simulated and observed peaks. A MAPTE of zero indicates that the observed and simulated peaks appear at the same moment in time. The timing difference found in waves from the Neckar that contribute to flood waves at Lobith is comparable to the timing difference of the threshold waves from this sub-basin. This effect is less visible in the Alpine region. The peak time difference of the waves that contribute to flood waves at Lobith is larger than the peak time difference of the threshold waves from these two basins. Apparently the difference in peak timing in the flood waves from the Alpine region in the winter months is larger than in the summer months.

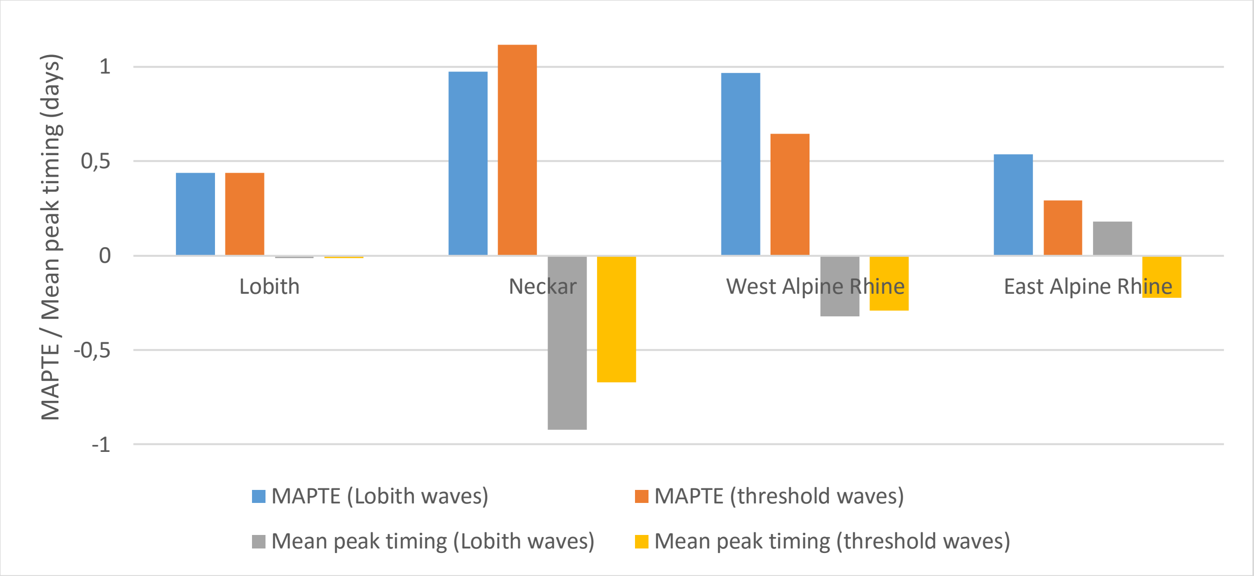

Figure 13 Performance of HBV in simulating the peak timing of flood waves from the Rhine upstream of Lobith and three upstream sub-basins (values of 0 indicate good correspondence between simulated and observed waves, MAPTE is the mean absolute peak time error)

In the Neckar and West Alpine Rhine the absolute time difference between observed and simulated peaks is largest. A reason can be the substantial human regulation of the discharge from the Neckar and Aare (West Alpine Rhine), which makes modelling more difficult. In the West Alpine Rhine the time difference is on average -0,3 days, whereas the MAPTE is almost 1 day. This indicates that the spread in the peak time difference is relatively large. The hydrological simulation of the runoff in the lower Rhine seems to compensate for timing differences further upstream. This can be concluded because the mean peak timing is around zero at Lobith, whereas in upstream basins mainly negative mean peak timings are found. So the calculated runoff from the Lower Rhine basin and/or the hydrological routing used in the HBV model of the Lower Rhine ensure that the time lag of the peaks from inflowing waves is decreased.

4.3 Wave volume

The HBV performance in simulating the flood wave volumes for different sub-basins can be seen in figure 14. Generally the flood wave volumes are underestimated. This same effect is seen when calculating the ratio between the total simulated and observed discharge. For example at Lobith just 92% of the total observed discharge is simulated, see appendix 1, so there is something wrong with the simulated water balance.

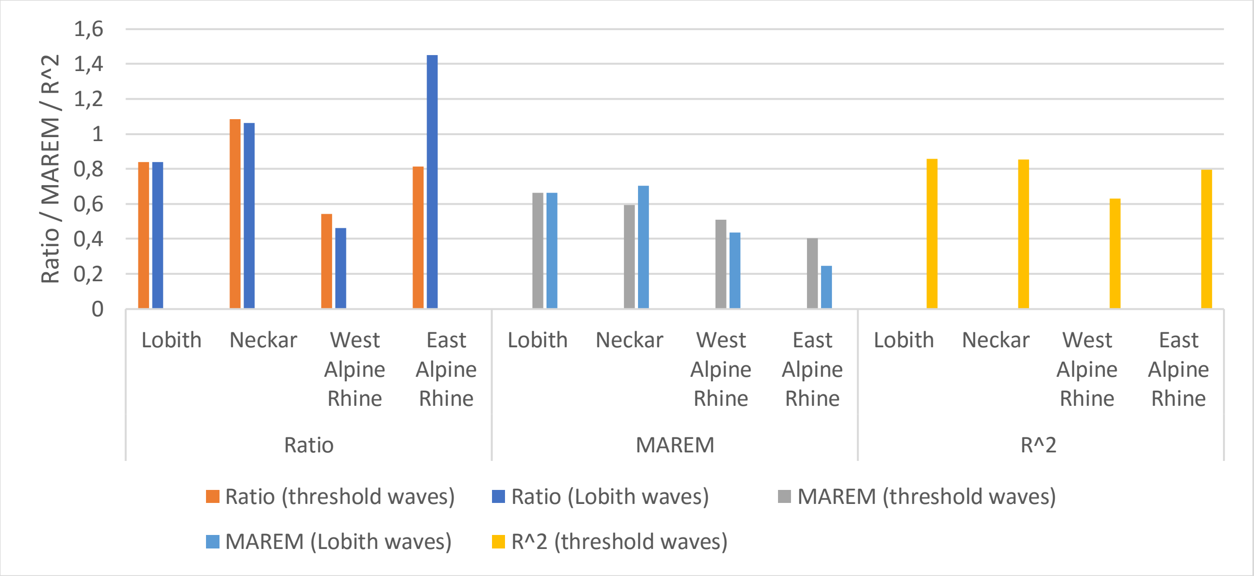

Figure 14 Performance of HBV in simulating the flood wave volumes of waves from the Rhine upstream of Lobith and three upstream sub-basins (values of 1 indicate good correspondence between simulated and observed waves, Ratio reveals if the model over- or underestimates, MAREM is a quantification of the absolute error and R^2 is the coefficient of determination)

Only volumes from flood waves in the Neckar and East Alpine Rhine waves that contribute to Lobith waves are generally too large. The volumes of the waves from the Alpine region that contribute to flood waves at Lobith are simulated worse than the volumes of waves selected by the use of the threshold only. The opposite effect is seen in all other sub-basins. This result might be attributed to the poorer performance in simulating the winter flood waves from the Alpine region. Not much can be said about the contribution of upstream basins to the performance at Lobith, because of the large differences in the performance of HBV in simulating the flood wave volumes from the upstream basins.

4.4 Wave duration

In figure 15 the performance of HBV in simulating the duration of the flood waves can be seen. Generally all flood wave durations are underestimated. Just like with the volume this effect might be attributed to the skill of HBV in simulating the water balance. The discharge is generally simulated to low by the HBV model. Too low flood wave discharges result in too small flood wave durations. Only the duration of waves from the Neckar do not show a structural underestimation when looking to all simulated wave hits from this region. Mainly the duration of waves from the Alpine region sub-basins is simulated poorly compared with those from the other sub-basins. Flood wave durations from the Neckar and Alpine region basins show the largest difference between the duration of Lobith waves compared with the duration of threshold waves. Only in the West Alpine Rhine durations from waves that contribute to flood waves at Lobith are simulated worse than those from the threshold waves.

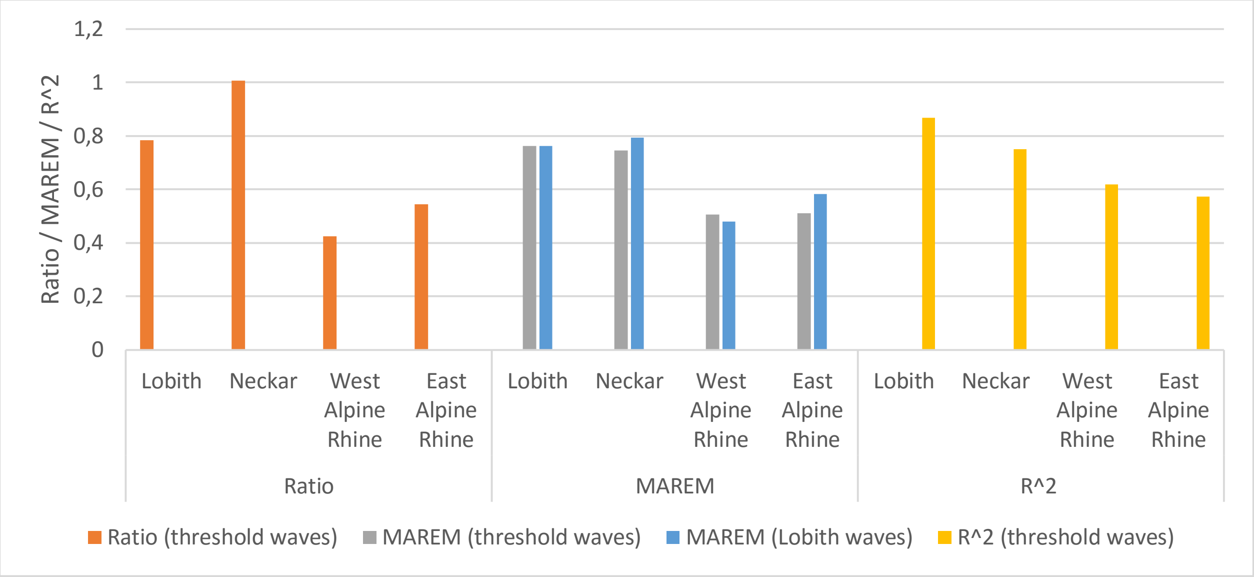

Figure 15 Performance of HBV in simulating the flood wave duration of waves from the Rhine upstream of Lobith and three upstream sub-basins (values of 1 indicate good correspondence between simulated and observed waves, Ratio reveals if the model over- or underestimates, MAREM is a quantification of the absolute error and R^2 is the coefficient of determination)

The performance at Lobith is highly influenced by the performance of upstream flood waves. The relatively good performance in the Moselle and Neckar combined with the relative poor performance in the Main and Alpine region basins, result in an average performance in simulating durations at Lobith. The performance at Lobith is better than at Andernach, suggesting that the flood wave duration from the flow entering the river in the Lower Rhine basin is simulated relatively well.

4.5 Number of waves

In figure 16 the performance of HBV in simulating the number of waves per hydrological year can be seen. In general the number of waves per hydrological year is underestimated. In the East Alpine region a strong overestimation of the number of flood waves can be seen. The low coefficient of determination in the East Alpine Rhine indicates that the number of simulated waves in a hydrological year has little to do with the number of observed waves in this same year.

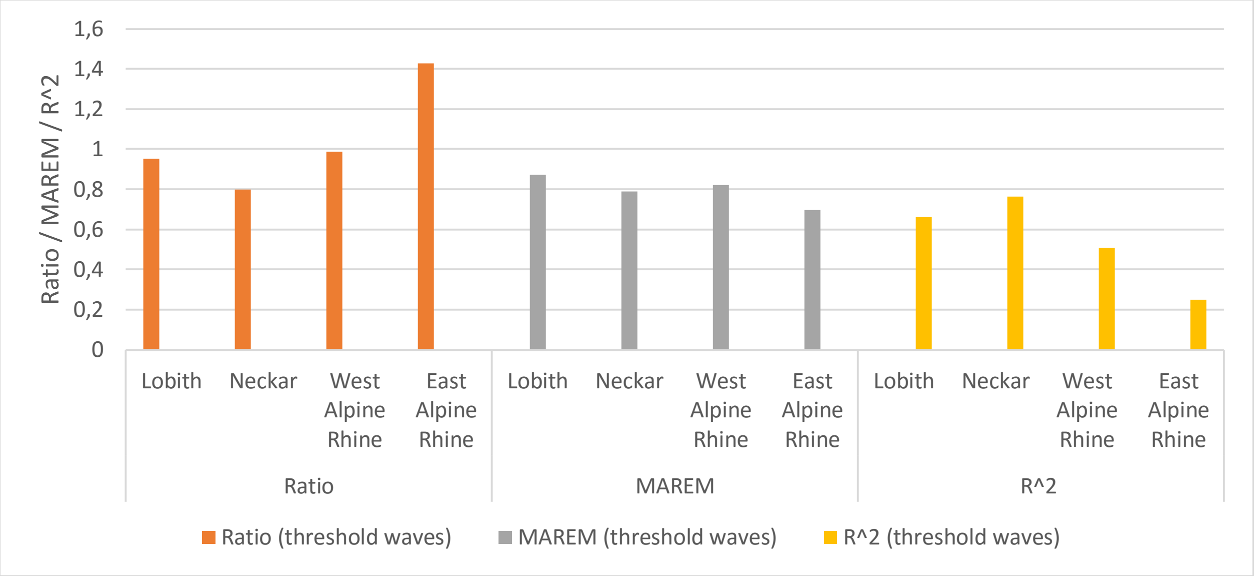

Figure 16 Performance of HBV in simulating number of wave per hydrological year from the Rhine upstream of Lobith and three upstream sub-basins (values of 1 indicate good correspondence between simulated and observed waves, Ratio reveals if the model over- or underestimates, MAREM is a quantification of the absolute error and R^2 is the coefficient of determination)

Figure 17 shows which percentage of the total simulated and observed waves is classified as wave hit. On average around 80% of the total simulated and observed waves are hits. A higher percentage of hits is found for waves from German sub-basins that contribute to flood waves at Lobith. This result is less visible in the Alpine region, in the East Alpine Rhine even the opposite can be seen. One reason is that Lobith waves from the Rhine upstream of Andernach, Moselle, Main and Neckar are on average a bit larger than threshold waves, see appendix 1. If for example an observed discharge is just above the threshold it often will not be found in the simulation, because the model generally underestimates the discharges from most sub-basins, see appendix 1. Therefore the relative large Lobith waves are more often simulated than the relative small threshold waves. In the East Alpine Rhine the percentage of hits is overall smaller than in the other basins. The model simulates more flood waves than have been observed in the East Alpine Rhine, see figure 18, so the percentage of hits is negatively influenced. The high percentage of hits found for Neckar waves that contribute to flood waves at Lobith might be due to the wet basin conditions during these periods. Human operation of weirs will disrupt the observed discharge less in periods with wet (winter) conditions, because surplus water should be discharged, which results in a more natural behaviour of the flow than in dry periods when water is retained to keep the river navigable.

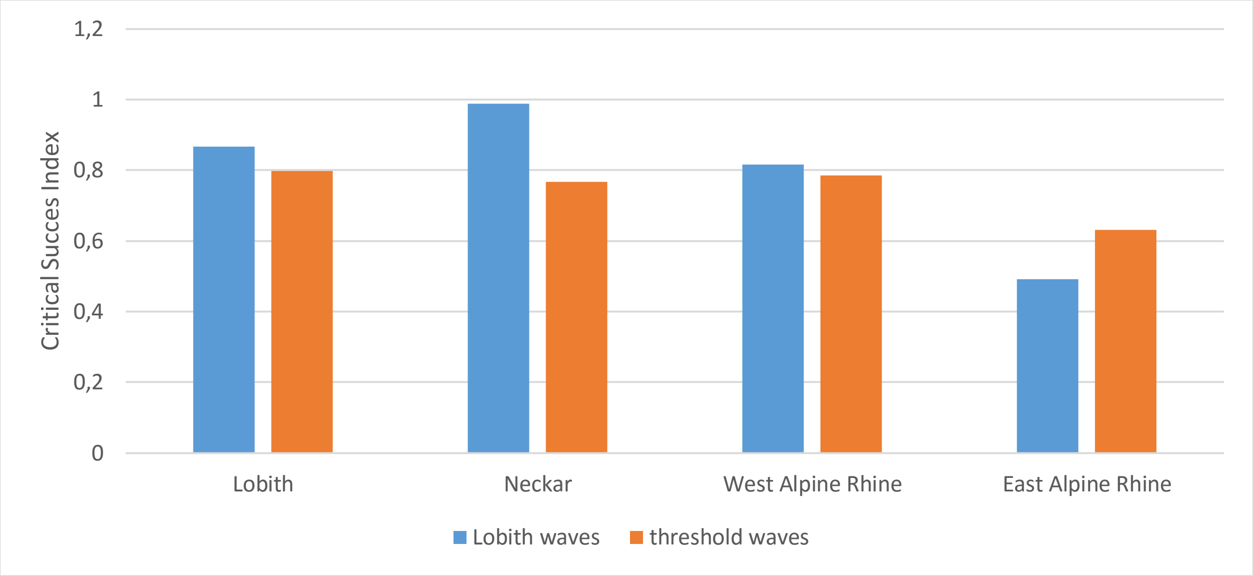

Figure 17 Critical success index of the HBV simulation of waves from the Rhine upstream of Lobith and three upstream sub-basins (values of 1 indicate that all simulated waves are hits)

The results shown in figure 17 only show the relative number of hits. The absolute differences between the numbers of waves found in the sub-basins cannot be derived from these critical success index outcomes. Therefore the absolute numbers of observed waves, wave hits, false alarms and wave misses per sub-basin are displayed in figure 18. It can be seen that the number of misses is generally larger than the number of false alarms in the German sub-basins, whereas the opposite is seen in the Alpine sub-basins. This suggest that discharge above Q5 from the Alpine region is generally overestimated, whereas these high discharges are underestimated in all other sub-basins. Only in the East Alpine Rhine a lot of false alarms (almost all from this region) are found for waves that contribute to flood waves at Lobith. This shows that the discharge from the East Alpine Rhine during floods at Lobith is often overestimated. Most waves are found in the Neckar, this suggest that the flow variation from this basin is higher than from the other sub-basins.

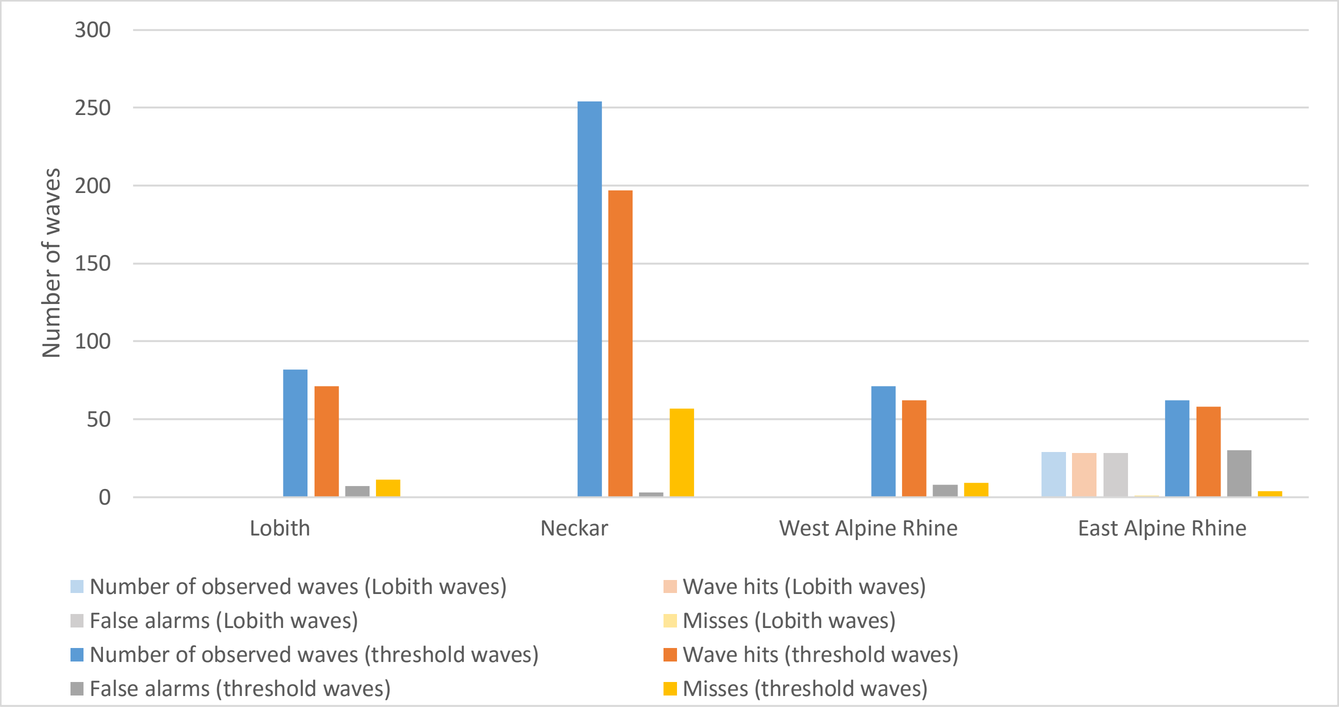

Figure 18 Number of observed waves, false alarms, wave hits and wave misses of the flood waves from the Rhine upstream of Lobith and three upstream sub-basins

4.6 Interpretation of the results

The performance of HBV in simulating the flood wave characteristics from the sub-basins in the German part of the Rhine basin is relatively good. The Neckar is the least well performing German basin. The peak discharges from this basin are underestimated by the model and the simulated peaks are on average approximately 1 day too late. The flood wave volumes and durations of waves from the Neckar are slightly overestimated, this effect is largest for waves that contribute to flood waves at Lobith. The main reason for the detected differences between the simulated and observed flood wave hydrographs is probably the human influence on the discharges from this region, which makes modelling the basin characteristics harder. Errors in the hydrological routing by the use of the Muskingum approach might also explain the detected differences in the simulated flood waves. Namely if the hydrological routing ensures slower runoff, then there is more time for the simulated peaks to attenuate than for the observed ones, resulting in a smaller discharge peak and a larger flood wave duration. The duration increases because the discharges at the boundaries of the simulated wave increases due to the attenuation of the simulated peak discharge. Simulated peaks will also appear later in time than observed ones when the hydrological routing ensures slower runoff. Shaw et al. (2011) discussed that by the use of the Muskingum routing the expected attenuation can be determined, whereas the translation, mainly for long river stretches, can only be accounted for in a non-physical theoretical way. This results in many cases that improving the translation worsen the attenuation. Sub-dividing the river reach in more sections can decrease such problems (Shaw et al., 2011).

The largest differences between observed and simulated flood wave characteristics are found in the discharge series from the two Alpine region basins. Simulated peak discharges from the East Alpine Rhine are mainly overestimated, the peak discharges from the West Alpine Rhine are slightly better simulated with no structural overestimation. The overestimation of the peaks from the East Alpine Rhine might be due to the skill of HBV in simulating the discharge from Lake Constance, which is mainly responsible for the discharge from this basin. An analysis of the performance of HBV in simulating the water levels of the four major lakes in the Alpine region shows that the water level in Lake Constance during flood waves from the East Alpine Rhine differs 0,36 meters from the observed water levels, whereas in the other lakes the water level difference is less, see appendix 4. This suggests that the performance of HBV in simulating the discharge from Lake Constance might be a cause of the overestimation of the peak discharges. Kersbergen (2016) showed that the differences in the lake water levels are probably due to the skill of upstream HBV models in simulating the snow storage. The error in the peak timing of waves that contribute to waves at Lobith is larger than the error in the threshold waves from the Alpine region. Most flood waves found at Lobith appear in the winter season, see appendix 2. So apparently the peak time error of flood waves that mainly appear in the winter season is largest. This might be due to the skill of HBV in simulating the accumulation of snow. The simulated peaks from East Alpine waves that contribute to Lobith floods are generally too early. Precipitation that is allocated to runoff instead of snow storage might result in too early peaks. Also the structural overestimation of Lobith waves and structural underestimation of threshold waves from the East Alpine Rhine can be explained by errors in the allocation of precipitation to snow storage. If precipitation is allocated to runoff instead of snow storage, then in winter there will be too much discharge and in early summer there will be too little. This assumption might also explain why there are so many false alarms detected during periods when flood waves at Lobith are detected. This effect is however not seen in the West Alpine Rhine, the flood wave volumes and durations from this region are generally strongly underestimated.

Another reason for the detected differences in the Alpine region might be the skill of HBV in modelling hydrological processes that are dominant on a small hydrological time scale. The steep hill slopes and shallow soils in the Alpine region influence the hydrological response of the basin. Net precipitation often directly runs off rather than infiltrates into the soil. This behaviour has a typical time scale of hours, instead of days (Hegnauer & Verseveld, 2013) and can thus not be correctly simulated with the daily HBV model. Also the operation of hydroelectric power facilities constructed in the Aare and Alpine Rhine and water abstraction might be a cause for the detected differences. Another reason for detected differences between observed and simulated flood wave characteristics from the Alpine region might be the quality of the precipitation data. Especially measuring the amount of snow is tricky because of relatively large spatial differences in snow quantities due to wind-induced losses. Therefore an undercatch (both over- or underestimation) of precipitation up to 10% might be expected in the Alpine region (Frei & Schär, 1998).

5 Results of HBV, HBV-WG and WG evaluation

In this chapter a description is given of the results of the performance assessment regarding the performance of HBV, HBV-WG and WG in simulating the flood wave characteristics of the different sub-basins in the Rhine basin. The performance assessment described in chapter 4 only focuses on the skill of HBV in simulating the characteristics of the flood wave hits. The results described in this chapter are calculated on the basis of all observed and simulated flood waves. The flood wave characteristics peak discharge, wave volume, wave duration and number of flood waves per hydrological year obtained from observed, HBV simulated and HBV-WG simulated discharges series are used in the assessment. P-values which show the equality of means and variances and cumulative distribution functions (CDFs) are used to compare the observed characteristics with the simulated ones. First in paragraph 5.1 the observed and simulated relative contributions to flood waves at Lobith are discussed, to assess if HBV and HBV-WG are able to simulate contributions comparable to the observed contributions. In paragraphs 5.2, 5.3, 5.4 and 5.5 the results regarding the peak discharge, wave volume, wave duration and number of waves per year are discussed respectively. Only the CDFs calculated for Lobith, the two alpine region basins and the worst performing upstream German basin (Moselle, Main or Neckar) are discussed. Finally in paragraph 5.6 the interpretation of the results is given.

5.1 Relative contributions to flood waves at Lobith

The pie charts in figure 19 show the relative sub-basin contributions to the discharge at Lobith during flood events. It can be seen that there are only some minor differences between the simulated and observed contributions. The up to 3 % difference between the contributions from the Lower Rhine cannot directly be attributed to flaws in the models or used data. This is because the contributions from the Lower Rhine and Middle Rhine are calculated only by abstracting the total volumetric contributions of all other sub-basins from the total flood wave volume at Lobith or Andernach, so the required residual contributions are in fact assumed to come from the Lower and Middle Rhine basins. Therefore flaws from upstream basins accumulate in the contribution of the Lower Rhine and Middle Rhine. The contribution from the Middle Rhine is always underestimated by HBV and HBV-WG, whereas the contributions from the Main and Lower Rhine are always overestimated by HBV and HBV-WG. The relative contribution of the Moselle simulated with HBV-WG is underestimated due to the generated weather series, because no difference is detected between the observed and HBV simulated contribution. From figure 19 it can be concluded that the relative contributions from upstream sub-basins to flood waves at Lobith is represented well by the HBV and HBV-WG simulations.

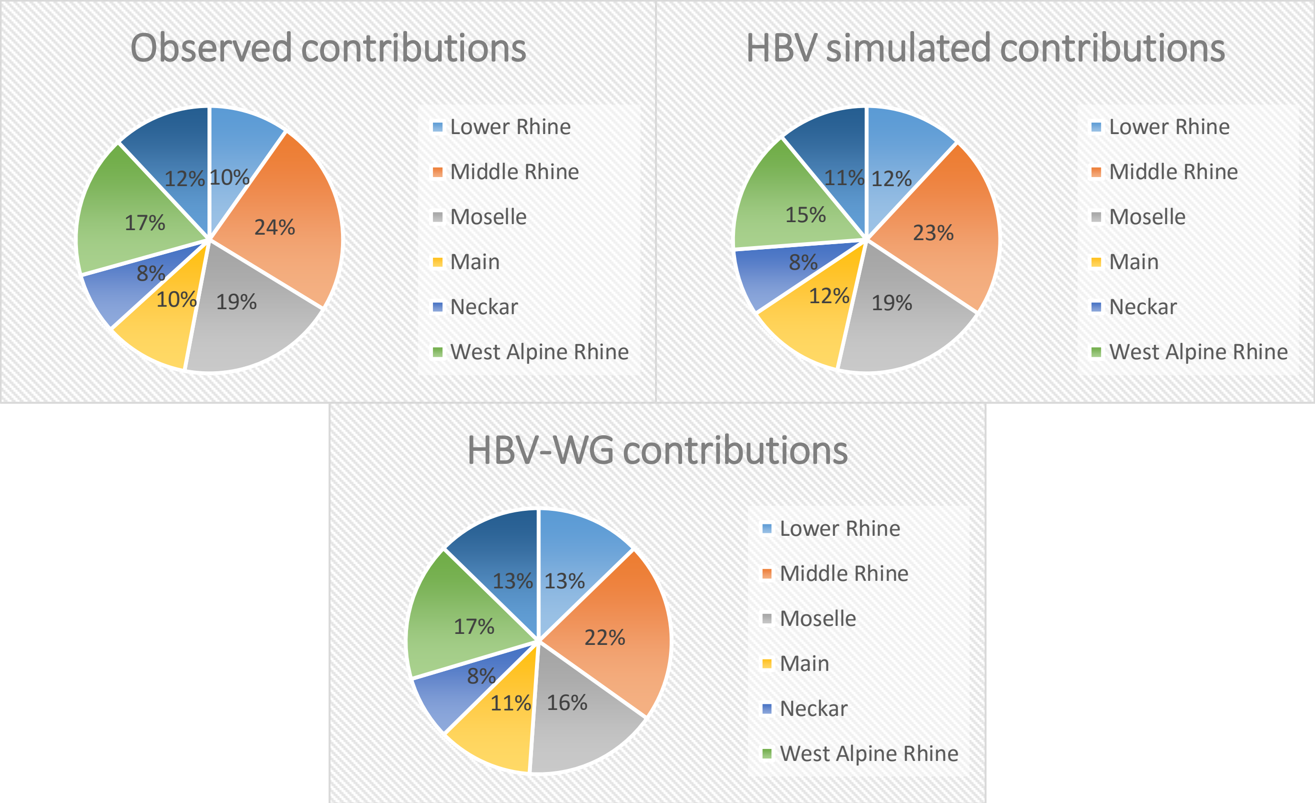

Figure 19 Relative observed and simulated sub-basin contributions to flood waves at Lobith

5.2 Peak discharge

Table 6 shows the results of the statistical tests conducted to assess the equality of the observed and simulated peak discharge means and variances. The table shows the p-values calculated with a t-test to assess if the observed and simulated mean peak discharges are equal and a F-test to assess if the variance of the observed and simulated peak discharge are equal. If the p-value is below 0,05 then the hypothesis of equality is rejected, meaning that there is a significant difference between the observed and simulated peak discharge means or variances. The overall picture is that the peak discharges are simulated relatively well. The variance of the peak discharges from the Rhine at Andernach simulated with HBV-WG differs significantly from the observations. This suggest that HBV fed with synthetic weather series enlarges peak spread differences that might be attributed to the skill of the HBV simulations, because the p-value of the equality of observed peak discharge variance with HBV peak discharge variance is much smaller than the p-value of the equality of HBV peak discharge variance with HBV-WG peak discharge variance. Very small p-values are found for peak discharges from the East Alpine region. Because water from this region is also discharged through the Middle Rhine, the significant difference in peak variances at Andernach might be due to the performance of HBV in simulating the peak discharges from the East Alpine Rhine. The skill of HBV from the East Alpine Rhine is obviously not good, this is concluded because the variances and means of peak discharges simulated by HBV and HBV-WG differ significantly from the observed ones. These differences can be attributed to the performance of the HBV model, because no significant differences are found between HBV and HBV-WG peak discharges. Only in the Moselle the WG seems to negatively influence the skill in simulating the mean peak discharges.

Table 6 Results of the t-test and F-test used to assess equality of the mean and variance calculated from the observed and simulated peak discharges for all sub-basins (Green = no significant difference, Red = significant difference) (Significance level = 0,05)

p-values peak discharge	Observed with HBV simulated		Observed with HBV-WG simulated		HBV-WG simulated with HBV simulated	
	mean	variance	mean	variance	mean	variance
Rhine at Lobith	0,6422	0,1154	0,6303	0,0631	0,8896	0,7704
Rhine at Andernach	0,2417	0,1550	0,3541	0,0326	0,5509	0,8671
Moselle at Cochem	0,0967	0,3813	0,7699	0,6328	0,0495	0,4553
Main at Frankfurt (from 1963 instead of 1951)	0,7791	0,2530	0,8504	0,6587	0,5719	0,2407
Neckar at Rockenau	0,2249	0,3005	0,3580	0,4218	0,4274	0,5080
West Alpine Rhine at Untersiggenthal	0,3203	0,1455	0,7921	0,1441	0,1077	0,5854
East Alpine Rhine at Rekingen	0,0068	0,0001	0,0006	0,0000	0,6154	0,7914

Figure 20 shows the CDFs calculated for the peak discharges from the Rhine upstream of Lobith, Moselle, West Alpine Rhine and East Alpine Rhine. At Lobith the peak discharges are simulated well by both HBV and HBV-WG, the 11 highest peaks are slightly overestimated, whereas the 50% lowest peaks are slightly underestimated. This is expected, because within the HBV calibration special attention is drawn to the simulations of extreme peak discharges at Lobith (Winsemius et al., 2013). The Moselle CDFs show that the HBV simulation generally overestimate the peak discharges. This overestimation is less present in the HBV-WG simulations, so the input of generated weather series in HBV results in peak discharges closer to those found in the observations. This shows that the WG coincidentally compensates HBV errors. It can therefore be concluded that in both HBV and WG errors are detected. The peak discharges from the West Alpine Rhine are simulated relatively well. The East Alpine Rhine peak discharges are simulated worst. The 80% highest simulated peaks are substantially larger than the 80% highest observed peaks. The larger peaks are overestimated more than the smaller peak discharges. No substantial differences between the peaks simulated with HBV and HBV-WG are found, which suggest that HBV is responsible for the found differences.

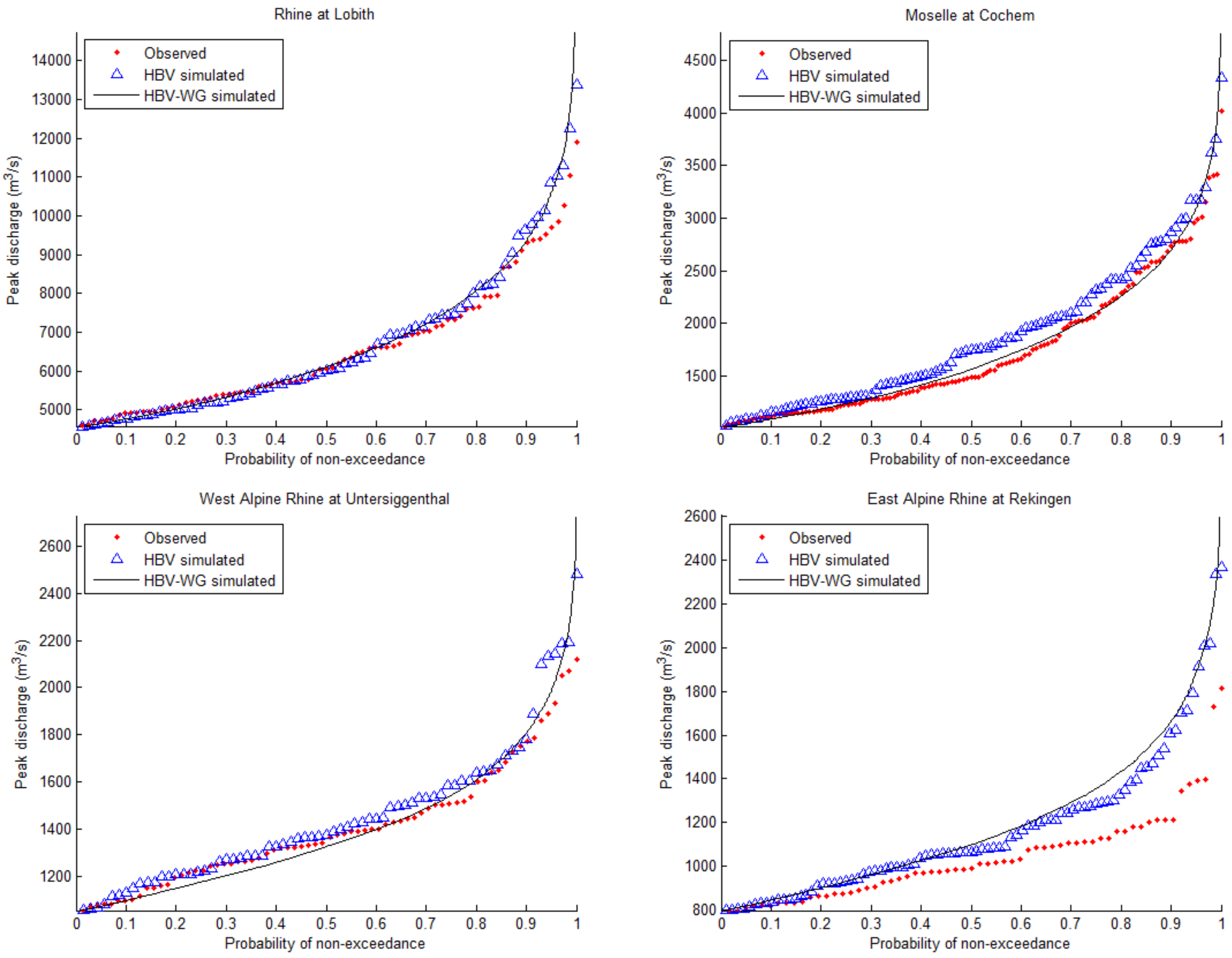

Figure 20 CDFs of the observed, HBV simulated and HBV-WG simulated peak discharges from the Rhine basin upstream of Lobith, Moselle, West Alpine Rhine and East Alpine Rhine

Spread in peak discharges at Lobith due to HBV-WG simulations

The graphs from figure 20 show the HBV-WG CDF for the 10.000 year period. Assessing equally long 56 year periods from this 10.000 year simulation should give comparable results as those from the observed and HBV simulated peak discharges. To assess this, 50 randomly picked 56 year periods are drawn from the 10.000 year HBV-WG simulation from the Rhine at Lobith. In figure 21 the CDFs of the observed, HBV simulated and HBV-WG 10.000 year simulated peak discharges compared with the CDFs of the peak discharges obtained from 50 different 56 years randomly selected from the 10.000 year HBV-WG simulations can be seen. It can be seen that the lowest 30% of the peak discharges are comparable. From this point on the spread in peak discharges obtained from the random 56 year periods increases. The spread is largest for the largest peak discharges. Both underestimation and overestimation is equally often present. The 10.000 year simulation CDF is more or less the average from all random 56 year CDFs, which is expected. Some CDFs from the randomly picked 56 years differ substantially from the CDFs of the observed and HBV simulated peak discharges. This shows that not in all 56 year HBV-WG simulations the peak discharge are comparable to the measurements. It seems that the WG sometimes simulates 56 year weather conditions that are more extreme and sometimes less extreme than observed. Single flood waves that are more or less extreme are expected, but whole 56 year periods that are more or less extreme not. A reason for this might be a bias in the WG to the initial weather conditions. The consistency in the WG is probably a bit larger than in reality, which means that the probability of selecting comparable weather as the previous day is larger in the WG than in the observations.

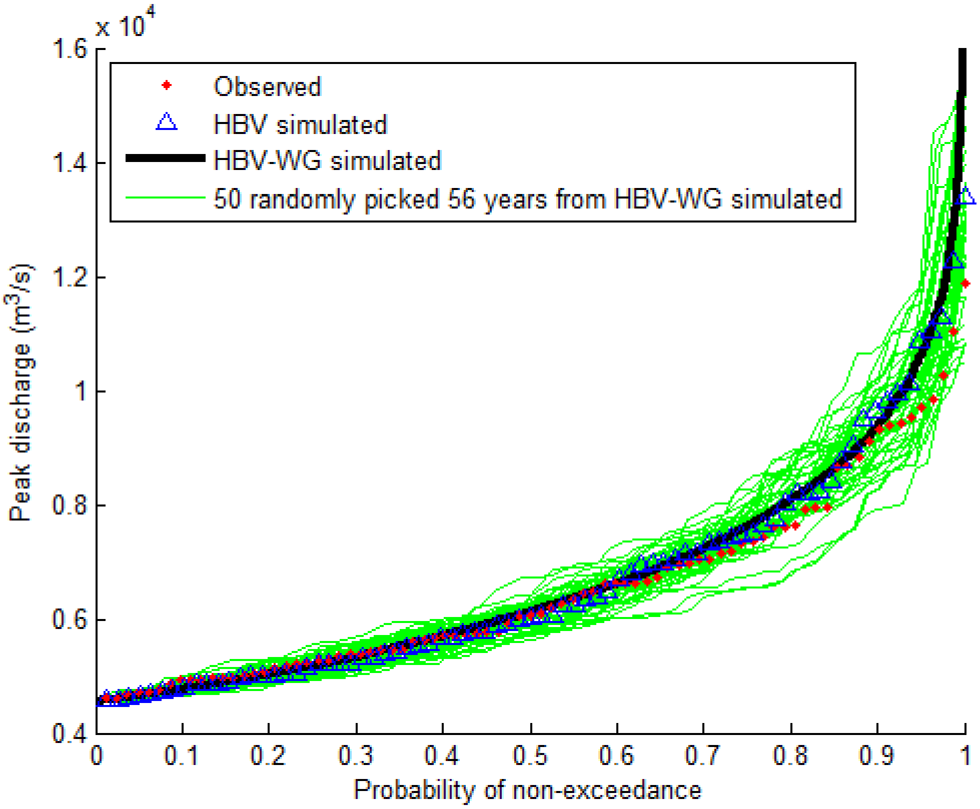

Figure 21 CDFs of the observed, HBV and HBV-WG simulated peak discharge of the Rhine at Lobith compared to the CDFs of the peak discharges at Lobith obtained from 50 random 56 year samples from the 10.000 year HBV-WG simulation

5.3 Wave volume

Table 7 shows the results of the statistical tests conducted to assess the equality of the observed and simulated flood wave volume means and variances. The mean volume simulated with HBV-WG from the Neckar differs significantly from the observed mean flood wave volume. The p-value of the mean observed volume compared with the mean HBV simulated volume is already close to 0,05, which indicates that the difference found is mainly caused by the HBV model and worsened by the WG. A lot of significant differences are found between simulated and observed flood waves from both Alpine basins. It is clear to see that the HBV and HBV-WG simulated volumes from the West Alpine Rhine compared with the observed ones differ significantly from each other, the analysis of HBV compared with HBV-WG does not show the same result, which reveals that detected differences are mainly due to the performance of HBV. It is furthermore remarkable that the spread between the volumes simulated with HBV differs significantly from the spread of volumes simulated with HBV-WG, whereas the p-value of the mean is remarkably high. This suggests that the WG is responsible for mainly differences in spread, which shows that the WG is capable of creating other extremes without changing the mean. Both HBV and WG are responsible for differences in East Alpine flood wave volumes. Feeding HBV with synthetic weather series results in smaller p-values. However no significant differences are found between HBV and HBV-WG, so the WG contribution to errors is smaller than that of HBV.

Table 7 Results of the t-test and F-test used to assess equality of the mean and variance calculated from the observed and simulated flood wave volumes for all sub-basins (Green = no significant difference, Red = significant difference) (Significance level = 0,05)

p-values volume	Observed with HBV simulated		Observed with HBV-WG simulated		HBV-WG simulated with HBV simulated	
	mean	variance	mean	variance	mean	variance
Rhine at Lobith	0,5030	0,4196	0,4869	0,2922	0,8876	0,0393
Rhine at Andernach	0,4974	0,0414	0,3270	0,6763	0,9627	0,0227
Moselle at Cochem	0,4416	0,8796	0,6622	0,4994	0,1463	0,4243
Main at Frankfurt (from 1963 instead of 1951)	0,6745	0,6389	0,2203	0,1890	0,6053	0,5894
Neckar at Rockenau	0,0510	0,0501	0,0212	0,0801	0,6090	0,3061
West Alpine Rhine at Untersiggenthal	0,0201	0,0000	0,0000	0,0000	0,9645	0,0223
East Alpine Rhine at Rekingen	0,1211	0,0000	0,0018	0,0000	0,8651	0,1961

The CDFs shown in figure 22 are calculated from the observed, HBV and HBV-WG simulated flood wave volumes from the Rhine at Lobith, Neckar, West Alpine Rhine and East Alpine Rhine. Only the two largest flood wave volumes at Lobith are slightly overestimated, the rest of the highest 25% of the volumes is slightly underestimated. Both HBV and HBV-WG slightly overestimate the flood wave volumes from the Neckar. The largest differences are found in the West Alpine Rhine, the 35% largest volumes are substantially underestimated by HBV and HBV-WG. This reveals that the significant difference in variance can be attributed to generally to small simulated extreme flood volumes. The largest observed volume is much bigger than the largest volume simulated with HBV. The hydrograph of the wave with this extreme volume can be seen in figure 23 and shows that most observed discharges during this period are remarkably larger than the HBV simulated ones. The differences between the East Alpine Rhine CDFs are comparable to those of the West Alpine Rhine ones. So mainly the HBV model is responsible for the underestimation of the extreme flood wave volumes from the Alpine region. A reason might be the performance of HBV in allocating precipitation to snow storage, probably too little water is allocated to the snow storage, so that too little melt water contributes to flood waves resulting from snow melt. Too much water might be allocated to flood waves resulting from precipitation during cold periods, because precipitation is allocated to runoff instead of snow storage.

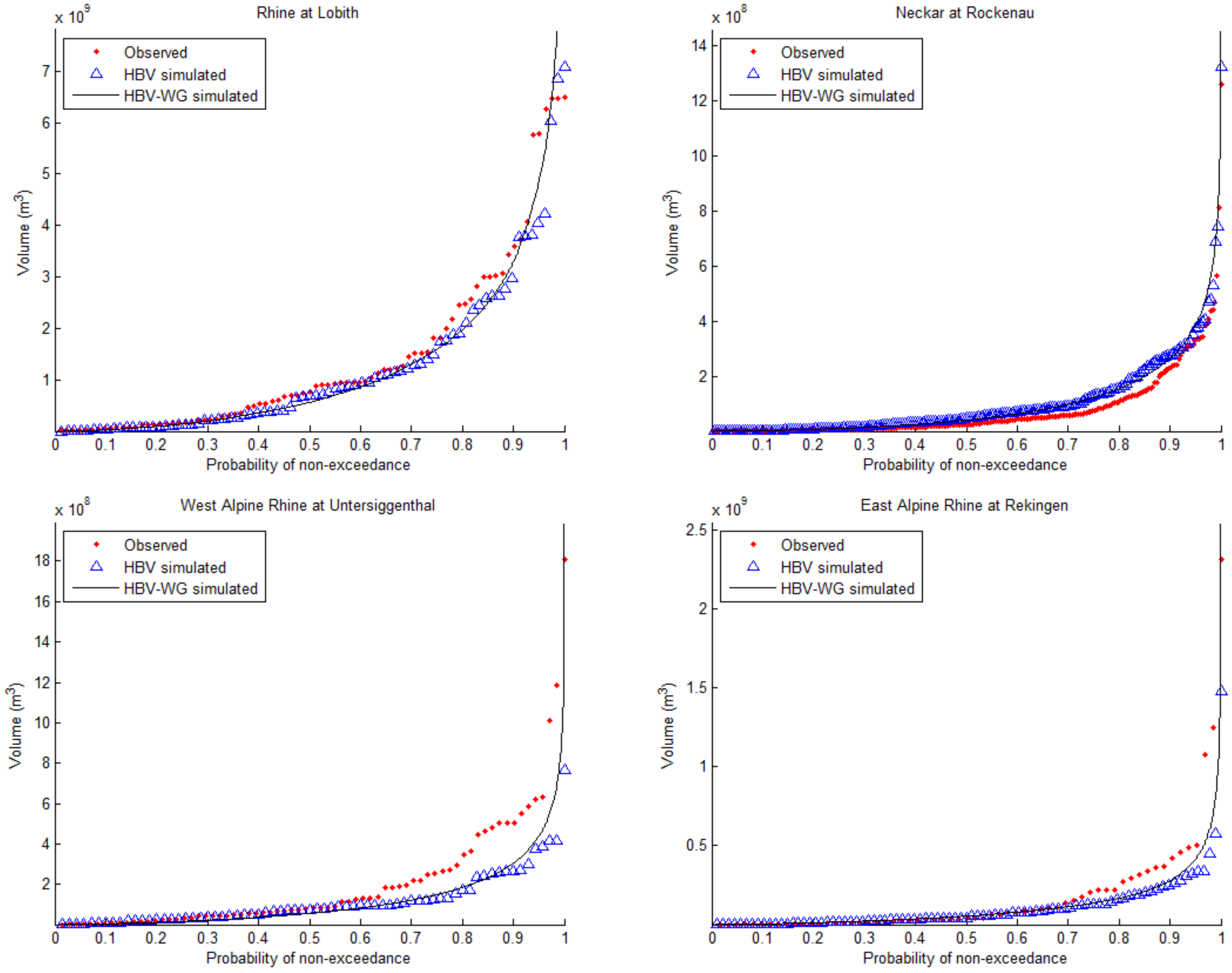

Figure 22 CDFs of the observed, HBV simulated and HBV-WG simulated wave volumes from the Rhine basin upstream of Lobith, Neckar, West Alpine Rhine and East Alpine Rhine

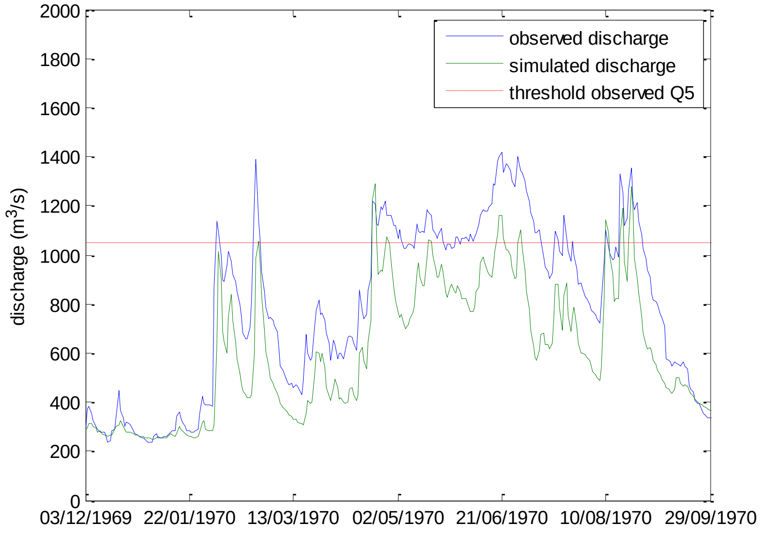

Figure 23 Hydrograph of the flood wave with the most extreme volume and duration detected in the West Alpine Rhine

5.4 Wave duration

Table 8 shows that the mean and variance of observed and simulated flood wave durations from many sub-basins differs significantly. Apparently the flood wave characteristic that is simulated worst is the duration. In the flood wave durations from waves at Lobith and Andernach significant differences are detected. These can mainly be attributed to the HBV model, however it says little about which sub-basin might be responsible. Probably the HBV models of the poor performing Alpine region are partly responsible for the differences found further downstream. HBV is mainly responsible for the differences in the flood wave durations from the Alpine basins. The HBV-WG flood wave durations from the Neckar only differ significantly from the observations. This suggests that the WG negatively influences the performance of HBV-WG in simulating durations. This influence is however mild, because no significant differences are found between durations simulated with HBV and those simulated with HBV-WG.

Table 8 Results of the t-test and F-test used to assess equality of the mean and variance calculated from the observed and simulated flood wave duration for all sub-basins (Green = no significant difference, Red = significant difference) (Significance level = 0,05)

p-values duration	Observed with HBV simulated		Observed with HBV-WG simulated		HBV-WG simulated with HBV simulated	
	mean	variance	mean	variance	mean	variance
Rhine at Lobith	0,1049	0,0323	0,0046	0,0689	0,6676	0,2067
Rhine at Andernach	0,0327	0,0001	0,0000	0,0000	0,6583	0,1293
Moselle at Cochem	0,8773	0,0709	0,1687	0,0283	0,1534	0,5683
Main at Frankfurt (from 1963 instead of 1951)	0,9265	0,3163	0,7197	0,0377	0,8252	0,5082
Neckar at Rockenau	0,0532	0,3718	0,0080	0,0125	0,6562	0,3369
West Alpine Rhine at Untersiggenthal	0,0001	0,0000	0,0000	0,0000	0,9124	0,0424
East Alpine Rhine at Rekingen	0,0002	0,0000	0,0000	0,0000	0,7154	0,1180

The graphs presented in figure 24 show the CDFs of the flood wave durations of flood waves from the Rhine at Lobith, Neckar, West Alpine Rhine and East Alpine Rhine. For all four sub-basins no substantial differences are detected between with HBV and HBV-WG simulated flood wave durations, which shows that HBV is mainly responsible for detected differences. The 70% largest durations of flood waves at Lobith are slightly underestimated. HBV structurally underestimates the overall discharge at Lobith, see appendix 1, this explains why the durations are often underestimated. The number of waves from the Neckar with a duration of 1 day is substantially underestimated by the HBV and HBV-WG simulations, whereas durations of 5 days or more are generally slightly overestimated. This suggests that the HBV model is responsible for simulating too long flood waves. The hydrograph shown in figure 23 shows the longest observed flood wave from the West Alpine Rhine. The CDFs of the flood wave durations from this basin reveals that even in the 10.000 year HBV-WG simulation no such an extreme flood wave duration is found. The 40% shortest durations are simulated well, whereas the 60% longest flood wave durations are extremely underestimated by HBV and HBV-WG, this is caused by the overall underestimation of observed flows. The CDFs of the durations of flood waves from the East Alpine Rhine are comparable to those of the West Alpine Rhine. The 50% longest durations are considerably underestimated by HBV and HBV-WG.

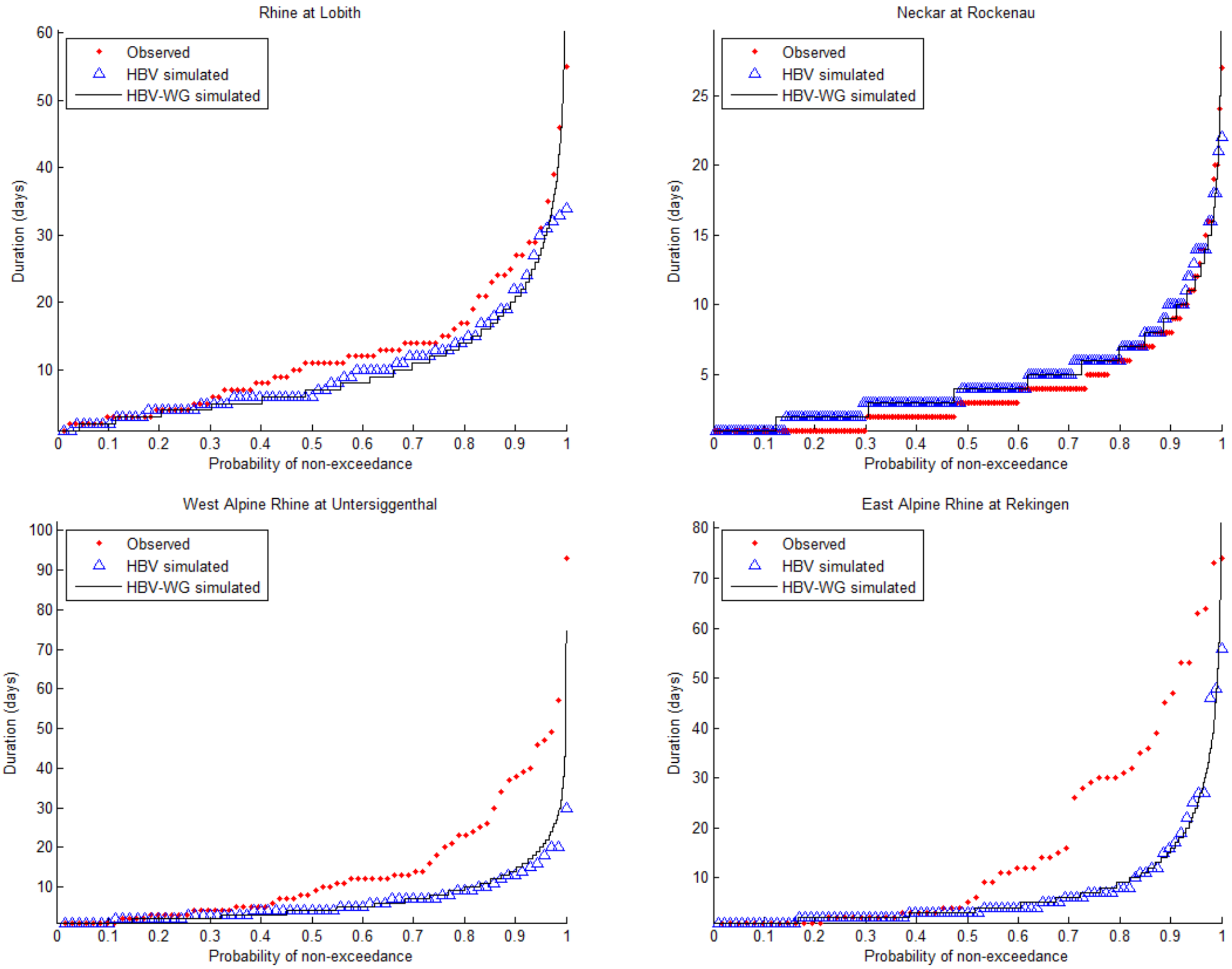

Figure 24 CDFs of the observed, HBV simulated and HBV-WG simulated wave durations from the Rhine basin upstream of Lobith, Neckar, West Alpine Rhine and East Alpine Rhine

5.5 Number of waves

Table 9 shows the results of the statistical tests done to assess the equality of the mean and variance of the observed and simulated number of flood waves per hydrological year. HBV-WG is mainly responsible for the detected differences. HBV or WG are less frequently responsible for differences, because just a little significant differences are detected between HBV and observations and HBV-WG simulations with HBV simulations. So mainly if both are assessed significant differences are found, so both HBV and WG are responsible for a portion of the difference. HBV and WG often strengthened each other's errors instead of compensate each other's errors, this can be concluded because a lot more significant differences are found when comparing HBV-WG with observations than when HBV and WG are compared with respectively observations and HBV simulations.

Table 9 Results of the t-test and F-test used to assess equality of the mean and variance calculated from the observed and simulated number of flood waves of all sub-basins (Green = no significant difference, Red = significant difference) (Significance level = 0,05)

p-values number of annual flood events	Observed with HBV simulated		Observed with HBV-WG simulated		HBV-WG simulated with HBV simulated	
	mean	variance	mean	variance	mean	variance
Rhine at Lobith	0,7123	0,6395	0,0815	0,5775	0,2390	0,2092
Rhine at Andernach	0,3269	0,6173	0,0204	0,1531	0,4253	0,4986
Moselle at Cochem	0,0525	0,7647	0,0005	0,0667	0,7789	0,1713
Main at Frankfurt (from 1963 instead of 1951)	0,3821	0,7708	0,0568	0,0093	0,6986	0,0334
Neckar at Rockenau	0,0663	0,2036	0,0001	0,0003	0,4751	0,1370
West Alpine Rhine at Untersiggenthal	0,9111	0,3707	0,3095	0,7180	0,2392	0,3552
East Alpine Rhine at Rekingen	0,0010	0,0887	0,0022	0,0238	0,2319	0,9802

In figure 25 the CDFs calculated from the observed and simulated number of flood waves per hydrological year from the Rhine at Lobith, Neckar, West Alpine Rhine and East Alpine Rhine can be seen. At Lobith only some minor differences between the numbers of waves per hydrological year are found, HBV and HBV-WG simulate more years with 4 or 5 waves than are found in the observations. The simulated number of waves per hydrological year from the Neckar is often smaller than the observed one. Years with 2 waves are overestimated and years with 6 waves are underestimated by HBV and HBV-WG. So years with large numbers of waves are mainly underestimated, which might be due to the fact that the routing by HBV is not good. Attenuation of high discharges due to the hydrological routing might result in fewer simulated flood waves, because the discharge may drop under the threshold. No major differences are found in the number of flood waves from the West Alpine Rhine. The number of hydrological years with only 1 flood wave from the East Alpine Rhine is notably underestimated by HBV and HBV-WG. An overestimation of the number of hydrological years with 2 or 3 simulated waves compensates for this effect.

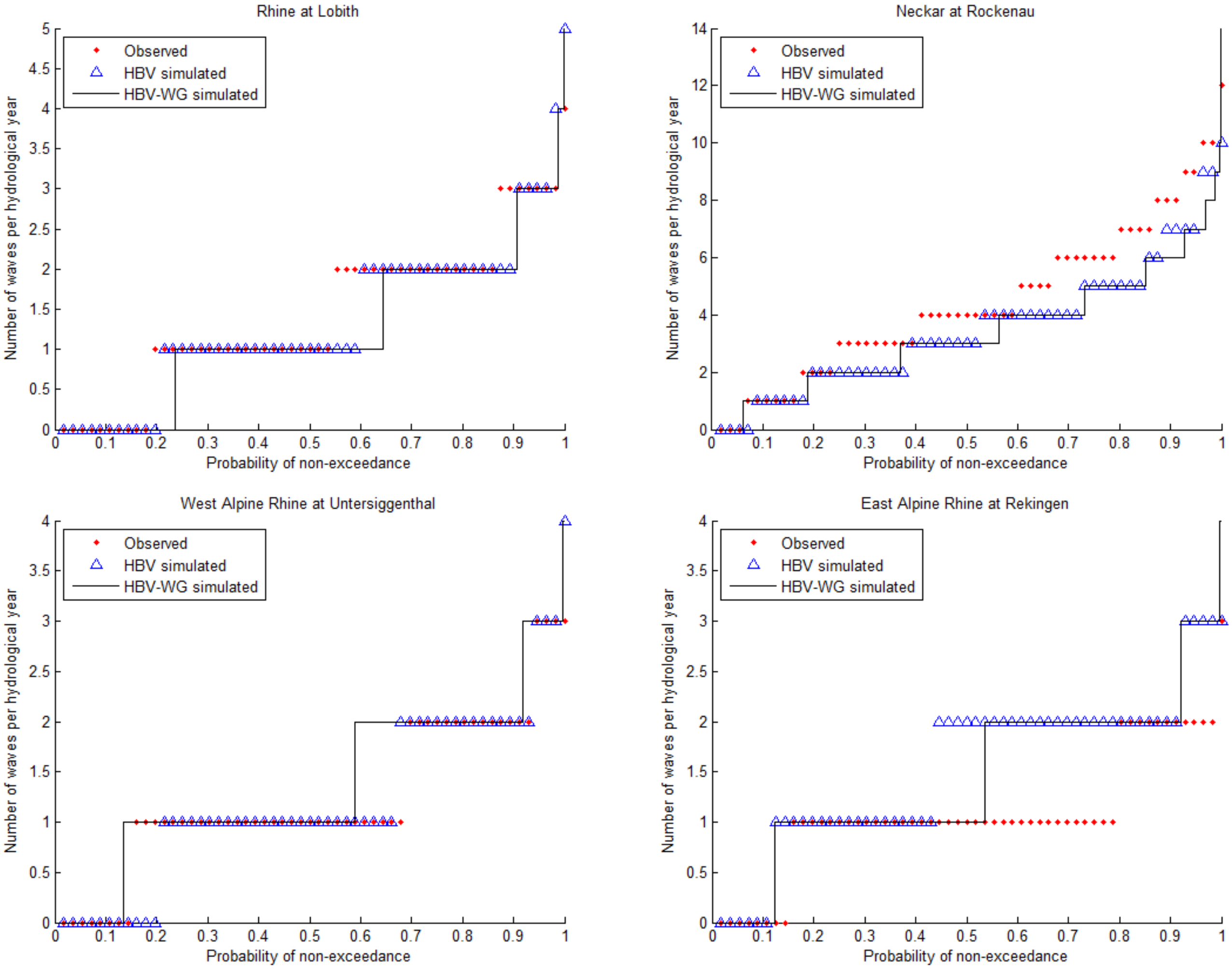

Figure 25 CDFs of the observed, HBV simulated and HBV-WG simulated number of flood waves per hydrological year from the Rhine basin upstream of Lobith, Neckar, West Alpine Rhine and East Alpine Rhine (The CDF points or lines represent the years)

5.6 Interpretation of the results

No major differences in the observed, HBV simulated and HBV-WG simulated relative contributions of the different sub-basins to flood waves at Lobith are detected. So on average the contributions are simulated well. The characteristics of flood waves at Lobith, simulated with HBV and HBV-WG, are both comparable with the characteristics of observed flood waves. The minor detected differences can generally not be attributed to the performance of HBV in simulating the runoff from the Lower Rhine, because most discharge originates from upstream sub-basins.

The flood waves from the East Alpine Rhine are simulated worst. Substantial differences are found for each of the flood wave characteristics. The HBV model is responsible for the largest differences. The peak discharges and number of flood waves per year are overestimated, whereas the volume and duration are underestimated. The underestimation of the volume and duration is probably due to the fact that not enough precipitation is allocated to the snow storage. If too little water is stored in snow, than the discharge driven by snow melt will be underestimated. Also the overestimation of the number of flood waves might be due to the errors in the allocation of precipitation to snow storage. If precipitation is allocated to runoff instead of snow storage than the discharge becomes higher, so extra flood waves might appear. Overestimation of the peak discharge might be due to the simulated snow melting rate. If snow is melted to fast than peak discharges will increase. Another possible reason might be the skill of HBV in simulating the damping of the discharge due to Lake Constance. Most fast runoff from the Alpine mountains in the East Alpine Rhine basin flows into Lake Constance and is discharged in the Rhine subsequently. If HBV is not able to transform the relatively fast fluctuating inflow to relatively steady outflow, overestimation of peak discharges might be the result. The analysis of the skill of HBV in simulating the water levels, included in appendix 4, does not reject this assumption, because for example during flood waves the water levels are on average up to 0,36 meters off.

In most circumstances the WG ensures for weather conditions that lead to comparable discharge as those simulated for the 56 year reference period. Feeding those synthetic weather series into the HBV model often increases already detected differences with the observations. The reason for this result is that the WG is in fact an extra source of uncertainty. It is therefore logical that the performance of HBV-WG is generally a little worse than the performance of HBV only. It should be noticed that errors in the WG might also lead to compensation of errors in HBV. This is not desirable, because the compensation leads to good results for the wrong reasons.

The performance of the WG in simulating weather conditions that lead to flood waves comparable to flood waves from the 56 year reference period is relatively good. Only some small difference have been detected, for example the peak discharges of waves from the Moselle become smaller due to the WG. An explanation might be the constraints used in the weather generator. The Moselle and Meuse basins both discharge water from the Ardennes. The weather conditions and basins characteristics are comparable. In the Meuse basin a different feature vector as in the Rhine is used for finding the nearest neighbours. One reason for using a different feature vector for the Meuse is the data availability. Only 7 precipitation stations and 2 temperature stations were available for the Meuse basin, whereas for the Rhine basin an interpolated grid of 5 km * 5 km has been used. The main difference in this feature vector is that in the Rhine the fraction of sub-basins with precipitation >0,3 mm is incorporated, whereas for the Meuse the average standardized daily precipitation of 7 rainfall stations, averaged over the four preceding days is used (Hegnauer et al., 2014). So in the Meuse the intensity of 4-day rainfall is used to select the nearest neighbours, whereas in the Rhine only the fraction of basins with precipitation is used. Not incorporating this 4-day rainfall in the Moselle basin might be a reason for the underestimation found. Furthermore Leander and Buishand (2004) used a 121 day moving window instead of a 61 day moving window (used for whole Rhine basin upstream of Lobith) for the Meuse basin to improve the simulation of extreme multi-day precipitation amounts. Using this criterion also for the Moselle might improve the performance of the WG. Another reason can be the quality of the weather measurements in the Moselle region, because of difficulties in measuring precipitation in mountainous areas. The flood wave characteristics are however relatively good simulated with HBV, so this is less likely.

6 Discussion

This chapter is about the discussion regarding the conducted research. A discussion on the validity of this research is described in paragraph 6.1. The applicability of the GRADE outcomes for Dutch river flood protection is discussed in paragraph 6.2.. The international applicability of the research outcomes is discussed in paragraph 6.3.

6.1 Research validity

Calibration

The GLUE method is used to calibrate the HBV model. The philosophy behind using this method is that instead of finding one optimal parameter set, multiple behavioural parameter sets are accepted for the possible realisation of the basin's hydrology (Winsemius et al., 2013). However the subjectivity involved in determining the threshold of the objective functions used to select behavioural parameter sets is a major drawback of this technique (Jin et al., 2010). Furthermore only one parameter set is used for the simulations used in this research. The parameter set used to calculate the median 1/10 year discharge is selected as reference parameter set and is used to do the HBV-WG simulations. This choice is however arbitrary and probably does not lead to the best results.

The Rhine catchment is divided into 7 sub-basins in order to assess the performance of HBV, HBV-WG and WG in simulating the flood wave characteristics. Hegnauer et al. (2014) calibrated the HBV model for 15 sub-basins separately. Implication of using only 7 sub-basins is that potential errors in the 15 different sub-basins might not be detected, because over- and underestimation might compensate each other. This possibility is however only present in the Middle and Lower Rhine, because for both the calibration done by Hegnauer et al. (2014) and evaluation done is study the other sub-basins Moselle, Main, Neckar, East Alpine Rhine and West Alpine Rhine are used. The HBV models are calibrated on simulating mainly the observed peak discharges for the period 1989-2006. The evaluation conducted in this study is done for the period 1951-2006. It is therefore obvious that evaluation results of mainly the peak discharges for at least the period 1989-2006 are the same as the calibration results. Because of this it might have been better to evaluate on the basis of observed discharges outside the calibration period.

Data

Data provided by the GRDC for the period 1951-2006 is mainly used in the research. The river geometry of the Rhine has changed since 1951. Canalization of the Rhine upstream of Lobith has mainly been executed between 1955 and 1977 (Parmet et al., 2001). The altering of the river geometry along with other human influences like for example urbanization are not accounted for in the used data and in the model. The incurred adjustments in the river basin will influence mainly the peak discharges. Peak discharges before the canalization will be smaller than after the canalization due to for example less damping of the flow velocity by meander bends. To cope with this effect homogenisation of the historical peak discharge can be executed as is described by Parmet et al. (2001). Homogenisation of the used data is however not incorporated in this research, whereas it is incorporated in the traditional extrapolation of annual peak discharges used to obtain the design discharges for the Rhine.

Rating curves are used to transform water levels measured at the gauging stations in the Rhine into discharges (Droge et al., 1992). The rating curves are established by calculating the discharge from measured water levels and flow velocities by the use of a velocity-area method. Extrapolation of the curve beyond the highest measured point is particularly needed for converting high water levels, for which often few water level-discharge measurements are present (Coxon et al., 2015). Because of this, assumptions are incorporated in the extrapolation of the rating curve. Therefore differences between especially the extreme observed and simulated flood wave characteristics might theoretically sometimes be due to the quality of the gauge station in converting the water level measurement into a discharge.

A relatively short period of 56 years is used in this evaluation. For some locations longer discharge series are available. Using these longer series gives insight in how well the performance of HBV, HBV-WG and WG is in simulating flood wave characteristics of periods on which the WG is not based on. Assessing the skill of HBV and HBV-WG in simulating the flood wave characteristics, when comparing them to the flood wave characteristics obtained from 107 year observations at Lobith, shows only minor differences, see appendix 5. These minor differences might however be found, because no homogenisation of the long discharge series is applied. The underestimation of peak discharges will for example namely be less, because annual maximum discharges will be increased by homogenisation (Parmet et al., 2001).

Method

The selection of the flood waves is done by the use of a threshold value and time window. For each sub-basin the magnitude of both boundary conditions is obtained from the observed discharge series in the same way. For the basins in the Alpine region this selection leads to long flood waves. Waves that are physically separated by approximately two months are allocated to one single wave, which is intuitively not right. The implication of the large windows is assessed by a sensitivity analysis, which can be seen in appendix 3. It shows that the influence on the performance assessment is minimal when looking at the performance of HBV in simulating the flood wave characteristics. Only the performance in simulating the right number of waves is substantially influenced by the window size.

Results

Most detected differences can be attributed to the performance of HBV. However in many cases the performance of HBV-WG is less good than only the HBV simulations. In most circumstances the WG ensures for weather conditions that lead to comparable discharge as the ones simulated for the 56 year reference period. The WG is furthermore capable to simulate extra spread in the flood wave and is therefore able to simulate different extremes without changing the mean of the characteristic. Feeding the synthetic weather series into the HBV model often increases already detected differences with the observations. The reason for this is that the WG is in fact an extra source of uncertainty. It is therefore logical that the performance of HBV-WG is generally a little worse than the performance of HBV only. It is also possible that the WG compensates for errors due to HBV, this is however not desirable because then the performance might be only accidently good, whereas it should be good because of physically reasonable outcomes.

6.2 Applicability of GRADE outcomes for Dutch river flood protection

GRADE is designed to simulate physically reasonable low probability flood events, which will be used as hydraulic boundary conditions in the next dike assessment round in the Netherlands. It is therefore important that simulated flood wave characteristics are physically plausible. This research shows that the flood wave characteristics of flood waves at Lobith are simulated well by both HBV and HBV-WG. Also the volumetric contributions of upstream sub-basins to flood waves at Lobith is simulated well by HBV and HBV-WG. However some errors are detected in upstream sub-basins.

The skill of HBV in simulating the flood waves from the two Alpine region basins is poor. However most flood waves from these basins occur in early summer, whereas most waves at Lobith occur in winter. The volumetric contributions from these basins are simulated well by both HBV and HBV-WG. The detected errors are therefore mainly important for low discharge conditions at Lobith. However the relative contribution to Lobith waves from the Alpine region is 29%, so errors in the flood waves from the Alpine region will also affect the skill of the model at Lobith.

The WG is responsible for an underestimation of the volumetric contribution to flood waves at Lobith of the Moselle basin. The flood wave characteristics from the Moselle are however simulated well, in fact the WG compensates for errors in simulating the flood wave peak discharges. Also the relative contribution to Lobith waves is just 19%. The impact of the errors detected in the Moselle at Lobith are therefore relatively small.

Other errors detected in flood waves from upstream sub-basins are negligible when looking at the flood waves at Lobith. It can therefore be argued that HBV and HBV-WG are able to simulate physically reasonable flood waves at Lobith if the errors detected in the HBV models from the Alpine region and the errors in the WG of the Moselle will be corrected.

6.3 International applicability of the research outcomes

Detected errors are mainly due to the performance of the HBV model, the applied WG is often only responsible for some minor differences. This shows that this type of WG is probably widely applicable for synthetic river flood wave simulation. The quality of the weather data in the river basin to apply this WG should however be good and detailed enough.

Several studies combine WGs with hydrological models to simulate low probability discharges for assessments concerning river floods. Validation of such models focuses mainly on the skill of the models in simulating peak discharges. This study shows that simulating the characteristics of low probability flood waves, from the relative hydrologically complex Rhine basin, with a WG combined with a hydrological model is possible. The evaluation results of the flood wave characteristics are however specifically about the performance of the WG and HBV model applied in the Rhine. Generalization of the results is difficult, because in other river basins different types of WGs and hydrological models are applied. However the method used to evaluate the performance is generally applicable for comparable methods, in which hydrological models and weather generators are coupled to simulate low probability discharges, if observed discharge series are available. This is for example the case for the models discussed by Blazkova and Beven (2004), Haberlandt et al. (2008), Kuchment and Gelfan (2011) and Falter et al. (2015).

7 Conclusions and Recommendations

The objective of this study is to assess the performance of the hydrological model HBV and the combined performance of the weather generator and HBV, used within GRADE, in simulating flood wave characteristics (peak discharge, peak timing, volume, duration and number of flood waves per year) and the contributions of 7 major Rhine sub-basins to flood waves at Lobith. The conclusions regarding this objective are discussed per sub-question in paragraph 7.1. Recommendations are given in paragraph 7.2.

7.1 Conclusions

Q1. How well are the flood wave characteristics peak discharge, peak timing, volume, duration and number of waves per year from the Rhine at Lobith and upstream sub-basins simulated with HBV when comparing the flood wave characteristics obtained from discharge observations and simulations?

The characteristics of the flood waves from most sub-basins are simulated well by the HBV model. The volumetric contributions are simulated well by the HBV model. Only a small underestimation of the relative contribution from the two Alpine sub-basins is found. This underestimation is compensated by the overestimation of relative contributions from the German sub-basins. The performance of HBV in simulating the characteristics of sub-basin flood waves that contribute to flood waves detected at Lobith is relatively good. The characteristics of contributing flood waves from the two Alpine sub-basins are simulated worse than the characteristics of other flood waves from these two sub-basins. Flood waves at Lobith mainly occur during the winter months, whereas flood waves from the Alps occur often in early summer. So the characteristics of winter flood waves from the Alpine region are simulated worst. The evaluation of all sub-basin flood waves reveals that mainly the performance of HBV in simulating the flood wave characteristics of flood waves from the Alps is poor. The peak discharges and number of flood waves per year from the East Alpine Rhine are generally overestimated, whereas the durations and volumes are underestimated. Similar results are found for the West Alpine Rhine, only the peak discharges and number of flood waves per year are not overestimated. The detected differences are possibly due to the skill of HBV in simulating the snow storage. If too little water is allocated to the snow storage than the volumes and durations of flood waves which are fed by snow melt will be underestimated. The overestimation of East Alpine Rhine peak discharges might be due to precipitation that is allocated to runoff instead of snow storage.

Q2. What is the performance of the combination of HBV and WG in simulating the flood wave characteristics peak discharge, volume, duration and number of flood waves per year of flood waves from the Rhine at Lobith and upstream sub-basin when comparing the characteristics of flood waves obtained from the observed and simulated discharge series?

The performance of HBV-WG in simulating the relative volumetric contributions is slightly different from the performance of only HBV in simulating the contributions. No relative underestimation is found anymore from the Alpine region. The contribution from the Moselle is underestimated, because of differences due to the synthetic weather series. The HBV-WG performance in simulating the flood wave characteristics from the upstream sub-basins is comparable to the performance of only HBV. This shows that the main detected differences are due to the performance of HBV. In many cases the WG only slightly worsens the performance of the HBV-WG. This effect is expected because implementing an extra model component increases the uncertainty and therefore the chance that the simulations differ more from the observations. Only calculated differences in flood wave characteristics from the Moselle can be allocated to the WG. Using criteria comparable to those used in the WG of the Meuse basin, might probably improve the skill of the WG in the Moselle. This should however be studied, because changing the WG of the Moselle only, might be difficult and changing the WG of the whole Rhine basin will probably worsen the performance of the WG in simulating proper weather conditions in the other sub-basins.

7.2 Recommendations

Scientific recommendations

The largest errors are found in the flood waves from the West Alpine Rhine and East Alpine Rhine. These differences are mainly due to the performance of the HBV model. An extensive validation of the HBV models applied in the Alpine region is recommended. Kersbergen (2016) showed that improvements for low discharges can be made by recalibrating some HBV models in the Alpine region. Within the validation special attention should be payed to the performance of HBV in simulating snow storage, because most detected differences are probably due to errors in the non-allocation of precipitation to snow storage.

This study focuses on the performance of HBV, HBV-WG and WG in simulating the flood wave characteristics. HBV is mainly responsible for detected differences. It is recommended to evaluate the skill of HBV in simulating the underlying hydrological processes to give a clear reason for detected differences in the flood wave characteristics.

Measuring the precipitation in mountainous areas is difficult, especially under snowfall conditions due to wind induced undercatch problems. It is therefore recommended to assess the implications of using uncertain snowfall (precipitation) series in the HBV models of the Alpine region on simulating the flood wave characteristics.

The WG is responsible for some errors in the simulated flood wave characteristics of the Moselle. An assessment on how the skill of the WG might be improved would be interesting. Especially because in the neighbouring comparable river basin of the Meuse different criteria are used within the WG.

This study shows that the ability of the WG in simulating weather conditions that result in flood waves comparable to observed flood waves is good. It is therefore recommended to analyse if this WG might also be applicable in other river basins.

The performance of GRADE in simulating the flood wave characteristics from the Meuse is not analysed. It is recommended to analyse the characteristics of simulated flood waves from the Meuse using the method presented in this study, because GRADE will also be used to determine the hydraulic boundary conditions required to assess the dikes around the Meuse River.

Recommendations for GRADE users

This study shows that the flood wave characteristics at Lobith are simulated well and to a large extent for physically plausible reasons. It can therefore be stated that GRADE is applicable for Dutch river management. One should however be cautious with this statement, because flood waves from mainly the Alpine region are poorly simulated. The effect of this is clearly visible when assessing the low flow conditions in the Rhine as is done by Kersbergen (2016), because most flood waves from the Alpine region occur in early summer. It is therefore recommended to be reserved in using GRADE in its current stage for assessments concerning river floods.

Only the duration of the flood waves simulated at Lobith is slightly underestimated by the model. It is therefore recommended to assess possible dike stability assessment implications of this underestimation.

For example too optimistic outcomes regarding the piping assessment might be a result of the underestimation of the flood wave duration.

This study shows that not only the most extreme flood waves are simulated well by HBV and HBV-WG, but in fact all flood waves at Lobith are simulated properly. Relatively small flood waves have mainly impact on the operational management of and around the dikes. These small flood waves deposit for example loads of mainly organic waste near the dikes. The frequency to clean up this waste can for example be predicted with this model. Furthermore the inundation frequency of the flood plains can be determined by the use of all flood waves simulated with GRADE. This might help landowners to manage their flood plains more efficient.

GRADE adjustments

The characteristics of flood waves at Lobith are simulated well. It is therefore recommended to assess possible model extensions, like for example a 2d inundation model extension to simulate up- or downstream flooding. Such an extension might give helpful information to determine for example the maximum inundation probability. Falter et al. (2015) added for example a flood-loss model to their model.

Quite some subjectivity is incorporated in the GLUE method used to calibrate the HBV models. It is recommended to investigate possible other calibration techniques for calibration of the Alpine HBV models on the basis of the flood wave characteristics.

The societal and economic consequences of using GRADE for determining the hydraulic boundary conditions might sometimes be large. Disapproved flood protection measures should namely be ameliorated, which might lead to public opposition. It is therefore important that the simulated flood wave characteristics are physically plausible. People negatively affected by projects concerning the improvement of flood protection measures might use the large differences found for the Alpine region flood waves as argument against using GRADE. Because the Alpine region is responsible for 29% of the total Lobith wave discharge, those people have a point. It is therefore recommended to improve the HBV models for the Alpine region. If the optimal calibration still results in large differences between observed and simulated flood wave characteristics, it is recommended to assess if applying another hydrological model might improve the performance in simulating the flood wave characteristics from the Alpine region.

Bibliography

Belz, J. U., Brahmer, G., Buiteveld, H., Engel, H., Grabher, R., Hodel, H., . . . Vuuren, W. v. (2007). *Das Abflussregime de Rheins und seiner Nebenflüsse im 20. Jahrhundret* (ISBN 978-90-70980-33-7). KHR. Lelystad, The Netherlands.

Bestuursakkoord Water. (2011). Unie van Waterschappen, Ministerie van Infrastructuur en Milieu, Vereniging van Nederlandse Gemeenten, ipo, Vewin. Den Haag, The Netherlands.

Blazkova, S., & Beven, K., (2004). Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic. Journal of Hydrology, 292: 153-172.

Bolwidt, L., Schoor, M., Hal, L. v., & Roukema, M., (2007). Hoogwater op de Rijn en de Maas. Rijkswaterstaat Riza.

Booij, M. J., (2002). Extreme daily precipitation in western europe with climate change at appropriate spatial scales. International Journal of Climatology, 22: 69-85. doi:10.1002/joc.715

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J., (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water resources research, 51(7): 5531-5546. doi:10.1002/2014wr016532

Davis, J. C., (2002). Statistics and Data Analysis in Geology, Third Edition. John Wiley & Sons, New York * Chishester * Brisbane * Toronto * Singapore.

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y., (2013). Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times. Hydrological Processes, 27(19): 2742-2758. doi:10.1002/hyp.9402

Donaldson, R. J., Dyer, R. M., & Kraus, M. J. (1975). *An objective evaluator of techniques for predicting severe weather events*. Paper presented at the 9th Conference on Severe Local Storms. 321-326.

Droge, B., Engel, H., & Golz, E., (1992). Channel Erosion and Erosion Monitoring Along the Rhine River. Erosion and Sediment Transport Monitoring Programmes in River Basins, 210: 493-503.

Eberle, M., Buiteveld, H., Beersma, J., Krahe, P., & Wilke, K. (2002). *Estimation of extreme floods in the river Rhine basin by combining precipitation-runoff modelling and a rainfall generator, CHR Report II-17.* Paper presented at the International Conference on Flood Estimation, Bern, Switserland.

Eberle, M., Buiteveld, H., Wilke, K., & Krahe, P. (2005). *Hydrological Modelling in the River Rhine Basin Part III - Daily HBV Model for the Rhine Basin, document number: BFG-1551*. Institute for Inland Water Management and Waste Water Treatment (RIZA), Federal Institute of Hydrology (BfG)

Ehret, U., & Zehe, E., (2011). Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrol. Earth Syst. Sci., 15(3): 877-896. doi:10.5194/hess-15-877-2011

EURECO, Ecologisch onderzoek & advies. (2015). Vierde toetsronde primaire waterkeringen vanaf 2017. http://www.zodenaandedijk.com/dijkentoetsing4.html

Falter, D., Schröter, K., Dung, N. V., Vorogushyn, S., Kreibich, H., Hundecha, Y., . . . Merz, B., (2015). Spatially coherent flood risk assessment based on long-term continuous simulation with a coupled model chain. Journal of Hydrology, 524: 182-193. doi:10.1016/j.jhydrol.2015.02.021

Frei, C., & Schär, C., (1998). A precipitation climatology of the Alps from high-resolution rain-gauge observations. International Journal of Climatology, 18(8): 873-900. doi:10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9

Haberlandt, U., Eschenbach, E. v., & Buchwald, I., (2008). A space-time hybrid hourly rainfall model for derived flood frequency analysis. Hydrology and Earth System Sciences, 12: 1353-1367.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M., (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Research: Atmospheres, 113(D20). doi:10.1029/2008JD010201

The HBV model. (2015). http://www.macaulay.ac.uk/hydalp/private/demonstrator_v2.0/models/hbv.html

Hegnauer, M., Beersma, J. J., Boogaard, H. F. P. v. d., Buishand, T. A., & Passchier, R. H. (2014). *Generator of Rainfall and Discharge Extremes (GRADE) for the Rhine and Meuse basins - final report of GRADE 2.0*. Deltares. Delft, The Netherlands.

Hegnauer, M., & Verseveld, W. v. (2013). *Generalised likelihood uncertainty estimation for the daily HBV model in the Rhine basin, Part B: Switzerland*. Deltares. Delft, The Netherlands.

Jin, X. L., Xu, C. Y., Zhang, Q., & Singh, V. P., (2010). Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. Journal of Hydrology, 383(3-4): 147-155. doi:10.1016/j.jhydrol.2009.12.028

Kersbergen, A. (2016). *Skill of a discharge generator in simulating low flow characteristics in the Rhine basin*. Universtity of Twente. Enschede, The Netherlands.

Knoeff, H., & Steffess, H. (2014). *Workshop WTI 2017*. Ministerie van Infrastructuur en Milieu.

Kuchment, L. S., & Gelfan, A. N., (2011). Assessment of extreme flood characteristics based on a dynamic-stochastic model of runoff generation and the probable maximum discharge. IAHS Publication, 347: 29-35.

Kundzewicz, Z. W., Hirabayashi, Y., & Kanae, S., (2010). River Floods in the Changing Climate—Observations and Projections. Water Resources Management, 24(11): 2633-2646. doi:10.1007/s11269-009-9571-6

Leander, R., & Buishand, T. A. (2004). *Rainfall generetor for the Meuse basin, Development of a multi-site extenstion for the entire drainage area*. De Bilt, The Netherlands.

Legates, D. R., & McCabe, G. J. J., (1999). Evaluating the use of ''goodness-of-fit'' measures in hydrologic and hydrodynamic model validation. Water resources research, 35(1): 233-241. doi:10.1029/1998WR900018

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S., (1997). Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology, 201(1-4): 272-288. doi:10.1016/S0022-1694(97)00041-3

Middelkoop, H., & Haselen, C. O. G. v., (1999). Twice a River, Rhine and Meuse in the Netherlands, RIZA report no. 99.003. RIZA, Arnhem - Lelystad, The Netherlands.

Ministerie van Infrastructuur en Milieu & Ministerie van Economische Zaken. (2014). *Deltaprogramma 2015 | Werk aan de delta*. Den Haag, The Netherlands.

Ministerie van Verkeer en Waterstaat. (2007). *Hydraulische Randvoorwaarden primaire waterkeringen*. Den Haag, The Netherlands.

Nienhuis, P. H., (2008). Environmental History of the Rhine-Meuse Delta, An ecological story on evolving human-environmental relations coping with climate change and sea-level rise. Springer Science+Business Media B.V., Nijmegen, The Netherlands, 639 pp.

Parmet, B., Buishand, T. A., Brandsma, T., & Mülders, R., (1999). Design discharge of the large rivers in The Netherlands--towards a new methodology. Hydrological Extremes; Understanding, Predicting, Mitigating. IAHS publication, 255: 269-272.

Parmet, B. W. A. H., Langemheen, W. v. d., Chab, E. H., Kwadijk, J. C. J., Diermanse, F. L. M., & Klopstra, D. (2001). *Analyse van de maatgevende afvoer van de Rijn te Lobith, Onderzoek in het kader van het randvoorwaardenboek 2001, RIZA rapport 2002.012*. Arnhem, The Netherlands.

Passchier, R. H. (1996). *Evaluation hydrologic model packages. Technical Report Q2044*. WL/Delft Hydraulics. Delft, The Netherlands.

Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., & Gratzki, A., (2013). A Central European precipitation climatology; Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS). Meteorologische Zeitschrift, 22(3): 235-256. doi:10.1127/0941-2948/2013/0436

Schmeits, M. J., Beersma, J. J., & Buishand, T. A. (2014b). *Rainfall generator for the Meuse basin: Description of simulations with and without a memory term and uncertainty analysis, KNMI publication 196-VI*. Royal Netherlands Meteorological Institute, Ministry of Infrastructure and the Environment. De Bilt, The Netherlands.

Schmeits, M. J., Wolters, E. L. A., Beersma, J. J., & Buishand, A. T. (2014a). *Rainfall generator for the Rhine basin: Description of simulations using gridded precipitation datasets and uncertainty analysis, KNMI publication 186-VII*. De Bilt, The Netherlands.

Shaw, E. M., Beven, K. J., Chappell, N. A., & Lamb, R., (2011). Hydrology in Practice, fourth edition. Spon Press, London and New York.

Shrestha, D. L., Kayastha, N., & Solomatine, D. P., (2009). A novel approach to parameter uncertainty analysis of hydrological models using neural networks. Hydrol. Earth Syst. Sci., 13(7): 1235-1248. doi:10.5194/hess-13-1235-2009

Steinrücke, J., Fröhlings, B., & Weißhaupt, R. (2012). *HYMOG Hydrologische Modellierungsgrundlagen im Rheingebiet. KHR Bericht I-24*. Lelystad, The Netherlands.

Tockner, K., Uehlinger, U., & Robinson, C. T., (2009). Rivers of Europe. Academic Press, Elsevier, Oxford, UK.

Tongal, H., Demirel, M. C., & Booij, M. J., (2013). Seasonality of low flows and dominant processes in the Rhine River. Stochastic Environmental Research and Risk Assessment, 27(2): 489-503. doi:10.1007/s00477-012-0594-9

Ulbrich, U., & Fink, A., (1995). The January 1995 Flood in Germany: Meteorological Versus Hydrological Causus. Physics and Chemistry of the Earrth, 20(5-6): 439-444. doi:10.1016/S0079-1946(96)00002-X

Walker, W., Abrahamse, A., Bolten, J., Braber, M. d., Garber, S., Kahan, J., . . . Riet, O. v. d. (1993). *Toesting uigangspunten rivierdijkversterkingen, Deelrapport 1: Veiligheid tegen overstromingen, Veiligheidsanalyse, kostenschatting en effectenbepaling*. Commisie Boertien, Ministerie van Verkeer en Waterstaat.

Whitehouse, D. J., (2011). Handbook of Surface and Nanometrology, second edition. CRC Press, Concentry, UK.

Winsemius, H., Verseveld, W. v., Weerts, A., & Hegnauer, M. (2013). *Generalized Likelihood Uncertainty Estimation for the daily HBV model in the Rhine Basin, Part A: Germany*. Deltares. Delft, The Netherlands.

Appendix 1 Overall performance of HBV in simulating sub-basin discharges

Table 10 shows the overall performance of HBV in simulating the discharges from the Rhine sub-basins for the period 1951-2006. Nash and Sutcliffe is used to assess the performance and is calculated with equation (1). The further from Lobith the worse the performance of HBV. Nash and Sutcliffe values for the discharge above threshold indicate that the relative high discharges are simulated relatively bad. Mainly the simulated high flows from the Alpine region seem to be very poor. The average total discharge of simulated upstream flood waves that contribute to flood waves at Lobith is larger than the average total discharge of all simulated flood waves from the Middle Rhine, Moselle, Main and Neckar. The average total discharge of waves from the Alpine region that contribute to flood waves at Lobith is smaller than the average total discharge of all waves from those regions. In both simulated and observed time series comparable effects are seen. The total simulated discharge divided by the total observed discharge shows that in most basins the total discharge is underestimated. Only for the Main and Neckar no major differences between the total observed and simulated discharge are found.

Table 10 Overall performance of HBV in simulating sub-basin discharges for the period 1951-2006 (Main 1963-2006)

	NSE	**NSE above Q5**	**Total simulated discharge of threshold waves averaged per wave / Total simulated discharge of Lobith waves averaged per wave**	**Total observed discharge of threshold waves averaged per wave / Total observed discharge of Lobith waves averaged per wave**	**Total simulated discharge / total observed discharge**
Rhine upstream of Lobith	0,92	0,57	1,00	1,00	0,92
Rhine upstream of Andernach	0,88	0,46	0,92	0,86	0,89
Moselle	0,89	0,53	0,82	0,68	0,86
Main	0,81	0,35	0,85	0,81	1,01
Neckar	0,75	0,13	0,48	0,38	1,00
West Alpine Rhine	0,71	-1,46	1,68	1,58	0,85
East Alpine Rhine	0,66	-1,33	2,55	2,78	0,87

Appendix 2 Comparison between the relative number of waves per month coming from the Rhine at Lobith, West Alpine Rhine and East Alpine Rhine

Figure 26 shows the relative number of waves per month, which are selected from the discharge series measured in the period 1951-2006. It show that flood waves from the Alpine Rhine are mainly present in late spring and early summer time, whereas waves at Lobith generally appear in the winter months.

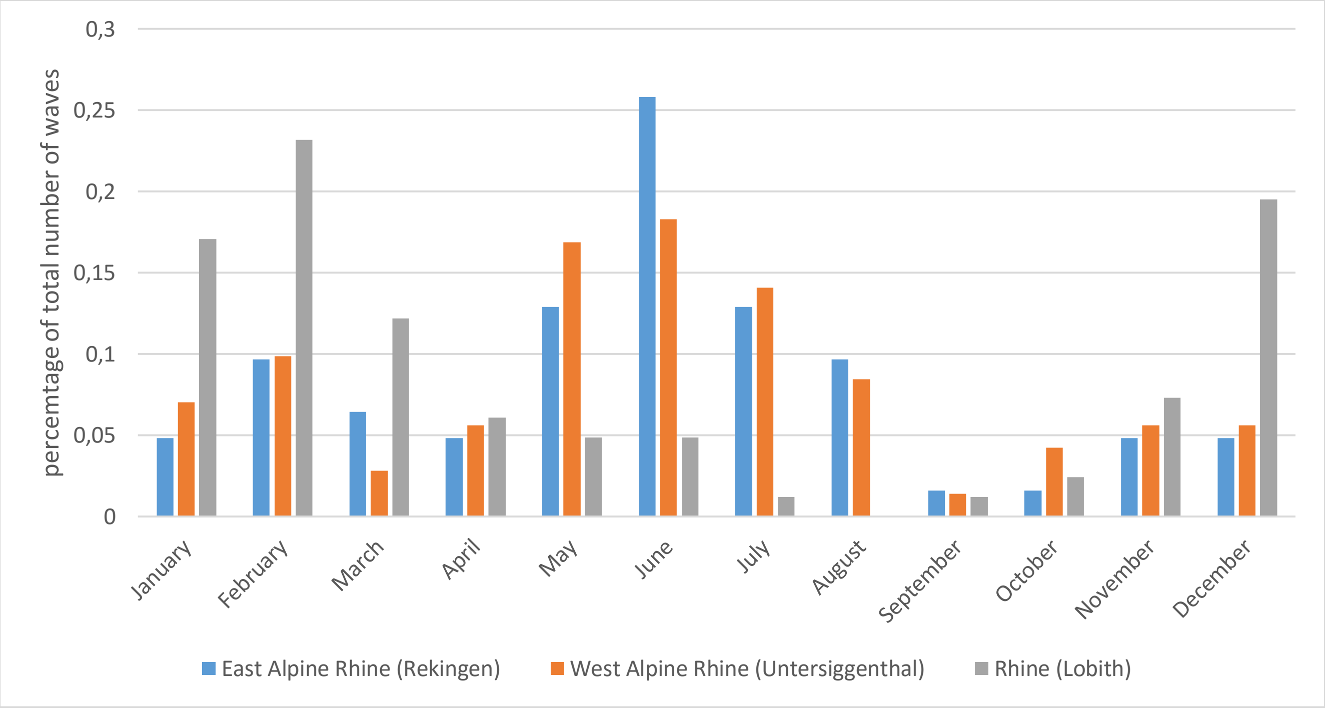

Figure 26 Relative number of flood waves per month from the Rhine at Lobith and the two Alpine region basins

Appendix 3 Sensitivity of the influence of window size on the performance of HBV in simulating the flood wave characteristics at Untersiggenthal for the period 1951-2006

Figure 27 shows that the number of detected waves is relatively strong influenced by the size of the applied time window. The sensitivity of the number of detected waves is largest when the windows are small. Larger windows result in a milder response of the number of detected waves.

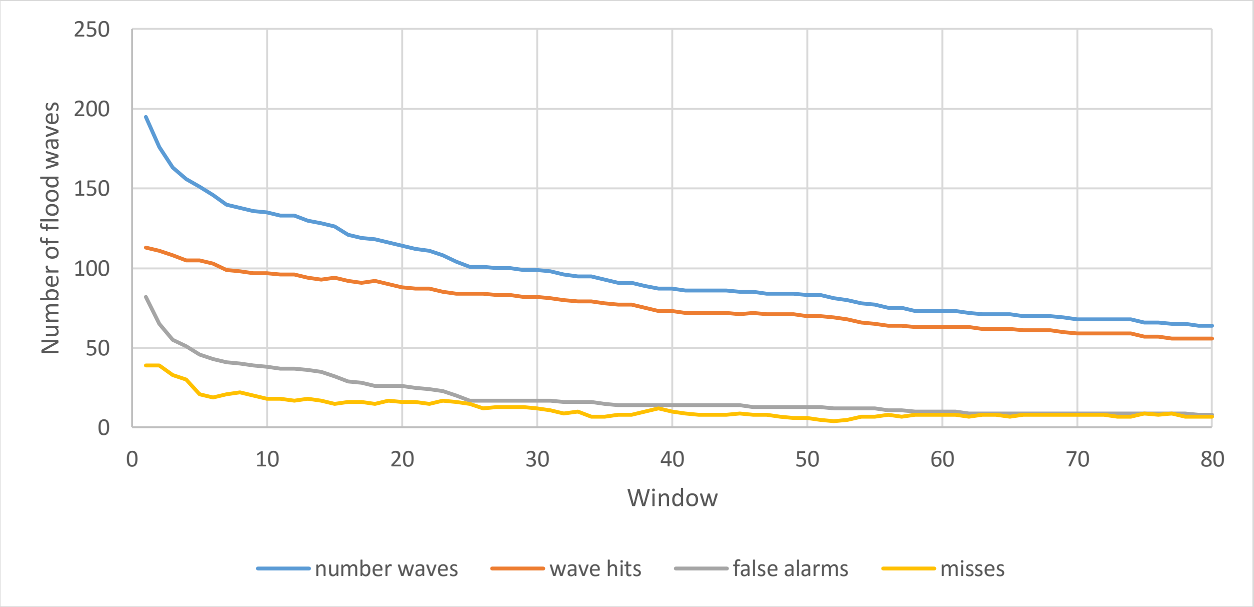

Figure 27 Sensitivity plot of the influence of window size on the performance of HBV in simulating the number of flood waves for the period 1951-2006

Figure 28 shows that the influence of the window size on the calculated performance of HBV in simulating the flood wave characteristics is small. This can be concluded because the lines representing the performances are relatively horizontal.

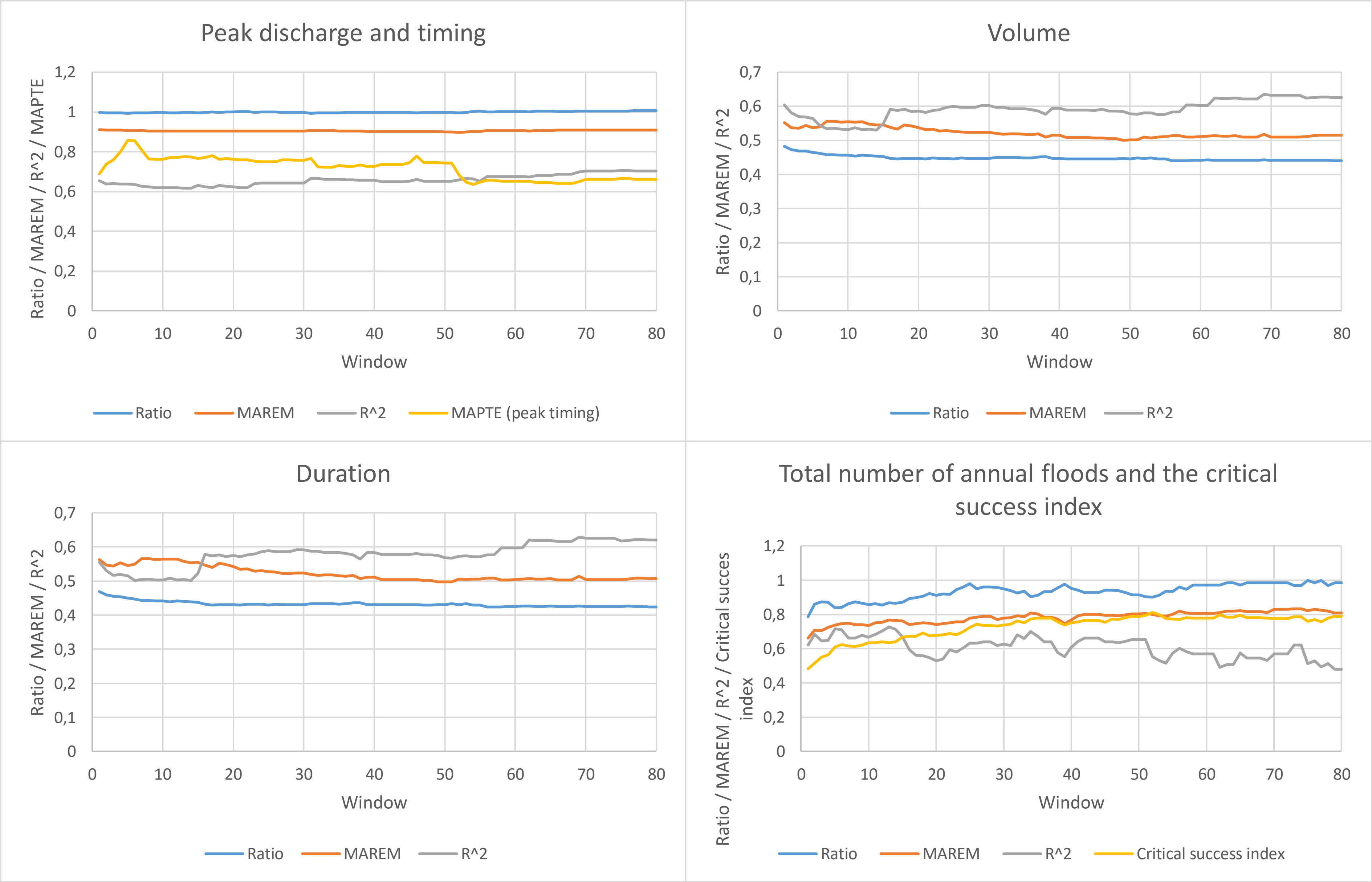

Figure 28 Sensitivity plots of the influence of window size on the performance of HBV in simulating the flood wave characteristics

Appendix 4 Performance of HBV in simulating the lake levels

The performance of the HBV model in simulating flood wave characteristics from the Alpine region is poor. The major lakes have a considerable effect on the discharges (Hegnauer et al., 2014). It is therefore decided to assess the performance of HBV in simulating the water levels of these four major lakes. Lake Constance, also known as the Bodensee, is located in the East Alpine Rhine basin. Lake Neuchâtel, Lake Lucerne and Lake Zürich are located in the West Alpine Rhine basin. Mr. M.C. Demirel provided water level measurements for the period 1978-2006, which were provided to him by GRDC. Both the measured and simulated water levels are given in meters above sea level. Because the lakes are hundreds of meters above sea level, the reference level hardens the interpretation of results from calculating for example the ratio. It is therefore decided to subtract the mean from each time step and to at 2 meters to all time steps to prevent the appearance of negative water levels.

For an indication of the overall skill of HBV in simulating the lake water levels, the Nash and Sutcliffe efficiency criterion is calculated for the whole period 1978-2008 with equation (1).

The observed and simulated water levels are furthermore compared for the periods in which flood waves are observed. A one day window around the observed flood waves is used to account for travel times to the gauging station. The ratio (equation (9)), the coefficient of determination (equation (10)) and the mean absoluter error (equation (18)) are calculated to assess the differences.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}|O_i - S_i| \qquad (18)$$

In which N is the total number of data points, S is the simulated water level, O is the observed water level and i is the index number of data point.

Figure 29 shows the performance of HBV in simulating the water levels of the four major lakes in the Alpine region. Only the overall performance of simulating the water level of Lake Lucerne is very poor, which can be seen from the very small Nash and Sutcliffe value. The skill of HBV in simulating the water levels of the lakes during periods in which flood waves are detected is relatively good. In Lake Constance the strongest structural underestimation is found, MAE indicates that the absolute error is relatively large, namely 0,36 meter. So during flood events the water level is underestimated by 0,36 meter. The pattern of the simulated water levels is comparable to that of the observed water levels. The lakes levels in the West Alpine Rhine area are all structurally underestimated by HBV. The absolute errors are however relatively small compared to the error in the water levels of Lake Constance. The pattern of the simulated water levels is comparable to that of the observed water levels, which can be concluded because the coefficient of determination is reasonably high.

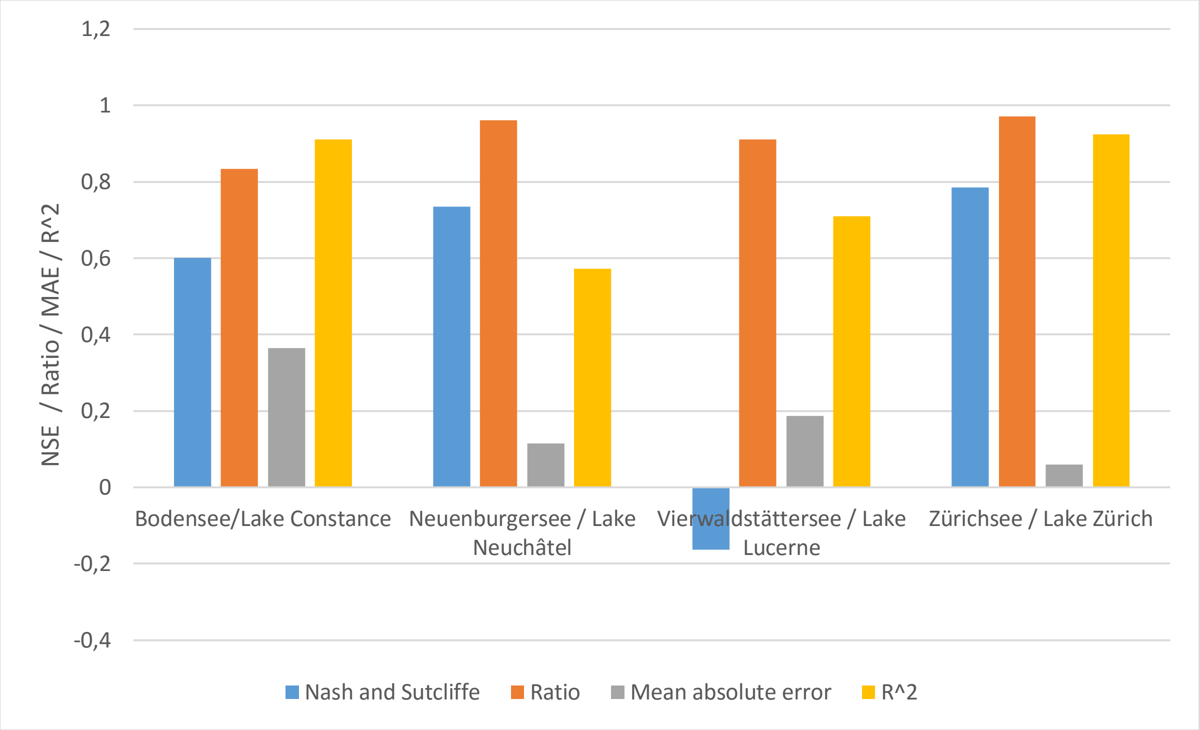

Figure 29 Performance of HBV in simulating the water levels in the four major lakes in the Alpine region

Appendix 5 Influence of a different reference period (1901-2008), on the performance of HBV and HBV-WG in simulating the flood wave characteristics at Lobith

HBV simulations that correspond to historical observations are only available for the period 1951-2006, because of the available weather data. Comparison by the use of statistical tests and CDFs is therefore done in this analysis. In table 11 the result of the statistical test used to assess the equality of mean and variances of the flood wave characteristics obtained from the 1901-2008 and 1951-2006 observations, the HBV simulations and the HBV-WG simulations can be seen. Mainly significant differences between the 1901-2008 observations and the HBV-WG simulations are detected. This might be because the generated weather is constructed on the basis of weather data for the period 1951-2006.

Table 11 Results of the statistical tests used to assess equality of the mean and variance calculated from the flood wave characteristics obtained from observed (1951-2006), observed(1901-2008), HBV and HBV-WG simulated discharge series at Lobith (Green = no significant difference, Red = significant difference) (Significance level = 0,05)

p-values Lobith 1901-2008	Observed (1951-2006) with observed (1901-2008)		Observed (1901-2008) with HBV simulated (1951-2006)		Observed (1901-2008) with HBV-WG simulated (10.000 years)	
	mean	variance	mean	variance	mean	variance
peak	0,3674	0,7849	0,1660	0,0337	0,0480	0,0024
volume	0,5704	0,9562	0,8566	0,3315	0,9371	0,1635
duration	0,4589	0,5530	0,2744	0,0640	0,0102	0,1416
annual floods	0,8079	0,9722	0,5065	0,6115	0,0042	0,3997

Figure 30 shows the CDFs calculated for the peak discharge, wave volume, wave duration and number of floods per hydrological yea, calculated from the observed discharge series of the period 1901-2008 and 1951-2006 and from discharge series simulated with HBV and HBV-WG. The CDF of the peak discharge shows no difference for the lowest 40% of the peaks, the highest 60% of the peaks are slightly overestimated when looking at the difference with the longer reference period and all other CDFs. This suggests that in the 1901-2008 more relative low peak discharges are found. This is expected, because human intervention in the Rhine basin done in the last century ensures faster runoff during periods with high discharges (Parmet et al., 2001). The division of extreme peaks is more or less the same, because the extra 52 years results in 2 extra extreme floods. These are the flood wave peaks from 1926 (highest ever measured) and 1920 (the third highest ever measured; besides this peak discharge is not found by Hegnauer et al. (2014). No major differences in the volumes is found. Two more extreme volumes are found in the longer reference period. The durations are a bit smaller in the reference period, probably also because of the alteration done in the basin. The number of flood waves found in the longer reference period is comparable with those from the initial reference period.

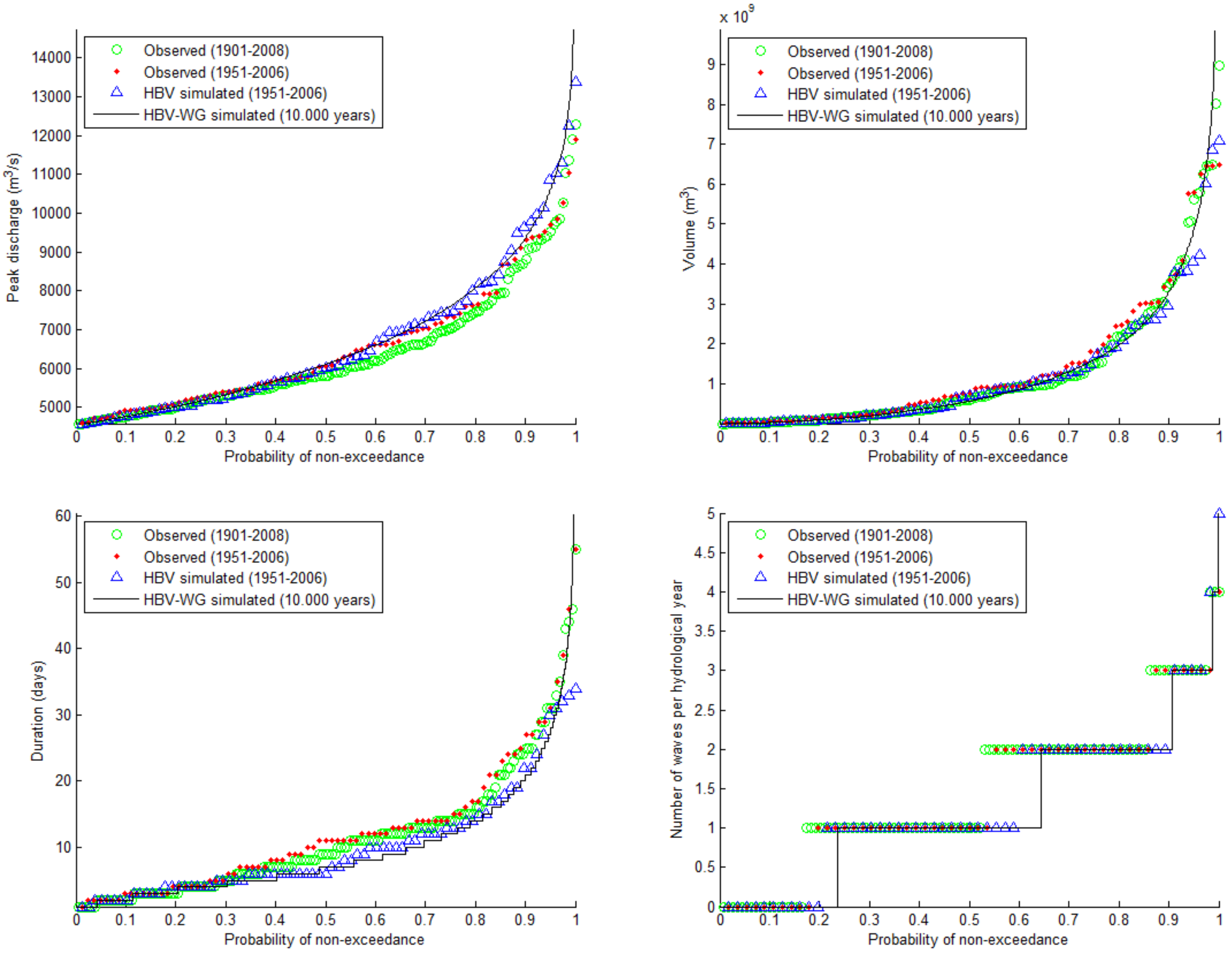

Figure 30 CDFs of the observed (1901-2008), observed (1951-2006), HBV simulated (1951-2006) and HBV-WG simulated flood wave characteristics at Lobith