Response to reviewer #1:
For clarity, the integral reviews are included in this response. Our answers to the questions raised by the reviewer are marked in italic text, and actual changes to the manuscript are indicated by a bold font.

General comments

This study presents potential future changes in hydrological response of the Ourthe catchment as projected by three different climate scenarios and simulated by the HBV model. The study has two parts. In the first, the authors evaluate the performance of four different bias correction approaches. The precipitation and air temperature outputs of the regional climate model are corrected according to the observations at one climate station. In the second, the bias corrected data are used as an input to the hydrological (HBV) model. Here, the authors compare the model simulations in both historical reference period and future time horizons.

Generally, the study has a good structure, is clearly written and within the scope of the journal. However, it should be also noted that the significance of potential future response simulations may be not so attractive for international audience (because of the results for only one specific catchment and the large uncertainty in climate scenarios). In order to learn more from this part, it may be worth to extend the discussion and put results in the context of existing studies (different/same scenarios from the same region, or the same scenarios in different regions).
We fully agree that the results of assessments like the present one are greatly influenced by the choice of the GCM (and to a lesser degree also the RCM). It is, however, in this case not feasible to repeat the model chain of GCM-RCM-bias correction-hydrological model using multiple models for each step. We also agree that we should have put our results into perspective, and therefore we added to the end of the discussion:
"In addition, our results are in line with studies using the same models in other basins (Hurkmans et al., 2010) as well as with studies applying different models in the same region (van Pelt et al., 2009), who all found higher discharges in winter and lower in summer for the end of the 21st century. Also Graham et al., (2007) found lower summer discharges and higher winter discharges both in the Rhine basin, and the Baltic region using a variety of RCMs and a few GCMs."

Much more interesting part is the application and assessment of the importance of the bias correction. The credence of the correction procedure is somewhat limited by the use of only one climate station (as it is already discussed by the authors), and it will be certainly interesting to see its performance in another regions and/or observations. However in comparison to the alternative evaluation, based on the assessment of relative changes between the reference and future scenarios, the approach used here is methodologically beneficial as it also allows to speculate about the changes in the absolute value of hydrologic response (e.g. a change in the storage in mm). Thus, I would suggest to discuss more how the hydrologic changes may differ according to the assessment used.
If we correctly understand the reviewer, he would like to see some discussion of how the results of climate change assessment would change had we used one of the other three methods of bias correction. Although there are certainly differences between the four bias correction methods employed, these between-method differences are much smaller than the differences between the observed, uncorrected and corrected datasets (see Figure 3). Because the direction of correction is similar as well, generally the results are not expected to change very much. Had we used the correction based on an individual pixel instead of the spatial average, discharge would have been higher in all seasons (hence more extreme peak flows and less extreme low flows) because these corrections "overshoot" the observations more than the correction based on spatial averages does.

Specific comments

Eq. 1: the description (and units) of the lakes and time constants in Eq. 1 is missing.

In the revised manuscript, we explain that "lakes" represent the water storage in lakes, and added units to all terms of Eq. 1.

Section 2.3: Please consider to add a table showing general characteristics of the climate scenarios used (e.g. what changes in precipitation and air temperature are projected by different scenarios).

We added a table to the manuscript showing mean annual values of precipitation, temperature and potential evaporation. In addition, we added the following to the last paragraph of Section 2.3: ***"Mean annual values of precipitation and temperature for all scenarios and the reference situation are shown in Table 1. In addition, potential evaporation values are shown that are derived using the method explained in Section 3.3."***

Table 1: Overview of average annual sums of precipitation and potential evaporation (E_p; derived using the method explained in Section 3.3), and average annual mean temperature, for the reference period and two 39-year periods in the 21st century according to three IPCC-scenarios.

Variable	*1962-2000*	*2002-2040*			*2062-2100*		
		A2	*A1B*	*B1*	*A2*	*A1B*	*B1*
Prec. [mm y^{-1}]	*1094*	*1103*	*1129*	*1199*	*1163*	*1107*	*1113*
Temp. [°C]	*7.19*	*7.69*	*7.83*	*7.73*	*10.29*	*10.53*	*9.31*
E_p [mm y^{-1}]	*674*	*665*	*673*	*666*	*690*	*712*	*680*

Section 3.1: Please give more details about the CRU dataset. Is it possible to evaluate the spatial variability of corrected data using the CRU dataset?

W*e agree that not much information was provided regarding the CRU dataset. The text at line 12, page 7152 was extended as follows:* ***"To check the representativeness of the meteorological station at St. Hubert for the catchment average value, a precipitation dataset from the Climate Research Unit (CRU) is plotted in Figure 3 as well. It contains monthly averages of observations, interpolated to a spatial resolution of 0.167 degree, or about 19 km. In Figure 3, it can be seen that the climatologies of the CRU dataset and the Saint Hubert station look very similar. Therefore, we conclude that the single observation site at Saint Hubert represents the entire catchment at the monthly time scale reasonably well."***

Because of the relatively low resolution of the CRU dataset (compared to the catchment size), and the fact that we do not have additional information on the interpolation method and the number of stations employed, we choose not to evaluate the spatial variability based on the CRU dataset and only use the catchment average value.

Section 3.2:

a) How is the elevation dependence introduced in model simulations consistent with the bias correction procedure?

In the bias correction procedure, we assume the spatial variability of REMO is correct, because we do not have additional information. As every pixel in REMO has its own mean elevation, the spatial pattern is also to a large extent dependent on topography. To reintroduce this dependency in the (lumped) model simulation, a constant lapse rate of 10% per 100 meter (HBV parameter PCALT) was chosen. This value is often used in studies using HBV (Seibert, 2005). In the revised manuscript, we included this in the first paragraph of Section 3.2: ***"Precipitation is lapsed with 10% per 100 m elevation difference, which is often used in studies using HBV (Seibert, 2005). By this lapsing, the dependency of precipitation on topography, which is present in the atmospheric forcing data, is reintroduced in the lumped model simulation."***

b) Please consider to present the calibrated model parameters and their ranges used in calibration. What is the weight of volume error in the objective function?
We added a Table showing all 15 parameters in the HBV model with their optimal values as determined by calibration and validation. The parameter ranges that are used in calibration are the same as used by Seibert (2000), and is this article is already referenced, we consider it not necessary to repeat those here.

In addition, we added to the last paragraph of Section 3.2: ***"The resulting set of optimal parameters is presented in Table 2."*** *To explain how the two objective functions are combined in the calibration, we added the following to the revised manuscript (Section 3.2):* ***"R_{eff} and |VE| are combined into the combined objective function R_c using Lindström (1997):***
$R_c = R_{eff} - 0.1|VE|$***"***

Table 2: Parameters of the HBV model and their optimal values resulting from calibration and validation.

Parameter	Symbol	Unit	Optimal value
Threshold temperature	TT	°C	0.7927
Degree-day factor	CFMAX	mm °C^{-1}d^{-1}	2.724
Snowfall correction factor	SFCF	-	1.2
Refreezing coefficient	CFR	-	0.03417
Water holding capacity	CWH	-	1.695e^{-5}
Maximum of soil moisture zone	FC	mm	119.91
Threshold for evaporation reduction	LP	-	1.0
Shape coefficient	Beta	-	2.012
Recession coefficient (upper stor.)	K_0	d^{-1}	0.2891
Recession coefficient (upper stor.)	K_1	d^{-1}	0.1563
Recession coefficient (lower stor.)	K_2	d^{-1}	0.0525
Threshold for K_0 to become K_1	UZL	mm	14.30
Maximum percolation	PERC	mm d^{-1}	1.495
Routing parameter	MAXBAS	d	3.565
Correction factor for pot. evaporation	CET	°C^{-1}	0.00980

Section 4: It may be interesting also to compare selected characteristics, as e.g. mean outflow, simulated by the calibration dataset (ERA) and the reference (bias corrected climate output). Are these consistent? To what extent may the uncertainty in model parameterization affect projected changes in hydrological response?
Although we agree that there might be some uncertainty related to the parameterization of the hydrological model, we certainly think that the meteorological forcing is the dominant source of uncertainty, especially when looking at monthly averages as we do in most of our analyses. Besides, results in (sub-basins of) the Rhine basin, very close by, using a different hydrological model (the Variable Infiltration Capacity (VIC) model; Hurkmans et al., 2010), yielded similar results.

We added the table showing some streamflow statistics of the calibration dataset, the reference dataset and the observations to put all datasets into perspective. In addition we added the

*following text, describing this table, to the first paragraph of Section 4: **"In order to put the different model simulations into perspective, Table 2 shows three streamflow statistics (mean streamflow, mean annual maximum and mean annual minimum), for 1) observed streamflow, 2) streamflow as simulated by the HBV model forced with observed precipitation (Saint Hubert), 3) streamflow as simulated by the HBV model forced by ERA; and 4) streamflow as simulated by the HBV model forced by the reference data. Because the latter is not constrained by observations, only statistics can be compared, not the actual time series. All values are calculated over the same, 17-year period (1980-1996), as this is the period of overlap between all datasets. In Table 2, it can be seen that the HBV model with observed forcing is quite similar to the observations, especially in terms of mean streamflow, but that ERA and the reference forcing datasets produce considerably more streamflow. It should be noted here that all discharges in Table 3 are obtained using the same set of model parameters (Table 2)".***

Table 3: Streamflow statistics of four employed sources: 1) observed discharge, 2) simulated discharge where HBV is forced by observations; 3) simulated discharge where HBV is forced by ERA, and 4) simulated discharge where HBV is forced by the reference run. Mean annual average, mean annual maximum and mean annual minimum values are shown.

Dataset	M. A. Average [$m^3 s^{-1}$]	M. A. Max. [$m^3 s^{-1}$]	M. A. Min [$m^3 s^{-1}$]
Observed	24	200	2.57
HBV-OBS	21.8	157.3	2.97
HBV-ERA	39.4	305.1	5.1
HBV-REF	35.9	262.3	4.35

Response to reviewer #2:

For clarity, the integral reviews are included in this response. Our answers to the questions raised by the reviewer are marked in italic text, and actual changes to the manuscript are indicated by a bold font.

This manuscript evaluated the impact of climate change on stream flow in the Ourthe catchment (1600km2). The authors used HBV model (a lumped distributed rainfall-runoff model) with bias corrected precipitation/climate information estimated by ECHAM5 (downscaled with a RCM). Four different bias correction methods were tested before their application. Even though I recognize the importance of this kind of challenge for local water management under climate change, I did not find any particular new method is proposed nor is any new surprising finding presented. Mainly following three major issues are pointed out after careful reviewing.

1: The study catchment has only one gauging station and meteorological station. Because of this limitation, a critical assumption had to be made in terms of the spatial variability: i.e. precipitation increase by 10% with every increase of 100m, without any detail discussion on the validity of this assumption. This must be critical because the spatial distributions are highly related to the structural errors in the RCM downscaling.

While we agree it is crude to assume that precipitation is increasing 10% per 100 meter, it is common practice in the HBV model to assume such a lapse rate (Seibert, 2000, 2005). Besides, with respect to the uncertainties that are involved in the bias correction (using only one observation point) and the RCM output, the results will most likely not change drastically because this assumption was made in all simulations (reference and scenarios).

Furthermore, there are still many debates about the minimum spatial scales for the direct application of GCM output (even if they are downscaled with a RCM). It is well known that as a study area becomes smaller, the error in the output of GCM tends to be larger. Of course, I agree that we eventually should be able to use GCM/RCM output for future streamflow assessment under climate change, but it is still necessary to discuss in detail potential uncertainty associated to the smaller scale application. Overall, I did not see clear reason why the author needed to choose this relatively small catchment which does not have much measurement information.

*We agree that there are many uncertainties in the output of GCMs and RCMs. However, because of the very high spatial resolution of the climate scenarios that are employed here, more of the topography is taken into account explicitly. It is thus, theoretically, possible to apply these scenarios to smaller catchments. Our analysis does indeed not yield very surprising results in that they are generally similar to other (larger-scale) studies in this region. This indicates that our results at least give an indication of the situation for the Ourthe catchment, but of course the uncertainties remain large, no matter how small or large the catchment. To discuss this more, the discussion at the end of the manuscript, placing our results in perspective, was extended with the following: **"In addition, our results are in line with studies using the same models in other basins (Hurkmans et al., 2010) as well as with studies applying different models in the same region (van Pelt et al., 2009), who all found higher discharges in winter and lower in summer for the end of the 21st century. Also Graham et al., (2007) found lower summer discharges and higher winter discharges both in the Rhine basin, and the Baltic region using a variety of RCMs and a few GCMs."***

One reason to choose the Ourthe is that recently many hydrological modeling studies have been or are being carried out in the Ourthe (e.g., Hazenberg et al., submitted), involving for example precipitation estimates from weather radar. Whereas at the moment few long time series of observations exist (only Saint Hubert), there will be more in the future. Assessing the influence of the spatial variability will then be possible. A second reason is that the modelling domain (for REMO) of the high-resolution climate scenarios that we used is mainly focused on the Rhine and Elbe catchments, although it also happens to cover the Ourthe (but not, unfortunately, the

southern part of the Meuse catchment).

2: Related to the above comment on the uncertainty, one of the ways to validate the applicability of GCM/RCM output for hydrologic simulation is to use the current climate condition outputs. Authors could simulate streamflow regimes with the output information during 1979 - 2003 to evaluate the uncertainty in the estimated variables. Especially, the ability to simulate extreme values (flood peaks and draught discharges) must be assessed first with the current climate condition. In addition, to reproduce flood peaks at this catchment scale, the simulation time step seems to be an important factor, but there is no information about the simulation time step.
*We indeed omitted to mention the model time step, which is one day. We added this to the first paragraph of Section 4: **"All model simulations are carried out using a time step of 1 day."***

Furthermore, for both the calibration and the validation period the model performance was good, (Nash-Sutcliffe values of 0.86 and 0.90 respectively). This gives some confidence in the ability of the hydrological model to simulate the current flow regime. In addition, we added a table showing streamflow statistics of model simulations for both the calibration and reference datasets, together with those of the observations (see below).

Table 3: Streamflow statistics of four employed sources: 1) observed discharge, 2) simulated discharge where HBV is forced by observations; 3) simulated discharge where HBV is forced by ERA, and 4) simulated discharge where HBV is forced by the reference run. Mean annual average, mean annual maximum and mean annual minimum values are shown.

Dataset	M. A. Average [$m^3 s^{-1}$]	M. A. Max. [$m^3 s^{-1}$]	M. A. Min [$m^3 s^{-1}$]
Observed	24	200	2.57
HBV-OBS	21.8	157.3	2.97
HBV-ERA	39.4	305.1	5.1
HBV-REF	35.9	262.3	4.35

The model description was added to the first paragraph of Section 4: ***"In order to put the different model simulations into perspective, Table 2 shows three streamflow statistics (mean streamflow, mean annual maximum and mean annual minimum), for 1) observed streamflow, 2) streamflow as simulated with the HBV model forced with observed precipitation (Saint Hubert), 3) streamflow as simulated with the HBV model forced by ERA; and 4) streamflow as simulated by the HBV model forced by the reference data. Because the latter is not constrained by observations, only statistics can be compared, not the actual time series. All values are calculated over the same, 17-year period (1980-1996), as this is the period of overlap between all datasets. In Table 2, it can be seen that the HBV model with observed forcing is quite similar to the observations, especially in terms of mean streamflow, but that ERA and the reference forcing datasets produce considerably more streamflow. It should be noted here that all discharges in Table 3 are obtained using the same set of model parameters (Table 2)".***

3: According to Figure 3, relatively large errors were introduced after the bias corrections, especially during winter time period. The authors should evaluate how significant this is with compared to the climate change. Figure 6 (2062-2100) showed certain increases in winter streamflow compared to the reference values. I am afraid that some part of this predicted change is associated to the bias correction. I wonder what happens if the authors just inputted the original GCM (RCM) output without any bias correction.
Figure 3 shows the bias correction as carried out for the ERA dataset, to compare the four different bias correction methods. For the subsequent hydrological analyses, the ERA dataset was

not used anymore, but the bias corrected reference and scenario datasets were used. Because the biases of ERA and the reference are different, separate bias correction parameters have been determined. The overcorrection of precipitation in winter that can be seen in Figure 3, was not present in the correction of the reference dataset. It, therefore, does not explain the increase in precipitation (and streamflow) in the climate change scenarios. Besides, we only investigated the changes in streamflow and other hydrological variables with respect to the reference period. Because both the scenarios and the reference were corrected using exactly the same parameter values, the influence of a possible overcorrection is limited. We added to the end of Section 3.1: ***"It should be noted that the overcorrection of the precipitation that can be seen in Figure 3 is not present in the correction for the reference period (not shown)."***

Response to reviewer #3:
For clarity, the integral reviews are included in this response. Our answers to the questions raised by the reviewer are marked in italic text, and actual changes to the manuscript are indicated by a bold font.

This paper is about the impacts of climate change on different hydrological responses (fluxes, stores, droughts, flood peaks). A conceptual hydrological model (HBV) is forced with bias corrected high resolution RCM results for three SRES scenarios. The paper is well written and structured and within the scope of HESS. General comments, specific comments and some technical corrections are given below.

General comments
- An important issue in the paper is the bias correction of RCM (REMO) outputs using correction parameters obtained by comparing precipitation and temperature observed at a station and precipitation and temperature from a re-analysis data set (ERA 15). It is thus assumed that the correction parameters can be used to correct the RCM outputs as well. It is not clear to me whether this assumption is reasonable or not. Moreover, why did the authors not compare the station observed and REMO climatologies directly to obtain bias correction parameters. It is clear that station observed and REMO simulated time series can not be compared, but climatologies can be compared. This seems to be a much more direct approach for bias correction. In the current setting, differences between ERA 15 and the reference climate of REMO may necessitate another bias correction in order to get a realistic simulation of the hydrological responses under current climate conditions.
*The reviewer is completely right in that different bias corrections are needed for the ERA dataset (ERA15 downscaled with REMO) and the reference dataset (ECHAM5 downscaled with REMO). This is what we did. However, we chose to show the analysis regarding different bias correction methods for the downscaled ERA15 data only, and use the 'best' method to correct the reference dataset (and the scenarios). To make this more clear, we changed the formulation at the beginning of the last paragraph of Section 3.1, which now reads: "**Before forcing the HBV model, the ECHAM5 reference and scenario datasets have to be bias-corrected using the observations. Because the bias will most likely be different, new correction parameters are determined by repeating the analysis described above (using the method correcting the spatially averaged data), with data from the reference period and the observations.**"*

- Another important point is the novelty of this study. Although the authors did an interesting study which is a nice contribution to the climate change impact literature in hydrology, it is not completely clear to me what exactly are the novel points in this study. Therefore, first it is important to clearly state the research objective in the introduction. Furthermore, the added value of this study should be clearly indicated. Is the novelty of this study in the use of high resolution RCM results in a hydrological climate impact study? Is the comparison of four different bias correction methods the most important issue (but then the paper needs to be revised according to this objective)? Is the impact analysis for different responses (including different stores within the catchment) the most important contribution? Finally, this study should be placed in a wider context than just the Rhine and Meuse basins. For instance, have other studies also used this type of high resolution RCMs in hydrological impact studies (e.g. in the US or Japan), including the bias corrections?
*The most important contribution of the present paper is the use of extra-ordinarily high resolution of the climate scenarios, and given this high resolution, the ability to investigate smaller catchments than other climate impact assessments, that typically focus on large river basins. This is emphasized more in the introduction by changing the formulation at line 20 of page 7145: "**The final resolution of the data that is used in this study is very high (...) compared to other similar studies**", and at line 27: "**...allows to zoom in over a smaller catchment than the previously mentioned studies.**"*

Finally, we are not aware of any study using similar model resolutions in the U.S. or Japan, or of any study outside the Rhine or the Meuse where a similar bias correction method was employed.

Specific comments

* Introduction

- p7145, l4-6: The evacuation of several hundreds of thousands of people in the Netherlands took place in 1995 as a result of the near flood in the Rhine (not the Meuse).

While it is true that the Rhine nearly flooded at the same time, the peak volume in the Meuse in January 1995 was the highest ever (Chbab, 1995; see the manuscript for the full reference), and most people were in fact evacuated from the area in between the main Rhine branche in the Netherlands and the Meuse.

- p7145, l16: Which alternative methods can be used to downscale coarse scale GCM information and why has this dynamical downscaling method (i.e. a RCM) been used in this study?

Alternative methods would be of statistical nature, but this requires higher resolution observations. As we have only one station available, statistical methods are not an option. That is why we used a dynamical approach. We did not change the text, because we consider a broader discussion on downscaling methods not relevant for the introduction.

- p7146, l1-2: The fast response of the Ourthe is not only due to its hydraulic gradient, but also due to its limited groundwater storage and steep sandstone slopes (see p7147, l114-16).

*We changed the text at lines 1-2 as follows: "...**due to its hydraulic gradient, limited groundwater storage and steep sandstone slopes, a fast responding river that**...".*

- p7146, l6-17: Please try to complete the outline of the paper, e.g. only sub-section 2.2 and sub-section 3.1 are mentioned.

*The last part of the introduction was extended to contain the complete outline: "**After a description of the study area (Section 2.1) and hydrological model (Section 2.2), more details about this bias and the correction process are provided in Section 3.1. In Sections 3.2 and 3.3, respectively, the model calibration and the calculation of potential evaporation are described, and in Section 4, the climate change effects on the hydrology of the Ourthe will be discussed in terms of average fluxes and storages, as well as extreme peak flows and stream flow droughts. In Section 5, finally, the conclusions will be presented.**"*

* Study area, model and data

- p7147, l27-28: What is the difference between HBV Light Version 2.0 (Seibert, 2005) and the commonly used HBV96 model version (see Lindström et al., 1997, Journal of Hydrology, vol. 201, p272-288)? And why can the HBV Light Version 2.0 be used in this study?

*According to Seibert (2005), there are basically only two differences between HBV Light and HBV96, namely 1) the routing parameter MAXBAS can not only take integer values but any real value; and 2) a warming up period is needed instead of providing initial states. As we use a warming up period for every simulation and calibrate the MAXBAS parameter, using HBV Light does not negatively influence the modeling results. We added this to the modified manuscript, after the first sentence of Section 3.2: **"The only two differences between HBV Light and the other versions are in the model initialization, which should be done using a warming-up period in HBV Light, and a routing parameter that can take all real values instead of just integer values (Seibert, 2005)**.*

- p7148, l4-6: It is questionable whether all HBV model parameters are either measurable or significantly correlated to easily measurable catchment characteristics, see e.g. previous regionalisation studies using HBV such as Seibert, 1999 (Agricultural and Forest Meteorology, vol. 98-99, p279-293) and Merz and Blöschl, 2004 (Journal of Hydrology, vol. 287, p95-123).

*We agree that this is the case. It is, of course, a problem inherent to all conceptual hydrological models. We therefore chose to remove this sentence ("**The model's parameters....characteristics (..).**").*

- p7148, l21-22: Why has only one meteorological station been used? Although apparently only one station is available in the Ourthe catchment, stations from adjacent areas could have been used for comparison and interpolation purposes.
Currently, more and more studies are undertaken in the Ourthe (e.g, Hazenberg, submitted), and observations are available at more locations. As opposed to the station at Saint Hubert, no long time series are available, making comparison and interpolation difficult. In future studies, however, it would be interesting to take into account more stations.

* Methodology
- p7150-7155: Given the length of section 3.1, it could possibly be divided into some sub-sections.
We agree with the reviewer that Section 3.1 is too long. We divided it into three subsections: 3.1.1: Bias correction methods, where the employed method is described and the 4 different configurations of observations and model simulations are introduced; 3.1.2: where the results of these 4 methods are described; and 3.1.3: a short discussion where our selection of one of the four methods is motivated.

- p7150, l20-22: Are the methods of Shabalova et al. (2003) and Hay et al. (2002) similar to the method of Leander and Buishand (2007)? And why has the method of Leander and Buishand (2007) been chosen for bias correction in this study?
*The methods by Leander and Buishand (2007), and Shabalova et al. (2003) are identical, although the latter used the method to obtain climate scenarios instead of bias correction. The method of Hay et al. (2002) is different and employs a gamma-distribution. We employ the method by Leander and Buishand (2007) because it has been applied to the Meuse, and because it has been applied successfully using exactly the same climate data in the Rhine basin by Hurkmans et al., (2010). To make this clear, we added to the beginning of Section 3.1: "**We select the method of Leander and Buishand (2007), which is similar to that of Shabalova et al., (2003), because it has been applied to the Meuse, and also to the Rhine using the same data as the present study (Hurkmans et al., 2010)**."*

- p7151, l9: Why 73 blocks?
*In this we follow Leander and Buishand (2007), who determine the scaling parameters for each of the 73 five-day periods in a year (taking into account 30-days before and after the five-day period) to reduce the sampling variability. However, we slightly changed the formulation at line 19, which now reads: "...**uses the same 65-day windows as precipitation**..." instead of just "73 blocks".*

- p7151, l18-20: Was the method as applied to the Rhine basin also successful for temperature?
*As temperature is generally much easier to correct than precipitation, it was indeed successful for temperature - the bias was almost completely removed (see Hurkmans et al., 2010 for details). We added to the end of this paragraph (page 7151, line 20): "**The temperature bias was almost completely removed in these studies**."*

- p7151, l26-27: How are the different bias correction parameters per grid cell calculated in method 1?
*This was done by calculating the bias between all individual grid cells and the meteorological station at Saint Hubert, thus obtaining correction parameters for every grid cell. The formulation at line 26 was slightly changed and now reads: "....**comparing every individual uncorrected cell data with the Saint Hubert observation**."*

- p7152-7154: The methods and results are mixed here, please try to separate these two parts and

put the relevant results in the results section (additional sub-section about bias correction?).
While we agree that the section is a bit long (it was splitted as described at a previous point raised by the reviewer), we choose to keep the bias correction together, as the choice of the bias correction method is an intermediate result that will be used in the remainder of the study.

- p7152, l10-12: Figure 3, shouldn't the period be 1979-2003 instead of 1979-1996?
The caption is actually correct; whereas the period 1979-2003 was used to carry out the bias correction, Figure 3 was created using data from 1979 through 1996 for consistence with Figure 4.

- p7152, l25-29: Is this part again about precipitation or still about temperature?
*Indeed, this part is again about precipitation, which is not clear. We changed the formulation of the beginning of line 25 into: "**The overall mean precipitation of the uncorrected ERA dataset....**" to make this more clear.*

- p7153, l1-2: Has the goodness-of-fit of the Gumbel distribution statistically been tested?
We have not specifically tested the goodness-of-fit of the Gumbel distribution, but to give an idea of the uncertainty bounds we included the 95% uncertainty intervals. As we do not use the distribution for extrapolation but for a better visualization of the differences between the scenarios, we do not consider such a test necessary.

- p7153, l2: What is shown in each of the four subplots?
*The subplots show the 4 different bias correction methods, as indicated by the title of each panel. To make this clear, we added to the caption: "**The four panels show the four employed bias correction methods**."*

- p7153, l11: Is the GEV distribution also used for the annual maximum daily precipitation amounts?
We agree that the wording: "GEV or Gumbel" at line 11 is confusing. For the maximum daily precipitation amounts in Figure 5, Gumbel distributions were used. The same method was, however, applied to GEV distributions later on (Figure 9). Therefore, we decided to delete the words: "..GEV or.."

- p7154, l9: What is the relative difference for 15 mm?
Since on average the datasets have a precipitation amount of about 80 mm corresponding to a return period of 100 years, the relative difference would be about 19%. However, this difference is calculated using extrapolated Gumbel distributions and is thus relatively extreme when compared to the actual data points.

- p7155, l16: In HBV96 (see Lindström et al., 1997) besides elevation also land use is included in the zones. Did the authors also include land use?
We also included land use, in that we divided the catchment into six land cover types, each with their own crop factor. The catchment average crop factor is than the weighted (based on areal coverage) average of the six crop factors. This is described in the last paragraph of Section 3.3.

- p7155, l17-19: Why are these lapse rate values used? Are they default values in HBV?
*The used lapse rates are indeed often used in studies that employ HBV (Seibert 2005). To provide more information about these lapse rates and their background, we included in the revised manuscript (first paragraph of Section 3.2): "**Precipitation is lapsed with 10% per 100 m elevation difference, which is often used in studies using HBV (Seibert, 2005). By this lapsing, the dependency of precipitation on topography, which is present in the atmospheric forcing data, is reintroduced in the lumped model simulation.**"*

- p7155, l23: The number of calibration parameters is substantial. Did the authors consider the use of a subset of this number, e.g. based on previous HBV studies or a sensitivity analysis?

It is true that there is large number of parameters, but the current calibration set-up using all parameters and the genetic algorithm (Seibert, 2000) yielded good results and was computationally not that intensive. We therefore have not considered reducing the number of parameters.

- 7155, l24: Are the parameter ranges of Seibert (2000), who applied the HBV model to two Swedish catchments, representative for the Ourthe catchment? Why not using parameter ranges based on former Meuse and Ourthe studies?
The ranges provided by Seibert (2000) are quite wide and are, to our knowledge, not specific for Swedish catchments. We assume that the values for the Ourthe would be in the same ranges, and the good calibration and validation results confirm that. In addition, the calibrated parameter values have been compared to those parameter values that were found by Velner (2000) who also performed an HBV study on the Ourthe catchment. They were found to be very similar.

- p7155, l24-28: Has the calibration been carried out using a local or global optimisation algorithm?
The optimisation routine is in fact a combination of a genetic algorithm and a local optimisation routine (Seibert, 2000). Because the genetic algorithm starts with a large (here 50) number of randomly chosen parameter combinations, assumed to cover the entire parameter space, it can be seen as a global optimisation algorithm.

- p7156, l5-7: How are the two single objective functions combined into a multiobjective function?
"To explain how the two objective functions are combined in the calibration, we added the following to the revised manuscript (Section 3.2): ***"R_{eff} and |VE| are combined into the combined objective function R_c using Lindström (1997):***
$R_c = R_{eff} - 0.1|VE|$***"***

- p7156, l14-17: It is remarkable that the validation results are better than the calibration results. Please comment on this.
The better results for the validation period are related to the fact that the validation period was much shorter than the calibration period (8 versus 20 years). We added at the end of Section 3.2: ***While the higher R_{eff} for the validation period is remarkable, it should be noted that the validation period is relatively short compared to the calibration period (8 versus 20 years)****.*

- p7156, l16-17: Is the correlation coefficient an additional, third, single objective function?
The correlation coefficient was not part of the objective function, but merely an extra way of evaluating the model performance. To make this more clear, we added at the end of Section 3.2 (the one but last sentence) the following: ***"The correlation coefficient was not part of the calibration procedure, but is just an extra measure of model performance.".***

- p7157, l13: What is reasonable in this context?
Here, "reasonable" indicates that the values resulting from the described method are plausible for this area (qualitatively). As we do not have any other estimates of evapotranspiration we cannot quantitatively confirm that. Therefore, we replaced the word "reasonable" by "plausible".

- p7157, l13-15: Why hasn't the potential evapotranspiration not been corrected for biases directly?
The bias correction method (all methods actually) requires observations. We did not have direct observations of potential evaporation, nor the observations necessary to calculate potential evaporation (radiation, windspeed etc.).

* Results and discussion
- p7159, l7-12: What happens with the range between the 25th and 75th percentile in the future?
The entire range shifts in a similar way as the mean values (higher in winter, lower in summer). We chose not to show this because it would reduce the clarity of the figure. The bandwidth of the

reference was shown to put the changes as a result of climate change into perspective.

- p7161, l1-2: What is the definition of the annual maximum cumulative deficit volume?
We agree that we did not define this properly. We added to the text at line 3 of page 7161:
"Here, a deficit volume is defined as the total volume of water during the period that the streamflow remains below the threshold, i.e., the cumulative intensity until the threshold is exceeded again."

- p7161, l3-4: Has the goodness-of-fit of the Generalised Pareto distribution statistically been tested?
We have not specifically tested the goodness-of-fit of the GP-distribution. As we do not use the distribution for extrapolation but for a better visualisation of the differences between the scenarios, we do not consider such a test necessary.

- p7161, l25-26: Figure 9, has the goodness-of-fit of the GEV distribution statistically been tested?
We have not specifically tested the goodness-of-fit of the GEV distribution, but to give an idea of the uncertainty bounds we included the 95% uncertainty intervals. As we do not use the distribution for extrapolation but for a better visualisation of the differences between the scenarios, we do not consider such a test necessary.

* Summary and conclusions
- p7164, l5-7: These comparisons probably have been done at a global scale. Is the same true at regional scales, in particular for North-western Europe?
All comparisons that we are aware off were indeed carried out using global averages. We are not aware of any specific comparison for northwestern Europe, but according to IPCC (2007) ECHAM5 agrees with other GCMs on the general trends projected for this area.

Technical corrections
- p7153, l10: “precipitation” instead of “discharge”
This has been changed as suggested.

- p7155, l3: second “the” should be removed at the end of the line
This has been changed as suggested.

- p7157, l2: “the slope of the saturated water vapour pressure as a function of temperature curve” instead of “the slope of the saturated water vapour pressure curve”
This has been changed as suggested.

- p7158-7161: the sub-section numbers should be corrected
The section numbers have been corrected.

- p7158, 10: “hydrology” instead of “climatology”?
*We agree that "climatology" is a bit confusing and rewrote the sentence as follows: "...**scenarios have monthly values that are similar to those of the reference**....".*

Response to the editor

For clarity, the comments of the editor are integrated in this document. Text in italic font denotes our response.

Most importantly, the goodness of fit of the bias correction is tested extensively in the present manuscript but this tells little about the predictive performance. The wording "perform best" p. 7153 is a little misleading as far as I can tell - this should really read "fit best". In order to use the bias correction for the projections in a meaningful way, a split sample test would be needed, i.e. calibrating the bias correction parameters for one period and comparing it to data from a different period. As runoff models are much more sensitive to precipitation than to air temperature Fig. 4 should be replaced by an analogue figure of precipitation (for the validation period). Similarly, Fig. 5 should be replaced by validation period results. This split sample test should, ideally, be carried out for both rainfall and runoff (as recommended by reviewer 3), but doing it for rainfall would be a minimum. Without properly taking the uncertainties involved into account it is really difficult to appreciate what can be learned from the simulation studies. Also, the new contribution should be made crystal clear and the uncertainty needs a proper discussion, perhaps in the context of the issues raised by Bloschl and Monanari (2010, Hydrol. Processes).

With the words "perform best", we did not intend to say that that method has the highest predictive performance, but indeed the best fit in terms of monthly averages. To make this more clear, we replaced them by "fit best", as the editor suggests. For a description of how we revised the paper based on the three reviews, we refer to our responses to these reviews.

As a final remark, we agree that for an evaluation of the predictive performance of the bias correction methods a split-sample test would be necessary. Another manuscript, currently under review in HESSD (Terink et al., 2010), did this using exactly the same climate data and correction methods for the Rhine basin, with good results. We therefore expect that results in the Ourthe will be similar. We included this in the revised manuscript, by adding to the 3rd paragraph of Section 3.1: ***"In addition, Terink et al., (2010) evaluated the bias-correction method using split-sample tests, where the correction parameters where determined using one part of the dataset and validated using another. They found a reduction of the precipitation bias in both the calibration and validation part."***