Dear Anonymous Referee #1,

we cordially thank you for your review, constructive comments and also the positive feedback, which motivates us to further improve the study. Our initial replies (**blue**) to your major comments (**grey**) are now extended by the description of our implementations during the revision (**green**). Furthermore, we provide an overview of all changes.

Thanks for your reviewing efforts,

Benjamin Poschlod, Laura Sailer, Alexander Sasse, Anastasia Vogelbacher, and Ralf Ludwig

- Please explain why you choose to opportunistically use a high-flow calibration of the model here.

There is a background information to this choice: The very demanding setup of this hydrological large ensemble was conducted during the project ClimEx with its focus on floods. There is a second follow-up project ("ClimEx II") with focus on droughts and low flows. This study was conducted at the transition phase from ClimEx I to ClimEx II as a preparation for the phase 2 in order to explore uncertainties due to climate change and internal variability to provide insights for the hydrological model setup, calibration, bias adjustment, and required ensemble size. Hence, we aimed for a sub-selection of catchments at near-natural state with sufficiently good validation scores to explore uncertainties regarding low flows – this choice is opportunistic, but delivered the insights that we need for ClimEx II. As we think that these insights are also valuable to the scientific community, we aim for this publication.

Furthermore, the calibration process considers NSE, KGE, logNSE and RSR. The weighting is tailored to better represent high flows but also considers metrics that are relevant to represent low flow events. The resulting calibration shows a sufficiently good performance for low flow conditions. We will briefly elaborate on this choice in the article.

➔ We added this reasoning to the manuscript in section 2.2 (L169 – 177). Furthermore, we added the validation of L7Q and event duration in the Supplement (Figs. S2 & S3) showing that the simulations are representing the statistics of observed droughts rather well. L7Qs are too wet by only 5 – 10% compared to discharge observations. Event durations are well captured for the Ammer in winter and summer, while they are slightly overestimated in the Wörnitz (summer) by 5 – 30 days. In sum, we have explained why we opportunistically chose the high-flow calibration and have shown the suitability of the simulations for low flow assessments.

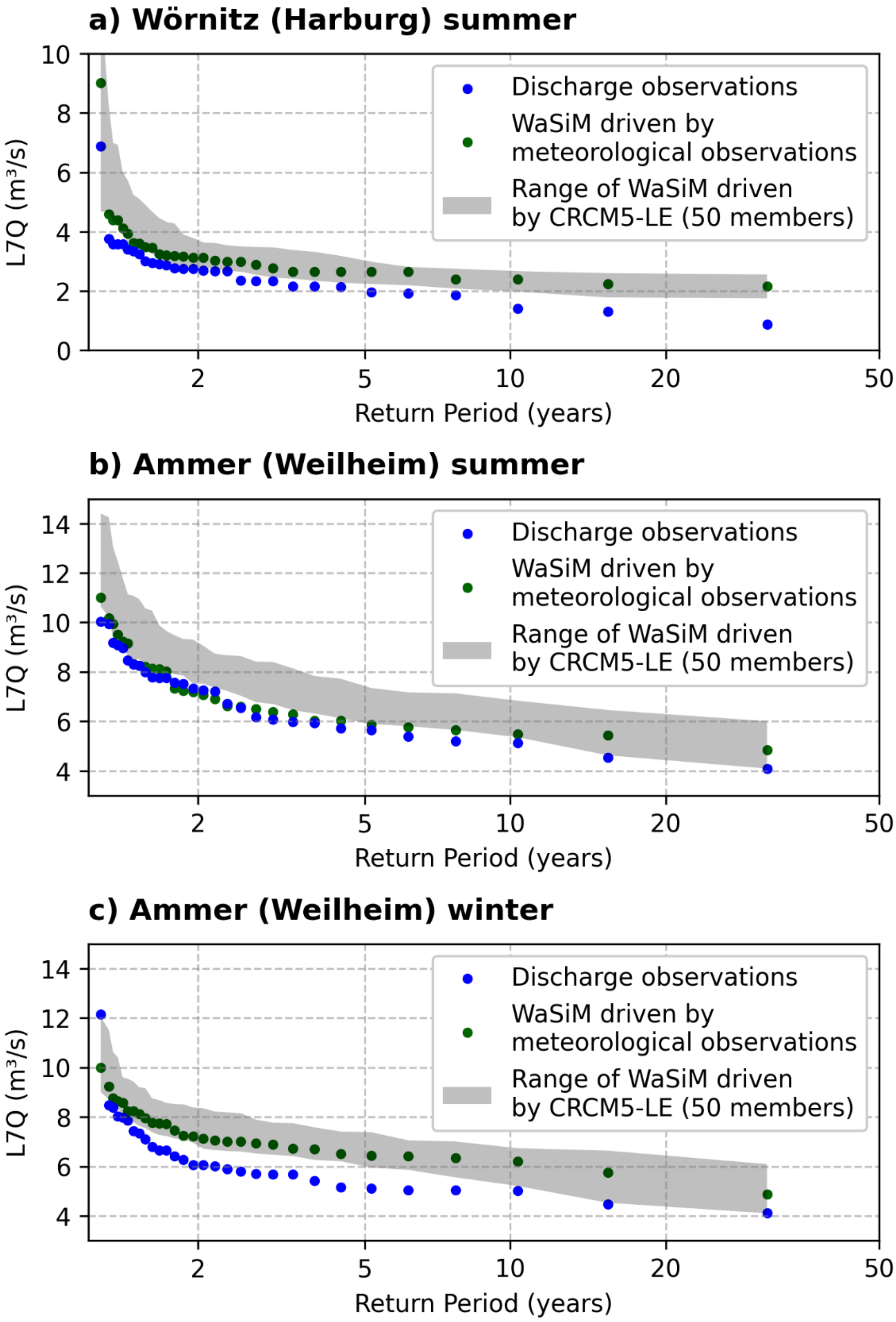

Figure S2: Evaluation of the L7Q during the reference period for the Wörnitz (a) and Ammer in summer (b) and winter (c). The return periods are plotted via Weibull plotting positions. Blue dots are based on discharge observations in the period 1981 – 2010. The green dots show the WaSiM simulation driven by meteorological observations. The grey area represents the full range of the 50 members of the WaSiM-LE simulations driven by the CRCM5-LE.

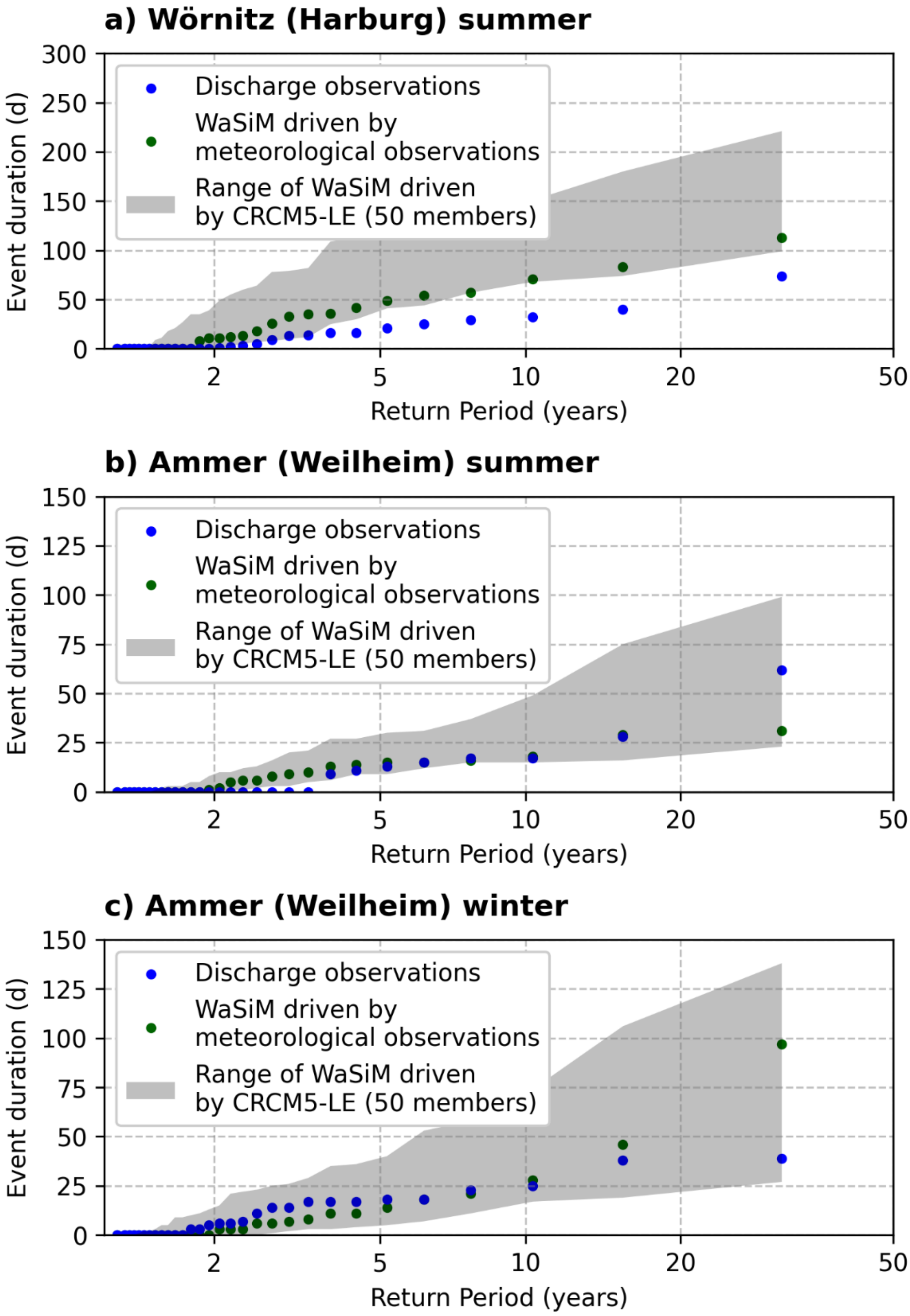

Figure S3: Evaluation of the event duration during the reference period for the Wörnitz (a) and Ammer in summer (b) and winter (c). The return periods are plotted via Weibull plotting positions. Blue dots are based on discharge observations in the period 1981 – 2010. The green dots show the WaSiM simulation driven by meteorological observations. The grey area represents the full range of the 50 members of the WaSiM-LE simulations driven by the CRCM5-LE.

- I definitely need a paragraph on that here, as you also address the bias adjustment in the limitations. All all forcins corrected or only O, T? Or do you use a Wasim configuration that needs only P and T?

The WaSiM setup requires precipitation, temperature, radiation, relative humidity and wind speed. All these climate forcings are bias adjusted via quantile mapping. We will add a paragraph on bias adjustment and the forcing variables in section 2.2.

➔ We added a more detailed description of the bias adjustment, the meteorological reference dataset and the choice of the CRCM5-LE as driving climate in section 2.2 in L135 – 148.

- a validation of L7Q and of the duration of events in the reference period should be included

Beyond the flow duration curves (Fig. S1), we will provide a validation of L7Q and event duration for the reference period. There, we will compare discharge observations, WaSiM simulation driven by meteorological observations and the 50-member WaSiM large ensemble driven by the bias-adjusted CRCM5-LE.

➔ See answer to the first point. Evaluation is added.

➔ All minor comments from your PDF annotations (wordings, additional references, additional figure versions in the Supplement) are addressed as proposed.

Dear Anonymous Referee #2,

we cordially thank you for the constructive and motivating review. Our initial replies (**blue**) to your major comments (**grey**) are now extended by the description of our implementations during the revision (**green**). Furthermore, we provide an overview of all changes.

Thanks for your reviewing efforts,

Benjamin Poschlod, Laura Sailer, Alexander Sasse, Anastasia Vogelbacher, and Ralf Ludwig

The study by Poschlod et al presents an analysis of future hydrological drought conditions in two catchments in Bavaria, Germany. The authors applied a single-threaded model chain with one climate model, one bias adjustment method, one RCP, one hydrological model, but driven by 50 ensemble members of the GCM. The future projections show a very strong intensification of river droughts.

The paper is very well written and I particularly like the innovative analyses and combination of methods to characterize the drought analysis. Also the figures are very nicely presented and convey the message very well.

However, I have some major comments which the authors hopefully find useful to improve the manuscript.

General and major comments:

The analyses and figures are of high quality. But it is a lot to digest and the dimensionality of some figures is high. I suggest to help the reader as much as possible to grasp the results efficiently. One point that would help, would be to mention in the figures (or caption) how many samples there are per group/class/bin/diagram (whatever is applicable), etc. (e.g. such as 50 dots per climate period in Fig 11).

Great idea, thank you. We will extend the figure captions to better communicate the dimensionality and explain the respective uncertainty assessments (confidence intervals).

- ➔ All captions extended

I generally like the additional analyses you show in the discussion. But I find it unusual that these weren't introduced in the methods and that these 'results' are shown in the discussion.

From my point of view, most of what you present until chapter 5.3 would rather be results and after that discussion. This requires including how the results were generated in the methods (e.g. it is quite unusual to introduce the Budyko Framework in the discussion with an equation or Ann van Loon's definition of winter droughts; and Figure 10 is an analysis of the climatic data which would be complementary to Figure 2). I am also open to merge the Results and discussion section into one section, though it might get very long.

We'd revise the article so that the additional analyses will receive their own sections within the Result chapter (so sections 5.1 and 5.2 will change to 4.4. and 4.5). The description of the Budyko framework will be added to the Methods section.
We would keep van Loon's categorization of winter river droughts in the new section 4.5, as it

describes not really an applied method but represents an explanatory introduction, which delivers context to better interpret the results of Figs 12 & 13.

- ➔ Structure changed accordingly

What I am missing in your discussion and conclusions is mentioning the implications on the hydrology of the region and contextualizing this given you chose one model/scenario. The changes you project are extremely significant, e.g. given the change in Return Periods of extreme drought events. The socio-economic, water availability and ecosystem implications that these conditions have would be enormous. I think it is important 1. to highlight that such changes are in the realm of possibility given the RCP8.5 scenario but 2. to contextualize this given that you choose the most extreme scenario and one of the hottest models in the CORDEX family.

Thanks for this suggestion. In the conclusion, we will add a paragraph on the implications of our findings:

- Contextualizing that RCP8.5 is “worst-case”, but the changes already occur in 2040-2069. This equals a global warming of slightly above ~3°C, which is realistic to be reached end of century (with less drastic emissions).
- Also contextualize the model uncertainty of the RCM (where CRCM5-LE is on the hotter & drier end during summer), but highlight that the climatic drivers of the hydrological results (given in Figs. 2, 8, 10, 11, 12, 13) can also be assessed directly in the climate models.
- Conclusions on the hydrology of the two regions with strong effects on river temperatures, ecology, and water availability.

- ➔ Based on the recommendation by the editor, we added a paragraph/section in the discussion (new section 5.3). We hope that global warming as reference framework can deliver the necessary context for the interpretation of the results and their severity.

Specific comments:

l.122 As the one model choice significantly impacts your results, I suggest to add more information here:

- Why was this model chosen?

- give a 1-2 sentence explanation what the 50 ensemble members represent and how they were developed (permutation of boundary conditions when driving the GCM, which ones, ...)?

- suggest mentioning that it is from the CMIP5 model stage

- Are there studies available that evaluated how CRCM5 ranked in comparison to other GCMs in terms of temperature (hotter, colder) and precipitation (drier/wetter?) - I had this question at this point. You give the answer in the discussion, which I think is too late

- short justification/reasoning behind using RCP8.5 (Again, I think at the current location in the discussion, this is too late)

We will extend section 2.2 with information on the RCM selection, the SMILE setup, CanESM2 as part of CMIP5. We will provide the context of RCP8.5 and CRCM5-LE compared to CORDEX already at this stage.

- ➔ We added the respective explanations already in the section 2.2 of the revised manuscript.

l.134 I think the WASIM calibration would have benefitted from a more targeted calibration and validation on droughts. Any reason why 2003 was left out and why stopping at 2010 given the most severe droughts occurred more recently?

We agree – the calibration process was driven by the focus of this project on floods. However, the presented results offer valuable insights for the capabilities of such a hydrological SMILE regarding low flows. In a second project, it is planned to implement a new bias adjustment, a complete WaSiM re-calibration/validation based on the insights of this study.

- ➔ We added a more detailed explanation on the opportunistically chosen high-flow calibration (L169-177). Furthermore, as suggested by Reviewer #1, we added an evaluation of L7Q and event duration for the whole reference period, comparing discharge observations, WaSiM driven by meteorological reference, and WaSiM driven by the CRCM5-LE. The according figures are in the Supplement (S2 & S3) and justify the usage of the simulations for low flow investigation.

l.176ff The methods in this chapter are too theoretically described for me to understand what the outcome of the analysis actually looks like. I think an example, graphical in diagrams / a flowchart or a more descriptive phrasing of the outcome would help to easier understand the analysis - this would include the additional analyses in the discussion as well.

Thanks for this suggestion. We will add a flow chart in section 3.2. that summarizes the extreme value theory workflow and visualizes schematic outputs, so the reader can better imagine the outcome of this analysis at this stage of the article already.

- ➔ Flowchart added as Fig. S4. We hope that this alternative visualization helps the reader to easier understand our extreme value statistical analysis.

l.299ff nice analysis and visualization. However, can you explain the white (no data?) areas in the diagrams? You earlier mentioned summer and winter as May-Oct and Nov-Apr, but the colored circles extend beyond these boundaries and also don't exist/are shorter for others.

The white areas in Figure 5 are a result of the event sampling. We divide the analysis in summer and winter half years (May-Oct and Nov-Apr). Each low flow event is assigned either as “summer” or “winter” event, based on the majority of the event duration. Hence, an event might start in late summer (September) and last until March – then it is counted as “winter event”, as the majority of Sept-March belongs to the winter half year.

Thereby, Fig 5a shows only summer events for the Wörnitz, 5b shows summer events for the Ammer and 5c shows winter events for the Ammer (as winter events in Wörnitz are not relevant due to neglectable snow dynamics and winter as wet season). If you’d combine 5b and 5c, you’d get a full year coverage without white areas. However, we opted for a separate visualization due to different drivers and different change signals in a warmer climate.

l.333 it is noticeable that the CUR 'overtakes' the NF and reaches the FF duration for RP's > 50. Did you look into why this is happening?

We guess, you're referring to Fig. 6f (winter low flow durations in Ammer). Our take-away message for this analysis is "During winter, no clear changes are detected, only a tendency for longer events in the future periods." – We argue that the confidence intervals overlap. Physically, we'd interpret this as climate variability. One can see that there are a few long-duration events in the CUR climate (also nicely visible in Fig. 5c), which are not there in the NF.

In 5c, we see that those events already start during the preceding summer and cover the winter half year until Feb/March. We will add a brief sentence on this in the revised article and diagnose the conditions of these events in the CUR period.

We would, however, also rephrase: "During winter, no clear changes are detected, only a tendency for longer events in the far future period." to be more precise.

- ➔ Rephrasing added

l.339 again, nice analysis. I find it hard to compare the return periods between the scenarios. Would it make sense plotting a few more isolines (light grey) for being able to compare this better (not sure, might also make the figure too busy)?

We think that the 10a and 100a isolines represent a good selection of moderate and rare extremity – more isolines will make the already complex figure even busier. We will think about an alternative (e.g. a supplementary figure, where only the isolines of the four time periods are in one figure for a direct comparison).

- ➔ We added Fig. S6 for an overview of bivariate return level changes

l.485 When comparing with other studies, I suggest to also highlight that you used the most extreme RCP and only one climate model. For a comparative analysis on the Ammer catchment, you could check the study by Kiesel et al. 2019, Ecol. Engineering, who seem to have found similar impacts on temporal shifts.

Thank you, based on the comments by the other reviewer, we will already highlight in chapter 2, that the combination of high-emission scenario and CRCM5-LE is more on the hot & dry end compared to EURO-CORDEX. Thanks for the study by Kiesel et al., which we will add to chapter 5.

- ➔ Added (the study by Kiesel et al. fits very well, thanks again).

Minor comments:

l.47-49 if these studies are directly linked to the enumeration, I suggest to place them directly behind the respective impact

Will be done.

→ Done

l.73 Check Clark et al. 2016, Characterizing Uncertainty, Curr Clim Change Rep. on uncertainty related to impact modelling

Thanks, will be added at this place.

→ Done

l.125 is this downscaling applied within your study?

The downscaling is applied and described in Leduc et al., 2019. The WaSiM simulations of this study are driven by these dynamically downscaled simulations.

l.133 which climate reference dataset?

It's described in the provided reference Willkofer et al., 2020. We will add a few details in the revised article.

→ Done

l.147-148 suggest to at least mention the bias correction method chosen and the main hydrological performance values.

We will add information on the bias adjustment (also mentioned by the other reviewer).

→ Done

l.170-171 I don't understand that sentence. What does the number of consecutive days have to do with the two metrics?

This sentence describes our definition of "event duration". We rephrase in the revised article:

"We further define the low flow event duration. Therefore, we count the number of consecutive days, where the 7-day mean discharge is below the ML7Q of the reference period 1980 – 2009."

→ Done

l.191 are you also applying back transformation?

Oh yes, thanks. We will add this for clarity.

→ Done

l.206 What is the 'threshold' u - is this the earlier described discharge value obtained from L7Q and ML7Q?

The threshold u applies only for the Generalized Pareto framework, used for peak-over threshold sampling. It is applied for the event duration in this study. We will further clarify this in the revised article.

→ Done

l.260 WASIM is a spatially distributed model. I assume you calculated averages over the model's hydrological objects upstream of the two gauges?

Exactly. We will add this information to section 2.2.

→ Done

l.284 average drying tendency and in the next sentence annual median. Clarify what exactly is averaged.

Thanks, the first sentence should already read “median drying tendency”. Will be revised.

→ Done

l.527-531 this seems unconnected to the previous sentences, but I think you want to express that the model is suitable to capture that process. Suggest to make that clearer.

Will be clarified, thank you.

→ Done

l.580 I suggest not to refer to the results again in the conclusion in this level of detail.

Will be removed.

→ Done