Replies to the comments by Referee #2

We would like to thank the anonymous referee for his/her interest and the comments on our manuscript.

Below, reviewer comments are in italic font and our replies are in normal font.

In their manuscript "Why does a conceptual hydrological model fail to predict discharge changes in response to climate change?", D. Duethmann et al. investigate possible reasons for the deficiencies of a conceptual hydrological model (HBV model type) in reproducing observed changes in discharge as a response to changing hydrometeorological conditions in 156 catchments in Austria. The authors set up hypotheses that belong to three groups of possible causes: (i) data problems, (ii) problems related to model calibration, and (iii) problems related to model structure. They test these hypotheses by comparing simulations generated by modified versions of the model according to the hypotheses against a baseline model. Data problems and model structural problems with respect to vegetation dynamics have been identified as the most relevant causes for the model deficiencies.

General comments:

The paper is well written and well structured. It addresses a relevant scientific question and provides valuable insights for hydrological modelling under changing climate conditions which surely is of broad interest. Still, I have a few comments and suggestions that may further improve the manuscript:

The results are mostly presented as averages over the investigated 156 catchments. I wonder if we could not learn even more if also the statistical and/or spatial distributions will be presented. As stated in the discussion, reasons for hydrological model deficiencies can be very site specific. By including more of the variability between the catchments, prominent cases could be identified which do not (or particularly do) support the conclusions which are based on the mean of all 156 catchments. This may also feed the discussion on possible further causes for model deficiencies which have not been tested in this study. The modified model versions V2, V7, and V8 have led to the best improvements. Maybe it is worth showing another figure on these results in the same manner as Fig. 3 (or the modified version of Fig. 3). This could be a nice illustration of the key results of this study.

> We agree with the reviewer that many details are hidden by aggregating the results to annual means over all catchments. While we need to aggregate the results to a large extent due to the large amount of data, we will show more spatial patterns and distributions across catchments in the revised manuscript. In particular, we will include maps similar to Fig. 3 (c) for selected model variants as suggested by the reviewer. We will further show distributions across the 156 study catchments of bias and NSE for the baseline model calibrated in the different subperiods, complementing Fig. 4 that shows changes in the mean value of bias and NSE averaged across the study catchments, as suggested by Yan Liu, Veit Blauhut, Amelie Herzog, Tunde Olarinoye and Ruth Stephan (second short comment).

Specific comments:

Title: The title is catchy but also provocative since it suggests that conceptual hydrological models in general are not suited/justified for climate change impact studies, which is not correct.

We will revise the title, please refer to the comment #1 by David Post. The new title reads 'Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change?' By referring to 'a conceptual hydrological model' and not 'conceptual hydrological models' we intend to indicate that we have tested one and not several or more models. The problems we found when applying the HBV-based model over catchments in Austria may, however, also be relevant for other hydrological models and other regions (also see SC3 by Taehee Hwang).

P2, ll9-11: what is meant by "minimum requirement". Passing or failing the test? How is this determined?

Passing the DSST can be seen as a minimum requirement for models applied for climate impact assessments. The text will be adjusted to make this clearer.

P3, l25: Please provide references.

Will be added to the manuscript (Fowler et al., 2018;Fowler et al., 2016;Westra et al., 2014).

P4, l14: The numbers show comparatively large differences in elevation ranges. I wonder if this has any influence on the testing result. Are there any altitude-dependent differences in the results of testing the hypothesis? This partly corresponds to my general comment.

The relationship between catchment elevation and the trend of the gap between simulated minus observed discharge is not very conclusive (Fig. 1). On the one hand, catchments in the lowest elevation class (median catchment elevations below 400 m) show clearly lower deviations between simulated and observed trends. Furthermore, there is a slight increase of the gap between simulated and observed trends with elevation for median catchment elevations up to 1200 m. On the other hand, this tendency largely disappears when the gap between simulated and observed trends is normalized by the mean annual observed discharge, and the group of catchments with median elevations below 400 m is based on only 7 of the 156 catchments.

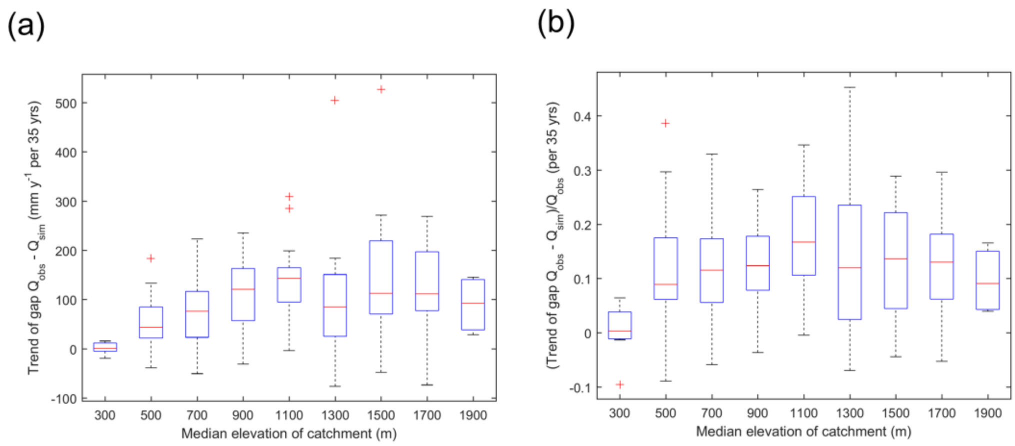

Figure 1 (a) Boxplots of the differences of simulated minus observed trends in discharge against median catchment elevation and (b) boxplots of the differences of simulated minus observed trends in discharge against median catchment elevation normalized by average annual discharge.

P4, Fig.1: When I look at this map, I am reminded to a paper that has identified (homogenous) hydrological regions in Austria (though it was probably with reference to flood types). Anyway, do the presented testing results show any systematic spatial differences regarding the major reasons for model performance losses or improvements? For the baseline model, Fig. 3 (c) presents a map in this regard. For the tested hypotheses, however, spatial information is not presented. I think, though, that this could be interesting. This also corresponds to my general comment.

We will show maps similar to Fig. 3 (c) for model variants V2, V7 and V8, as suggested by the reviewer in the general comments. This shows that using the precipitation data set P2 resulted in reduced gaps between trends of simulated minus observed discharge particularly for catchments with large trends in simulated minus observed discharge, whereas considering vegetation dynamics for the calculation of evapotranspiration resulted in a much more even effect between catchments. V8 combines both of these effects, reducing the trend of simulated minus observed discharge in most catchments with large reductions in catchments that showed large trends of simulated minus observed discharge in the baseline model.

P5, ll9-13: I remember from other regions and countries that their official meteorological data products are already corrected for potential undercatch. I am not familiar with the SPARTACUS data; I just want to be sure that no "double-correction" is performed here.

The SPARTACUS data are not corrected for undercatch (Hiebl and Frei, 2017).

P6, Section 2.3.1 could also make a reference to Table 1.

Will be added.

P7, l19, and P8, ll9-10: "(E3)" confuses me. Did I miss E2? On P8, E3 is compared to E2. Later, only results for E0-E2 are reported (e.g. Table 3). I assume that E3 is E2. Please check. Also, "than" instead of "tha" (P8, l9).

Thanks for pointing this out. This will be corrected.

P8, Eq.8: Is fbeta the same as fp? Otherwise, fbeta is not explained. Is the same objective function applied in Merz2011?

Thank you, yes this is the same and this will be corrected in the manuscript. In this study, we added a penalty for the volume bias in order to keep it low, which was not considered by Merz2011.

P9, l4: One more sentence on how the shuffled complex evolution algorithm works would be nice.

Ok, we will add a short explanation.

P9, l9: It could be highlight that the seven 5-year calibration periods have no temporal overlap.

Will be added.

P10, ll13-16ff: I agree that such problems will probably not affect many catchments. For selected catchments, particularly in mountainous areas, it still might be a cause for problems in calibrating and evaluation the hydrological model. Does the HZB provide information in this regard?

Information on abstractions and flow diversions is provided in the hydrological yearbooks (BMLFUW, 2015). Catchments where flow diversions were introduced before the beginning of the study period were included in the data set, since we did not expect large effects on simulated discharge trends. We excluded catchments where diversions were introduced during the study period.

P11, Figure 2: You may add to the figure caption to which number of stations P1 (P2) and T1 refer.

The data sets P1, P2 and T1 are based on a constant number of station series that extend over the entire period. The number of stations they refer to will be added to the text.

P14, l18: Does Esim refer to the model estimation based in Eq.2? Or does it refer to the difference between P-Qsim? Would it make any difference (also regarding the consideration of the same uncertainties that refer to the estimation of Ewb)?

The calculation of E_{sim} is described in Section 2.3.1 (P6, L16ff). For the baseline model, E_{ref} is calculated using Eq. 2. E_{sim} is then calculated as a function of E_{ref} and soil moisture. Thus, E_{wb} also includes storage changes, wheras E_{sim} does not. This will be pointed out in the manuscript. This difference is relevant at short time scales. For example, the large year-to-year variations of E_{wb} in Fig. 3 (b) are likely due to storage changes. The mean values over a 5-years subperiod and the trend over the entire study period is much less influenced by any storage changes. We will add more explanation to the text. We will also add an additional figure that shows the differences between precipitation minus runoff for observations and simulations to the supplement.

P15, ll3-6: How has this been done?

We will add some more information on how we calculate changes in simulated storage.

"For this, we analysed the sum of all simulated storages, i.e. soil moisture store, upper zone and lower zone groundwater store and snow water equivalent, and calculated trends of annually average values (based on hydrological years). Trends in simulated storage changes were, on average over all catchments, 8 ± 20 mm over 1978–2013. This shows that the overestimation of the discharge trend is not generated by an opposite trend in a storage component. Of the simulated storage groundwater is the largest component. Small changes in simulated storage are also in agreement with no consistent large scale groundwater changes in the observations (Blaschke et al., 2011; Neunteufel et al., 2017)."

P16, Figure 3 (and others): I see that these figures are designed to match the presentation by Merz2011. However, I think that by presenting only the mean a lot of information is hidden. Boxplots or additional maps (as in Fig. 3) would be more appropriate. This also refers to my general comment.

We will add further maps similar to Fig. 3 (c) for selected model variants, as suggested by the reviewer in the general comments. We will further add violin plots showing distributions of the bias and NSE complementary to Fig. 4.

P17, Figure 4: Do the seven 5-year calibration- and evaluation periods show any marked differences in terms of hydro-meteorological conditions?

Yes. Over the study period, precipitation, air temperature and E_{ref} increased, as shown in Fig. 5 (a–b) and Fig. 7. We will add a description of the changes in the hydro-meteorological conditions to Section 2.1.

P18, Figure 5 (also Figure 7): You could add to the figure caption that the impacts of altering these variants in the hydrological model are summarized in Table 4.

Results of these model variants are further also shown in Table 5 and Fig. 6. We will think of adding cross-references to the figure captions.

P19, Figure 6: You may indicate that Fig. 6 (a) is the same as Fig. 4 (a).

Will be added.

P20, Table 5: This table (in combination with Table 2) is really nice since it provides a good summary of the tested hypotheses. Maybe the result of V8 can also be summarized here.

We did not include V8 in Table 2 or Table 5 because it was not part of the original set of hypotheses. However, we will provide the same information that we provide for the other model variants in Table 5 in the text (where it is missing in the current version).

P21, ll25-27ff: It could be emphasized more clearly why you choose to combine V2 with V7 to V8.

Ok, will be added.

P22, Discussion: The discussion reads nicely, and I agree with the main conclusion that the consideration of interrelations between climate, vegetation, and hydrology is an important further step for hydrological modelling in transient climate. Still, I have a few remarks and thoughts regarding the discussion.

a) The discussion in its current form gives the impression that model structure deficiencies regarding vegetation dynamics is the most important reason for model performance deficiencies in transient climate, although fixing problems in the precipitation data have led to improvements of similar magnitude. Finally, it could be highlighted that the combination of both approaches has led to the largest improvement (reduction in mismatch by about 95%).

Thanks for the feedback; it was not our intention to give this impression. We will adjust the discussion to avoid giving this impression. We will also pick up the results of combining the modifications for the precipitation data and considering vegetation dynamics.

b) For good reasons, model structure improvements are restricted to incorporating vegetation dynamics only. Still, what could be further model structural issues that cause model performance losses in this particular study region? Maybe it is worth highlighting that glaciated catchments have not been considered here. Have they been considered by Merz2011?

Good point. We will mention possible model structural problems with respect to changes in glacier extent and glacier volume in the discussion.

P23, ll1-3: Considering my complaint regarding the title: This is a good example for the benefit of a conceptual hydrological model. By applying a rather simple approach, vegetation dynamics can be considered to some degree for hydrological simulations in changing climates.

Please see our comments regarding the title above. While we tried to include changes in vegetation dynamics into a conceptual hydrological model in this study in order to derive a first order estimate of the possible effects, changes in vegetation dynamics are not considered by most conceptual hydrological models.

P24, l1: I think this refers to V2 which indeed had a considerable effect.

V2 had a considerable effect when compared to V0. However, it builds on V1 and the differences between V2 and V1 are small and not significant. This will be clarified in the manuscript.

P24, l4: One ")" is missing.

Will be corrected.

References

BMLFUW: Hydrographisches Jahrbuch von Österreich 2013, 121. Band - Daten und Auswertungen, Wien, 2015.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R., and Zhang, L.: Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement, Water Resour. Res., 54, 9812-9832, 10.1029/2018wr023989, 2018.

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., and Peterson, T. J.: Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, Water Resour. Res., 52, 1820-1846, 10.1002/2015wr018068, 2016.

Hiebl, J., and Frei, C.: Daily precipitation grids for Austria since 1961—development and evaluation of a spatial dataset for hydroclimatic monitoring and modelling, Theor. Appl. Climatol., 10.1007/s00704-017-2093-x, 2017.

Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M.: A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resour. Res., 50, 5090-5113, 10.1002/2013wr014719, 2014.