Authors' response - Manuscript "Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change?" by D. Duethmann, G. Blöschl and J. Parajka

---

Replies to the comments by Mojca Sraj

We would like to thank Mojca Sraj for her interest and for her comments on our manuscript.

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

General comments

The paper investigates the reasons for failure of the conceptual hydrological model predicting changes in discharge as a response to observed increases in precipitation and air temperature for 156 catchments in Austria. The authors considered three groups of possible causes, namely data problems (precipitation, temperature), problems related to the model calibration (length of calibration period, objective function), and problems of the model structure (ET calculation method, vegetation changes). Hypotheses of the possible causes were evaluated using simulations with modifications of the baseline model. The paper is in the scope of the journal. It is well written and structured. The data seem to be of appropriate quality. References are up to date and appropriate. There are, however, some areas that require minor corrections for further improvement.

*1. Penman-Monteith method – Please, check the calculation and the equation for the net radiation at the crop surface (Eq. 4). Net radiation (Rn) is the difference between the incoming net shortwave radiation (Rns) and the outgoing net longwave radiation (Rnl), and Rns is derived from the balance between incoming and reflected global radiation (Rs) given by (1-_)*Rs (see Allen et al., 1998).*

> Thank you very much for pointing this out. This will be corrected. The error only occurred in the manuscript (not in the calculations).

Specific comments

1. Page 1, line 21: I would suggest being more specific in defining the size of the impact. How much is "little"? Is it negligible?

> We will add "(less than 5 mm yr^{-1} per 35 yrs)" to be more specific.

2. Page 3, line 29: Using abbreviation Merz2011 for the reference is unconventional. I would suggest to use the usual way of citing, namely Merz et al. (2011). This issue should be corrected throughout the document.

> Since Merz et al. (2011) is referred to very often, it seems a useful abbreviation and we will adjust this according to any guidance by the journal.

3. Page 7, line 16: Please, check equation 4. It does not seem ok to me.

See above.

4. Page 7, line 19: Modified Eref calculated using a variable surface resistance based on changes in a satellite-based vegetation index should be marked as E2 (not E3) in order to be consistent with Table 3.

Thank you for pointing this out. This will be changed.

5. Page 8, lines 9-10: Abbreviations E1 and E2 should be explained when first mentioned. Furthermore, correct the numbering of E1, E2 and E3 to be consistent with Table 3.

E0 and E1 are defined in Eq. 2 and Eq. 3. The numbering will be corrected to be consistent with Table 3.

6. Page 14, Table 3: It would be useful for readers to include the exact years of the 5-year calibration period in the table or in the table caption. Is it the first 5 years of the considered data or any other? Is it the same for all model variants?

Model calibration and the calibration periods are described in Section 2.3.3. We used seven 5 year calibration periods (based on hydrological years), during 1978−2012. As a modification, we also tested using a 25-year period as calibration period (1978−2002). We will change the header of the respective column in Table 3 from "Calibration period" to "Length of calibration periods".

7. Page 15, lines 3-6: Analyses of simulated changes in storage should be explained in more detail since they are mentioned only in this paragraph.

We will add some more information on how we calculate changes in simulated storage and will extend this paragraph in the manuscript.

"For this, we analysed the sum of all simulated storages, i.e. soil moisture store, upper and lower zone subsurface 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 simulated storage. Small changes in simulated storage are in agreement with no consistent large scale groundwater changes in the observations (Blaschke et al., 2011; Neunteufel et al., 2017). "

8. Page 15, lines 13-16: As seen from Fig. 4b, an increase in model performance loss with increasing distance of evaluation periods from the calibration period could be observed in almost all cases, regardless of the calibration period. Please, rewrite the sentence.

Will be rewritten: "In many cases, model performance decreases with increasing distance between the calibration and the evaluation period, particularly for model evaluations in subperiod S1 and S2."

9. Page 20, Table 5: It would be useful for readers to add a corresponding model variant to each individual result.

Good idea, will be added.

10. Page 21, line 3: Please add which 5 years and 25 years were used for calibration of the mentioned model variants.

Will be added.

11. Page 21, line 21: It would be useful for readers to add a model variant in brackets.

Will be added.

Technical corrections

1. Page 15, line 18: It should be "is reversed".

(Note, the comment apparently refers to page 14, line 18). Ok, can be changed.

2. Page 15, line 14: Correct the structure of the sentence.

See above, we will rewrite the sentence. "In many cases, model performance decreases with increasing distance between the calibration and the evaluation period, particularly for model evaluation in subperiod S1 and S2."

3. Page 24, line 4: Bracket is missing at the end of the sentence.

Thanks, will be corrected.

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 Figure 2c for selected model variants as suggested by the reviewer as Supplementary Figure S6a-c. We will further show distributions across the 156 study catchments of bias and NSE for the baseline model calibrated in the different subperiods as Supplementary Figures 4 and 5, complementing Figure 3 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 and observed discharge is not very conclusive (Figure 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.

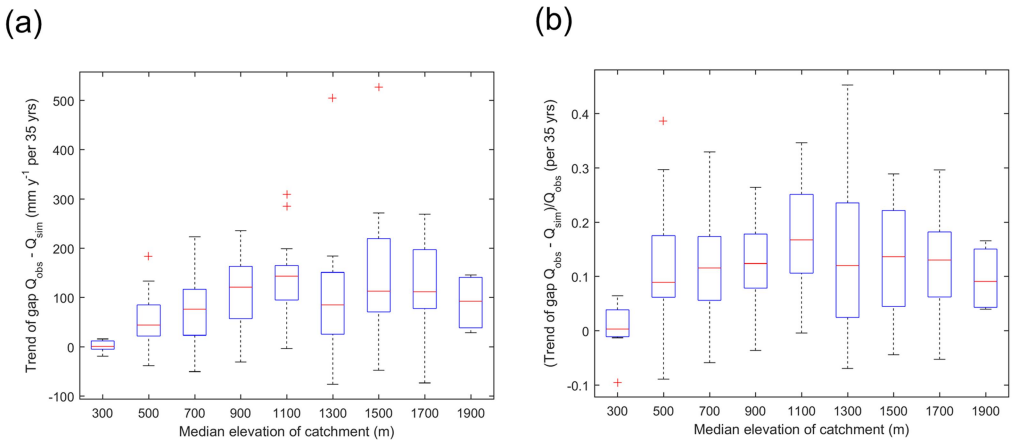

Figure 1 (a) Boxplots of the differences of simulated and observed trends in discharge against median catchment elevation and (b) boxplots of the differences of simulated and observed trends in discharge against median catchment elevation normalized by average annual discharge.

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 Figure 2c for model variants V2, V8 and V9 (V2, V7 and V8 in the original manuscript) as Supplementary Figure S6 (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 and 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. V9 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. 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 Figure 2b are likely due to storage changes. The mean values over a 5-year 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 and runoff for observations and simulations to the supplement (Supplementary Figure S3).

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. 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 Figure 3c for selected model variants, as suggested by the reviewer in the general comments, to the supplement (Supplementary Figure S6). We will further show violin plots with distributions of the bias and NSE in Supplementary Figures S4–S5 complementary to Figure 3.

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 Figure 4 a–b and Figure 6. 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.

Yes, this could be an idea. In the end, it was not added as there would be more information needed (i.e. which model variant links to which precipitation or air temperature data set), and because this is just one out of many possible cross references between the figures and tables (and adding all of them seems a bit much).

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

Figure 5a (6a in the original manuscript) has changed and we added a reference to Figure 3a (Figure 4a in the original manuscript).

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 V9 (V8 in the original manuscript) 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 was missing in the original 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 (see first paragraph in '4 Discussion').

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.

"Changes in glacier volume may cause deviations between simulated and observed discharge trends if not accounted for by the model. Therefore, glacier covered catchments were excluded in our study. Model structural deficits with respect to glacier dynamics may be responsible for further deviations between simulated and observed discharge trends in the study by Merz2011, which did not exclude glacier covered catchments, although the total glacier cover of Austria is small (0.5 %; Fischer et al. (2015))."

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 (we added "when compared to the simulation using the same precipitation data without undercatch correction.").

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.

Replies to the comments by David Post

We would like to thank David Post for his generally positive comment and for his suggestions for improving our manuscript.

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

This is a nice example of a study that attempts to determine exactly what it is about rainfall-runoff models that means they are not capable of predicting well runoff under changed climate conditions. One major thing that would improve the paper would be to quantify for the reader what the change in relevant hydroclimatological characteristics during the verification period actually are. The authors state that the area was subject to significant climate changes, but do not tell us what these actually were. Were the evaluation periods drier/hotter? If so, by how much. What were the relative runoff coefficients?

> Thanks for this advice. Over the study period, precipitation, air temperature and E_{ref} increased, as shown in Figure 4a–b and Figure 6 (Figure 5a–b and Figure 7 in the original manuscript). We will add a description of the changes in the hydro-meteorological conditions to Section 2.1.
>
> "Over the period 1977–2014 annual precipitation increased by 32 ± 23 mm yr^{-1} or 2.4 ± 1.7 % per decade (based on undercatch corrected SPARTACUS data), air temperature increased by 0.45 ± 0.09 °C per decade and global radiation increased by 5.1 ± 0.9 W m^{-2} per decade on average over the study catchments. In contrast, discharge did not show strong trends and the average trend over the study period was 0.2 ± 3.1% per decade (Duethmann and Blöschl, 2018)."
>
> Relative annual runoff coefficients (based on observed discharge data and undercatch corrected SPARTACUS precipitation data) vary in the range of 0.22 and 0.86 between the study catchments. The runoff coefficients significantly decreased ($p \leq 0.05$) in 29 and significantly increased in 4 of the 156 study catchments. (A large number of catchments showed insignificant decreases, probably due to the large interannual variations of the annual runoff coefficients).

Despite these issues, I have just three comments on improving the paper:

1. The title is misleading. Almost every model will predict discharge changes in response to climate change. The questions is why they do not 'accurately' predict discharge changes? The addition of a qualifier like 'accurately' would be useful.

> We will revise the title and add 'correctly' as a qualifier. The title then reads 'Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change?'.

2. Changes in anthropogenic influences are largely ignored as the authors claim that the catchments are largely unregulated and existing diversions were introduced before the beginning of the study period. I would question this. While the diversions may be in place before the beginning of the study period, are there operating rules related to this diversions which may vary from year to year, for example allowing larger diversions during periods of low flow (or vice-versa). I ask as we have identified catchments in Australia that not only behaved abnormally (gave lower than predicted yields during the Millennium drought), but that have not returned to 'normal' yields post-drought. One hypothesis for this is that farmers sank groundwater bores to access an alternative water supply during the drought when they were unable to pump from surface water. Any lowering of the groundwater table resulting from this activity would obviously lead to lower than expected yields. Once this 'sunk cost' had been incurred, there would be no benefit to farmers in ceasing the pumping of water from these bores, thus they may still be doing so post-drought. Such anthropogenic influences are of course hard to determine (and even harder to quantify), but the authors would do well to keep them in mind.

> Changes in private abstractions by households or farmers are indeed difficult to get hold of. In Austria, water abstractions for irrigation are much less important than in Australia. With respect to anthropogenic impacts on water resources, diversions for hydropower generation are much more relevant than abstractions for irrigation. Irrigation in agriculture is most relevant in small areas east, southeast and northwest of Vienna, where estimated irrigation amounts of agricultural areas exceed 10 mm/year (BMLFUW, 2011). In most parts of Austria, estimated irrigation amounts of agricultural areas are less than 1 mm/year. The fraction of arable land in our study catchments is only small (5% on average over the catchments) and the catchments in our study hardly overlap with those areas where agricultural areas receive a large amount of irrigation. At this stage we therefore assume that changes in irrigation amounts are not a major source for the deviations between the simulated and observed discharge changes.

3. The assessment that problems with the model calibration can be the source of the poor performance during the evaluation period is a good one. In particular, that processes that are relevant in the calibration period are not present (or 'activated' to use the author's terminology) in the calibration period. I am not sure that extending the calibration period from 5 to 25 years will actually evaluate whether this is the case. It may be that these processes will be seen in the 25 year period, but it may not. One thing that could be done is to compare the model that is calibrated on the evaluation period (or perhaps part of it) to the model that is calibrated on the calibration period. If different processes are dominant in the evaluation period, this would be seen in how these models perform on an independent data set.

In the original version of the model, the model is calibrated in a 5-year period and then evaluated in 6 other 5-year periods. For example, the model calibrated in 1978–1982, is evaluated in 1983–1987, 1988–1992 and so on. If this model performs well in calibration (e.g. in 1978–1982) but performance is worse in evaluation (e.g. 1983-1987), this might be due to a process that was seen in the evaluation but not in the calibration period. If the model is calibrated over a 25 yrs period that includes both 1978-1982 and 1983-1987, the process that was relevant only in 1983–1987 is now included in the calibration period. If there are potentially additional processes that are however not seen in the 25-year calibration period, these processes cannot explain the decrease in model performance when e.g. calibrating in the first 5 yrs of this period and evaluating over the other 20 yrs of this period.

References

Duethmann, D., and Blöschl, G.: Why has catchment evaporation increased in the past 40 years? A data-based study in Austria, Hydrol. Earth Syst. Sci., 22, 5143-5158, 10.5194/hess-22-5143-2018, 2018.

BMLFUW (2011): Irrigated areas in Austria – final report (Bewässerte Flächen in Österreich – Endbericht), in German. https://gruenerbericht.at/cm4/jdownload/download/28-studien/470-39-abschaetzung-der-bewaesserungswuerdigen-flaechen (last access 11. March 2020).

Replies to Chang Liao

We would like to thank Chang Liao for his interest in our paper and for uploading his comments.

Below, comments by Chang Liao are in italic font and our replies are in normal font.

This manuscript tries to untangle one of the most challenging problems in hydrology, and it has implications to more than hydrology models: why even a calibrated hydrology model is not reliable for future simulations? While the authors lay down quite great effort to test and examine some hypothesis, its vision and credibility may be shorten by some major limitations. It is great to see authors went through input driving data (precipitation, temperature, etc.) to all the way up to discharge. The whole analytical process was very convincing. Regardless of the model details, I only have a couple of concerns and comments.

First, various spatially distributed hydrology models were used across scales. The authors need to justify why HBV is representative here. There are models considering vegetation dynamics for example.

We are not claiming that the applied model is representative for all hydrological models and acknowledge that there are models that consider vegetation dynamics. However, conceptual HBV-type models are often used in the context of national scale climate change impact assessments. The fact that in this study, HBV did not result in reliable discharge simulations in a transient climate is thus concerning and very relevant for studies that apply HBV-type models (or similar models that neglect changes in vegetation dynamics).

Second, as authors pointed out many sources may contribute to model low performance, I suggest there should be at least more evaluations of various hydrological processes. For example, the spatial maps of snow cover, SWE, canopy interception, runoff, snowmelt, soil moisture, etc. A cost function only focus on discharge will likely miss a lot of information. We all know a combination of different parameters can produce the similar results but only one of them is the correct set. The only way to reduce this uncertainty is to examine every single step.

> We agree that including more data on other variables than discharge in the objective function is a good idea. However, for most of the suggested fluxes or state variables there are no observations to compare to (or, available observations are not directly comparable to the modelled variable, as for example for remotely sensed soil moisture). Since many of the study catchments are in a mountainous region, snow data are a relevant data source and we will add a model variant where snow data are included in the objective function (model variant V6, see Tables 2 and 3). The results are described in section 3.3.2.
>
> "Including a snow related criterion into the objective function (model variant V6) improved the model performance with respect to snow without deteriorating the model performance for discharge (Supplementary Table S1). The performance of the model compared to observed snow cover derived from interpolated snow depth was comparable to Parajka et al. (2007), when considering the same set of catchments. Model performance with respect to long-term trends was not improved, with an average gap between simulated and observed discharge trends of 91 ± 50 mm yr^{-1} per 35 yrs over 1978–2013 (Table 4)."

Replies to comments by Yan Liu et al.

We would like to thank Yan Liu, Veit Blauhut, Amelie Herzog, Tunde Olarinoye and Ruth Stephan for their interest in our paper and for posting their comments on our manuscript.

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

Comments are from the discussion during a workshop by: Yan Liu, Veit Blauhut, Amelie Herzog, Tunde Olarinoye, Ruth Stephan

The study "Why does a conceptual hydrological model fail to predict discharge changes in response to climate change?" by Duethmann et al presents a very interesting topic, which tries to find important factors that influence the prediction capability of conceptual hydrological models, especially under climate change. In this study, the HBV model was used as one representative of conceptual hydrological models. Three aspects regarding precipitation input, model calibration period, and potential evapotransporation (LAI and NDVI were used to consider changes of vegetation dynamics and land cover) were investigated to discuss the causes why HBV model fails to predict discharge under changing climate. This study is in the scope of HESS and well written.

After reading and discussing this manuscript during a workshop, we thought that posting our comments might be helpful for improving the manuscript. We have following major and specific points:

Major points: 1) Title and abstract are a bit misleading because the results are not generalising for all hydrological models but using HBV as one representative. It would be better to explicitly state that the results are based on HBV model in the abstract. Using subtitle may also help clarifying this issue.

We will revise the title, please also 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. We will make the abstract clearer and mention explicitly that the results are based on a HBV-type model (and catchments in Austria). To avoid abbreviations in the title we did not add 'a HBV-type model' there. Our results are based on a specific model and catchments in Austria, the problems we found may, however, also be relevant for other hydrological models and other regions (also see SC3 by Taehee Hwang).

2) The prior distribution of model parameters was assumed to be the beta-distribution. In such way, by giving shaping parameters _ and _ for the beta distribution, it seems that the optimal parameter ranges (high probability density part of the beta distribution) are known for the prior. That will affect the model calibration. To justify why using a beta distribution not a uniform distribution for the parameter prior distribution is needed in the method section.

The a priori distributions for the model parameters were applied to be consistent with the study by Merz2011. It is assumed that we have more information on the likely parameter values than just the parameter range. We checked that including the penalty for deviating from the prior distributions does not have much influence on changes in model performance over time (see Figure 1 below, compared to Figure 3 in the manuscript). When the penalty for deviating from the prior distributions was omitted from the objective function, calibrating the model in subperiod S1 and applying it to 1978–2013 resulted in an average discharge trend of 118 ± 86 mm yr^{-1} per 35 yrs and thus virtually no effect compared to the original model.

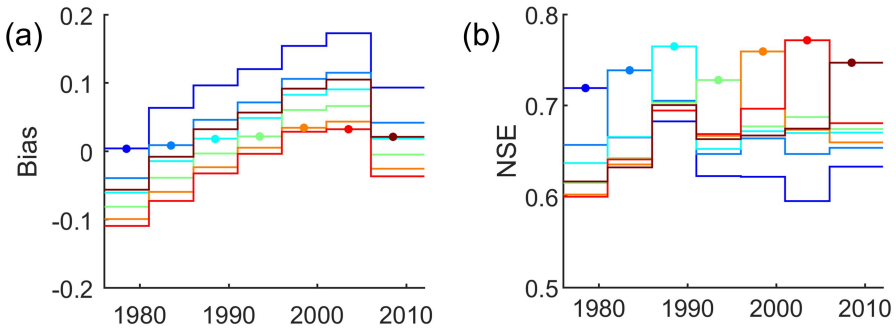

Figure 1 (a) Bias and (b) NSE for the different subperiods averaged over all study catchments when omitting the penalty for deviating from the prior distributions. Each line refers to models calibrated in one subperiod, showing bias and NSE during calibration (marked by the filled circle) and during evaluation in the other six subperiods.

3) Since the results were analyzed for the averages over 156 catchments, it would be better to see the probability density distribution of the bias (Qobs-Qsim) of all catchments for the prediction periods to support that the low predictability of the averages of all catchments is not due to several catchments that bring very big bias. Providing this information in the supplement will strongly support the results.

This is a good idea and we will add violin plots showing the distribution of the bias and NSE as Supplementary Figures S4 and S5. Figure S4 shows that the changes in the average bias were not caused by few catchments with very large changes.

Specific points:

1) A northern arrow is missing in Fig.1, the elevation legend is normally vertical. Fig. 2 is not very informative, maybe merge it with Fig. 1.

A northern arrow will be added. The elevation legend was set horizontal to better use the space. Figure 2 will be moved to the supplement.

2) In Fig. 4, how was the bias calculated.

This will be added. The bias was calculated as $bias = \left(\sum_{t=1}^{n} Q_{sim,t} - \sum_{t=1}^{n} Q_{obs,t}\right) / \sum_{t=1}^{n} Q_{obs,t}$

3) What does the unit "mm yr-1 per 35 yrs" mean? Is that the mean discharge (mm yr-1) over the 35 years?

The unit "mm yr-1 per 35 yrs" refers to trends, such as the trend in mean annual discharge over a period of 35 yrs.

4) In equation 8, definition of fbeta is missing. fp was not used.

Thanks, this will be corrected.

5) In Sect. 2.3.1, many model parameters were introduced, such as CR and Bmax, but these two parameters are not provided in Table 1.

Table 1 lists parameter ranges of the a priori distribution for the model parameters that were included in the model calibration. The parameters T_R, T_S, C_r and B_{max} were not included in the model calibration and set to constant values (see section 2.3.3). They are therefore not listed in Table 1 and we will add a not to the table header to clarify this.

6) Table 2 contains almost all the details of hypotheses. But there is also quite long text in Sect. 2.4.2 that repeats the table. Table 2 is clear, try to reduce the duplicate text in Sect. 2.4.2.

We include small changes in 2.4.2. However, since the text contains further explanations of the hypotheses that are summarized in Table 2, large parts were left as in the original manuscript.

7) Hypothesis should be a result out of the introduction and be mentioned at the last paragraph in the introduction.

That is an alternative we have also thought about. However, the reason why we decided to introduce the detailed hypotheses after the model description and not in the introduction was

that we assume it to be easier for a reader to follow the hypotheses and the corresponding modifications to the model after the model (including the input data) has been introduced.

8) In the discussion, very good literature review was done. But it should more highlight the findings of this study and relate and compare to literatures.

We modified the discussion, also in response to a comment by Referee #2. We discuss the results of this study in the context of existing studies on this topic.

9) It is not clear that how the trend was calculated when using 25 years as the calibration period. Please clarify that.

As in the cases where the model is calibrated for 5-year periods, the parameters are applied to the entire study period and the trend over the entire study period is calculated.

Reply to Taehee Hwang

We thank Taehee Hwang for his entirely positive comment on our paper. Taehee Hwang replied to a comment by Dr. Liu. He underlined the relevance of our study by emphasizing the problem that many hydrologic models neglect changes in vegetation dynamics even though vegetation responses to climate change may have important influences on hydrologic systems.

Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change?

Doris Duethmann[1, 2], Günter Blöschl[1], Juraj Parajka[1]

[1]Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Karlsplatz 13/223, 1040 Vienna, Austria.
[2]IGB Leibniz Institute of Freshwater Ecology and Inland Fisheries Berlin, Müggelseedamm 310, 12587 Berlin, Germany.

Correspondence to: Doris Duethmann (duethmann@igb-berlin.de)

Abstract. Several studies have shown that hydrological models do not perform well when applied to periods with climate conditions that differ from those during model calibration. This has important implications for the application of these models in climate change impact studies. The causes of the low transferability to changed climate conditions have, however, only been investigated in a few studies. Here we revisit a study in Austria that demonstrated the inability of a conceptual semi-distributed HBV-type model to simulate the observed discharge response to increases in precipitation and air temperature. The aim of the paper is to shed light on the reasons of these model problems. We set up hypotheses for the possible causes of the mismatch between the observed and simulated changes in discharge and evaluate these using simulations with modifications of the model. In the baseline model, trends of simulated and observed discharge over 1978–2013 differ, on average over all 156 catchments, by 92 ± 50 mm yr^{-1} per 35 yrs. Accounting for variations in vegetation dynamics, as derived from a satellite-based vegetation index, in the calculation of reference evaporation explains 35 ± 9 mm yr^{-1} per 35 yrs of the differences between the trends in simulated and observed discharge. Inhomogeneities in the precipitation data, caused by a variable number of stations, explain 37 ± 26 mm yr^{-1} per 35 yrs of this difference. Extending the calibration period from 5 to 25 yrs, varying the objective function by including annually aggregated discharge data or snow cover data in the objective function, or estimating evaporation with the Penman-Monteith instead of the Blaney-Criddle approach, has little influence on the simulated discharge trends (less than 5 mm yr^{-1} per 35 yrs). The precipitation data problem highlights the importance of using precipitation data based on a stationary input station network when studying hydrologic changes. The model structure problem with respect to vegetation dynamics is likely relevant for a wide spectrum of regions in a transient climate and has important implications for climate change impact studies.

1 Introduction

A vast number of studies employ hydrological models to estimate climate change impacts on hydrology. In these studies, hydrological models are typically calibrated in the present climate and then run with climate input derived from climate models. However, hydrological predictions under changed climatic conditions are challenging as it is not clear whether the current generation of hydrologic models performs well under change (Blöschl and Montanari, 2010). By definition, testing models under future climate conditions is not possible, as future observations are not available. However, climatic changes have already been observed in the last decades. Hindcast simulations during periods with climatic variations in the past allow testing the suitability of hydrological models under changing climatic conditions. In the differential split sample test (DSST), suggested by Klemeš (1986), a hydrological model is evaluated in a period with climate conditions that differ from those during calibration. Though climatic contrasts between current and future conditions are likely larger than those in the observed record and future conditions will involve higher air temperatures and higher atmospheric CO_2 concentrations, which increases the uncertainties (Stephens et al., 2020), ~~this~~ passing the DSST can be seen as a minimum requirement for models applied ~~for~~ in climate impact assessments.

Studies that investigated the performance of hydrological models this way, evaluating them in periods with climatic conditions that differ from those of the model calibration, largely found a decrease in model performance (Seibert, 2003; Vaze et al., 2010; Merz et al., 2011; Coron et al., 2012; Seiller et al., 2012). In a study on four catchments in Sweden, large flood peaks in the evaluation period were strongly underestimated by the HBV model if the calibration period only contained small flood peaks (Seibert, 2003). Vaze et al. (2010) analysed the model performance of four lumped hydrological models in 61 catchments in southeast Australia when the model was calibrated to selected wet or dry periods of variable length. The reductions in model performance were greater with increasing difference in rainfall between calibration and evaluation periods. While most studies report reduced model performance in contrasting climates, Vormoor et al. (2018) did not find reduced model performance under contrasting conditions in terms of flood seasonality and flood generating processes, when applying a conceptual hydrological model in five catchments with changes in flood seasonality and flood generating processes in Norway.

Low model performance in contrasting climates is often characterized by biased discharge values (Coron et al., 2014; Kling et al., 2015). This is a serious concern since changes in discharge volume are of high interest in climate change impact studies. Merz et al. (2011) calibrated and evaluated the HBV model in 5-year periods in 273 catchments in Austria. They found that median flows were overestimated by 15 % and high flows by 35 % when parameters calibrated during 1976–1981 were applied to 2001–2006. Several studies found increased differences in discharge bias between the calibration and evaluation period with increasing differences in precipitation (Coron et al., 2012; Sleziak et al., 2018).

The problem of poor model performance in contrasting climates has been observed for various model structures. While most studies that investigate the transferability of hydrological models focus on lumped conceptual models, low transferability in contrasting climate has also been observed for semi-distributed conceptual models (Merz et al., 2011; Coron et al., 2014) and

process-based models (Magand et al., 2015). The application of a DSST to three different lumped conceptual models in five catchments in Tunisia showed similar problems of model transferability under contrasting climate conditions for the three models (Dakhlaoui et al., 2017). Seiller et al. (2012) tested the transposability of 20 lumped conceptual hydrological models between periods with contrasting precipitation and air temperature for two catchments in Canada and Germany and they were not able to identify a specific model structure that performed well in contrasting climate for all their test conditions.

Understanding the causes of poor performance in a transient climate is a key question since this determines the way forward for hydrological modelling in a transient climate. Possible causes include data problems, poor parameterization of the model, or structural inadequacy (Coron et al., 2014; Westra et al., 2014; Fowler et al., 2018). In case of data problems, the model should be calibrated with corrected data; however, apart from this, simulations with projections of future climate should not be affected by this problem. In case of parameterization problems, efforts should be invested in choosing calibration methods that result in reliable parameterizations in a transient climate. If the problem is related to the model structure, it will be important to understand what parts of the model structure result in reduced performance in order to avoid these structural components in climate change impact analyses. An example of data problems that may cause poor model performance under contrasting climate conditions are inhomogeneities in the precipitation data, which lead to biased estimates of the precipitation changes. Such inhomogeneities may be caused by inhomogeneities in the station data itself, a variable number of stations included in a gridded data set (Fawcett et al., 2010), or climate variations that lead to changes in the undercatch error (Forland and Hanssen-Bauer, 2000). A poor parameterization may be caused by a too short calibration period. However, in several studies that observed poor performance in contrasting climate the problem could not be solved by using a longer calibration period (Luo et al., 2012; Brigode et al., 2013; Coron et al., 2014). Too low sensitivity of the objective function to the long-term dynamics of discharge may be another cause for a poor parameterization that results in poor performance in a transient climate. Hartmann and Bárdossy (2005) observed increased transferability of a distributed conceptual hydrological model under contrasting climate conditions when including annually aggregated discharge data in the objective function in addition to daily discharge data. A thorough approach to test whether the problem may be solved by improving the parameterization is by applying multiobjective calibration to the different periods with contrasting climate (Fowler et al., 2018). Model structural inadequacy in the context of a transient climate includes changes in catchment characteristics or dominant hydrological processes that are not reflected by the model. For example, changes in the glacier volume or a longer vegetation period may alter the hydrologic response of the catchment and result in deviations between simulated and observed discharge if not accounted for in the model. Despite their relevance for hydrological modelling in a transient climate, the causes of poor performance under contrasting climate conditions have only been investigated in few studies (Westra et al., 2014; Fowler et al., 2016; Fowler et al., 2018).

This study aims at contributing to closing this gap by analysing the causes of the poor performance of a hydrological model in a transient climate for a case study on a large number of catchments in Austria. Due to a strong climate signal over the last decades (Schöner et al., 2011), Austria is well suited for studying climate-induced hydrologic changes. We applied a semi-distributed

hydrological model based on the HBV concept, which is widely used for operational and scientific purposes including climate impact assessments. However, in the study by Merz et al. (2011) (Merz2011 in the following), the model was not able to estimate changes in mean discharge in response to the observed increases in precipitation and air temperature. Applying the model calibrated during 1976–1981 with climate data of 2001–2006 resulted in an increase of simulated discharge of on average 15 %, whereas observations show relatively stable annual discharge volumes. Here, we revisit the study by Merz2011 and investigate what causes the differences between simulated and observed changes in discharge. For that purpose, we set up hypotheses that are tested using modifications of the model. In particular, we analyse the effect of varying the input data for precipitation and air temperature, increasing the calibration period, ~~varying the objective function to~~ includinge annually aggregated discharge data or snow cover data in the objective function, and varying the calculation of reference evaporation (E_{ref}) to consider changes in global radiation and vapour pressure ~~and~~ as well as changes in vegetation dynamics.

2 Data and methods

2.1 Study area

This study was carried out using data from 156 catchments in Austria. The catchments were selected based on the availability of daily discharge data for 1977–2014 (hydrological years, November to October; maximum of two years missing). We excluded catchments with substantial anthropogenic influences from dams or water withdrawals (Viglione et al., 2013), glaciers, and catchments where discharge exceeded the precipitation estimate. The more rigorous selection resulted in smaller set of catchments compared to Merz2011, who used a set of 273 catchments. The median (interquartile range) of the catchment sizes is 198 (96~~5~~/3698) km². The data set includes lowland and mountain catchments and the median elevation range is 5198 (3732/6647)–158293 (98475/2126) m, (numbers in brackets refer to the interquartile range). The most frequent land cover is forest, which covers on average 5~~1~~2(~~38~~40/67) % of the catchment area (~~numbers in brackets refer to the interquartile range,~~ based on Corine 2000 data; European Environment Agency (2016)~~(European Environment Agency, 2016)~~), and grassland, which covers 23(14/334) % of the catchment area. In most catchments the fraction of arable land and heterogeneous agricultural areas is small with a median of 5(0/29) % of the catchment area. The study region shows strong climatic changes over the recent decades. On average over the study catchments annual precipitation increased by 32 ± 23 mm yr^{-1} or 2.4 ± 1.7 % per decade, air temperature increased by 0.45 ± 0.09 °C per decade and global radiation increased by 5.1 ± 0.9 W m^{-2} per decade over the period 1977–2014. In contrast, discharge did not show strong trends and the average trend over the study period was 0.2 ± 3.1 % per decade (Duethmann and Blöschl, 2018).

2.2 Hydrometeorological data

Discharge data were provided by the Central Hydrographical Bureau (HZB) in Vienna. Climate data required by the hydrological model are air temperature, precipitation, and, depending on the model variant, relative humidity, global radiation and wind speeds. Furthermore, interpolated snow depth data were used for model calibration in one model variant. The baseline precipitation data set (P0) was derived by spatially interpolating daily precipitation values of the available stations from HZB and the Austrian Central Institute for Meteorology and Geodynamics (ZAMG) using external drift kriging with elevation as auxiliary variable to a 1 km^2 grid, as in Merz2011. Due to variations in the station network, the number of stations included in the interpolation varies over time. In addition, two alternative precipitation data sets were used. As the first alternative (P1), we used the gridded SPARTACUS data set (Hiebl and Frei, 2016). It has a temporal and spatial resolution of 24 h and 1 km and is based on a two-step interpolation scheme. In the first step, a daily background climatology for 1977–2006 was obtained based on 1249 stations (including 119 totalizer precipitation gauges), and in the second step, a constant number of 523 stations was used for interpolating ratios between the daily precipitation and the background climatology. For the second alternative precipitation data set (P2), we added a correction for systematic underestimation from gauge undercatch to the SPARTACUS data set using the following equation (Richter, 1995)

$$P_{\mathrm{corr}} = P_{\mathrm{orig}} + b \cdot {P_{\mathrm{orig}}}^{e} \quad (1)$$

where P_{corr} is undercatch corrected precipitation, P_{orig} uncorrected precipitation, and b, e are coefficients that depend on season, precipitation type and wind exposure. We estimated the precipitation type as snow for mean air temperatures below −1°C, as mixed precipitation between −1°C and 3°C, and as rain for mean air temperatures above 3°C (ATV-DVWK, 2001). The coefficients of Richter (1995) for very sheltered locations were applied to all grid points. On average over all catchments, the undercatch correction increased precipitation by 7.2 % compared to the original data without undercatch correction.

The baseline data set for mean daily air temperature (T0) was derived by spatially interpolating mean daily air temperatures of the available stations from the ZAMG using local ordinary least-squares regression with elevation, as in Merz2011. In addition, we used the gridded SPARTACUS data set (Hiebl and Frei, 2016), which is based on a constant station network of 150 stations, as alternative input (T1). Air temperature and precipitation were aggregated to averages by elevation zone for each catchment, as used by the hydrological model.

For model variants that applied the Penman-Monteith approach for estimating E_{ref}, relative humidity, global radiation and wind speeds were needed as further input data. Measured global radiation was used rather than global radiation derived from sunshine duration since for this study our interest is in the changes over time and, due to e.g. changes in the atmospheric aerosol concentrations over time (Norris and Wild, 2007), trends in sunshine duration may differ from those in global radiation. Measurements of relative humidity at 7:00 and 14:00 and global radiation were obtained from the ZAMG. Stations with more than

5 % (15 % for global radiation) missing data during 1977–2014 (hydrological years, November to October) were excluded, which resulted in 125 and 6 stations for relative humidity and global radiation, respectively. Data gaps were filled using linear regression to the station with the highest correlation. The data were interpolated onto a 1 km^2 grid using local ordinary least-squares regression with elevation. The local neighbourhood was set to a default radius of 100 km for relative humidity and 200 km for global radiation, adjusted to include at least 10 (global radiation 4) and at most 40 stations. Due to a strong influence of inhomogeneities, long-term changes in wind speed from measured wind speed data are highly uncertain (Böhm, 2008). This is also reflected in the fact that annual anomalies of wind speed data from 85 stations in Austria are hardly related to each other (Duethmann and Blöschl, 2018, see Supplement S1). Uniform monthly wind speeds averaged over all years from all stations in Austria were therefore applied in this study.

For an additional calibration to snow data, snow depth data from HZB were interpolated by external drift kriging with elevation and aggregated to averages by elevation zone for each catchment (Parajka et al., 2007).

2.3 Hydrological model

2.3.1 Model description

In this study, we applied the same hydrological model as Merz2011, which is a semi-distributed conceptual model that follows the structure of HBV (Hydrologiska Byråns Vattenbalansavdelning) (Bergström and Singh, 1995). The model equations can be found in Parajka et al. (2007). The model parameters are listed in Table 1. The model operates on a daily time step and the spatial discretization is based on 200 m elevation bands. Precipitation is partitioned into snow or rain based on air temperature using a threshold temperature T_r. A snow correction factor SCF corrects undercatch of the precipitation gauges during snowfall. Snowmelt is calculated using a temperature-index approach based on the degree-day factor DDF and the melt temperature T_M. Actual evaporation (E_{sim}) is estimated as a function of E_{ref} and soil moisture. It equals E_{ref} if soil moisture is above a calibrated threshold LP. Below this threshold, it linearly decreases to zero at a soil moisture level of zero. The fraction of the sum of rain and snowmelt that results in discharge is calculated as a nonlinear function of soil moisture. This involves the parameters FC, the maximum soil moisture storage, and the nonlinearity parameter B, where a larger B is associated with a smaller fraction of direct runoff and vice versa. The runoff module consists of a hillslope component and a river routing component. The hillslope component is represented by two linear soil stores that are connected through a constant percolation rate C_p. Fast runoff is generated if the state of the upper zone store is above a threshold LSUZ, using a fast storage coefficient K_0. Medium and slow runoff components are calculated as outflow from the upper and lower zone store, using the storage coefficients K_1 and K_2. In the river routing component, runoff routing in streams is simulated using a triangular transfer function involving the parameters C_R and B_{max}.

2.3.2 Estimation of reference evaporation

Despite being technically external to the applied HBV model, the estimation of E_{ref} is considered part of the hydrological model rather than part of the input data since it is calculated and not available as measured data. E_{ref} is computed on a 1 km^2 grid and aggregated to elevation zones for each catchment, as used in the hydrological model. For the baseline model, E_{ref} was derived based on a modified Blaney-Criddle method (DVWK, 1996), following Merz2011, denoted as E0

$$\mathrm{E0} = -1.55 + 0.96 \cdot (8.128 + 0.457 \cdot T) \cdot \frac{S_0 \cdot 100}{S_{\mathrm{year}}} \quad (2)$$

wWhere T is the mean daily air temperature at 2 m height (°C), S_0 the potential daily sunshine duration (h), and S_{year} is the mean yearly sum of potential daily sunshine duration (h).

In order to consider interannual variations in global radiation and vapour pressure deficit, in addition to air temperature, we calculated E_{ref} using the Penman-Monteith equation for well-watered short grass vegetation (Allen et al., 1998), denoted as E1

$$\mathrm{E1} = 0.408 \cdot \frac{\Delta \cdot (R_\mathrm{n} - G) + \gamma \cdot \frac{185400}{(T + 273) \cdot r_\mathrm{a}} \cdot (e_\mathrm{s} - e_\mathrm{a})}{\Delta + \gamma \cdot (1 + \frac{r_\mathrm{s}}{r_\mathrm{a}})} \quad (3)$$

where R_n is the net radiation at the crop surface (MJ m^{-2} d^{-1}), G is the soil heat flux density (MJ m^{-2} d^{-1}), r_a is the aerodynamic resistance (s m^{-1}), r_s is the surface resistance (s m^{-1}), e_s is the saturation vapour pressure (kPa), e_a is the actual vapour pressure (kPa), Δ is the slope of the vapour pressure curve (kPa °C^{-1}), and γ is the psychrometric constant (kPa °C^{-1}). According to the reference conditions of a vegetated surface with a height of 0.12 m, r_s = 70 s m^{-1} and $r_\mathrm{a} = 208/u_2$ where u_2 is the wind speed at 2 m height (m s^{-1}), which was derived from the wind speed at 10 m height based on a logarithmic wind speed profile (Allen et al., 1998). The ground heat flux was neglected. The vapour pressure deficit $e_\mathrm{s} - e_\mathrm{a}$ was calculated as the average of the vapour pressure deficit at the minimum air temperature (using relative humidity at 7:00 LT) and at the maximum air temperature (using relative humidity at 14:00 LT). R_n was estimated from global radiation (R_s; MJ m^{-2} d^{-1}), albedo (α; set to 0.23) and net longwave radiation (R_nl; MJ m^{-2} d^{-1})

$$R_\mathrm{n} = (1 - \alpha) \cdot R_\mathrm{s} + R_\mathrm{nl} \quad (4)$$

where R_nl was estimated according to Allen et al. (1998) based on minimum and maximum air temperature, clear-sky solar radiation, measured R_s, and the mean daily vapour pressure.

In order to consider additionally changes in the vegetation dynamics, we calculated E_{ref} using a variable surface resistance based on changes in a satellite-based vegetation index (E32). We used observed 15-day maximum value composite data of the Normalized Difference Vegetation Index (NDVI) at a resolution of 8 km from the Advanced Very High Resolution Radiometer

(AVHRR) from Tucker et al. (2005). For each point in time of this biweekly series, we aggregated the NDVI data to 200 m elevation zones based on the NDVI data for a rectangle around Austria. As the NDVI data is only available starting in 1981, we applied the data of July 1981–June 1982 for 1976–1981, where the NDVI data is not available. We used the parameterization from Sellers et al. (1996) to estimate a variable r_s from the NDVI data. This involved estimating the fraction of photosynthetically active radiation (FPAR) from transformed NDVI data (Eq. (5); Sellers et al. (1996)), estimating the leaf area index (LAI) from the FPAR data (Eq. (6); Sellers et al. (1996)), and estimating r_s from the LAI data (Eq. (7); Allen et al. (1998)).

$$\mathrm{FPAR} = \frac{(S - S_{\min})}{(S_{\max} - S_{\min})} \cdot (\mathrm{FPAR}_{\max} - \mathrm{FPAR}_{\min}) + \mathrm{FPAR}_{\min} \quad (5)$$

where S is a transformed NDVI value $(1 + \mathrm{NDVI})/(1 - \mathrm{NDVI})$, and $S_{\min}$ and $S_{\max}$ are the 5 % and 98 % quantiles of S for a given land cover class.

$$\mathrm{LAI} = \mathrm{LAI}_{\max} \cdot \frac{\log(1 - \mathrm{FPAR})}{\log(1 - \mathrm{FPAR}_{\max})} \quad (6)$$

where $\mathrm{LAI}_{\max}$ is the maximum LAI of a land cover class. In Eq. (5) and Eq. (6), we applied the following coefficients for grassland: $\mathrm{NDVI}_{\min}$ = 0.039, $\mathrm{NDVI}_{\max} = 0.674$, $\mathrm{FPAR}_{\min} = 0.001$, $\mathrm{FPAR}_{\max}$ = 0.95, and $\mathrm{LAI}_{\max}$ = 5 (Sellers et al., 1996).

$$r_s = r_l \cdot (\mathrm{LAI} \cdot 0.5)^{-1} \quad (7)$$

where r_l is the leaf surface resistance. We applied a value of r_l = 100 s m^{-1} for well-watered grass (Allen et al., 1998). Since the satellite based LAI values derived this way are often lower than the value of 2.88, which is assumed in the Penman-Monteith equation for well-watered short grass by Allen et al. (1998), E23 generally resulted in lower annual E_{ref} than E21. Based on the annual average ratio of E23 to E12 averaged over all catchments, E23 was multiplied by 1.2 to avoid water balance problems in the hydrological model. Such an adjustment of E_{ref} may be justified based on the fact that our study catchments are dominated by forest, and the maximum possible evaporation under well-watered conditions ($E_{\max}$) of forests is typically higher than E_{ref} that assumes short grass. For example, analyses from non-weighable lysimeters suggest $E_{\max}$ to be 20 %–30 % higher for sites with pine forests at typical stand ages of 80–100 years compared to sites with grass (ATV-DVWK, 2001).

2.3.3 Model calibration

The objective function applied for model calibration consisted of three parts. An average of the Nash-Sutcliffe efficiency of linear and logarithmic discharge values (f_Q) was applied in order to achieve a balanced model performance for high and low flows. In order to keep the volume bias low the absolute value of the relative volume bias (f_{bias}) was added as a penalty. Furthermore, a penalty for model parameters that deviate from an a priori distribution (f_{Pbeta}) was added. The penalty function f_{Pbeta} is based on a Beta distribution for each parameter, as described in Merz2011. The a priori distributions for the model parameters were applied

since, on the basis of the literature and previous applications of the model, we believe to have more information on the likely parameter values than just the parameter range. Including this criterion in the objective function has very little influence on the difference between simulated and observed discharge trends (Supplement S1). These objectives were combined in the following way

$$f_1 = w_1 \cdot (1 - f_Q) + w_2 \cdot f_{bias} + w_3 \cdot f_{beta} \quad (8)$$

setting the weights $w_1 = 0.8$, $w_2 = 1$, and $w_3 = 0.2$.

In order to test whether including annually aggregated discharge data in the objective function improves the model performance under transient climate conditions we additionally applied a modified objective function

$$f_2 = w_1 \cdot (1 - f_Q) + w_2 \cdot f_{bias} + w_3 \cdot f_{beta} + w_4 \cdot (1 - f_{annual}) \quad (9)$$

where f_{annual} is the Nash-Sutcliffe efficiency calculated for annually aggregated discharge data. The weights were set to $w_1 = 0.4$, $w_2 = 1$, $w_3 = 0.1$, and $w_4 = 0.5$.

In a further model variant, we tested whether including snow data improves the model performance under transient climate conditions. The snow related part of the objective function aims at minimizing the number of days with poor snow cover simulations and was defined following Parajka et al. (2007). Observed snow cover was derived from maps of interpolated snow depth. An elevation zone was considered as snow covered if the average interpolated snow depth was greater than 0.5 mm, and snow free otherwise. In the model, an elevation zone was considered snow covered if the simulated snow water equivalent was greater than 0.1 mm, and snow free otherwise. If the difference between simulated and observed snow cover on a particular day was greater than 50 % of the catchment area, it was considered as a day with poor snow cover simulations. The snow related part of the objective function f_{snow} was defined as the ratio of the number of days with poor snow cover simulation and the number of days with observed snow cover. The overall objective function was then defined as

$$f_3 = w_1 \cdot (1 - f_Q) + w_2 \cdot f_{bias} + w_3 \cdot f_{beta} + w_4 \cdot f_{snow} \quad (10)$$

The weights were set to $w_1 = 0.7$, $w_2 = 1$, $w_3 = 0.1$, and $w_4 = 0.2$, following Parajka et al. (2007).

The objective function was minimized automatically with the shuffled complex evolution algorithm (SCE-UA) (Duan et al., 1992), a global optimization method based on the simplex downhill search scheme (Nelder and Mead, 1965). The calibration included 11 parameters. The upper and lower bounds and two further parameters of the Beta distribution for each parameter were selected following Merz2011 (Table 1). Four parameters that showed little sensitivity were pre-set to the following values: T_R = 2°C, T_S = 0°C, C_r = 25 d^2 mm^{-1}, and B_{max} = 10. As the focus of this study was on calibrating the model many times for different calibration periods, catchments and model variants, characterizing parameter uncertainties was beyond the scope of this study. For the baseline model, we used seven consecutive 5-year calibration periods without temporal overlap (based on hydrological years),

during 1978−2012. Each simulation was started with an additional 22-month warm-up period. As a modification, we also tested using a 25-year period as calibration period (1978−2002).

2.4 Analysing model problems for simulations under changing climate conditions

2.4.1 Metrics for evaluating model performance under changing climate conditions

Model performance was evaluated using the relative bias in discharge volume and the Nash-Sutcliffe efficiency (NSE). The relative bias in discharge volume was calculated as

$$bias = \left(\sum_{t=1}^{n} Q_{sim,t} - \sum_{t=1}^{n} Q_{obs,t}\right) / \sum_{t=1}^{n} Q_{obs,t} \quad (11)$$

where $Q_{sim,t}$ and $Q_{obs,t}$ are respectively the simulated and observed discharge on day t and n is the number of time steps.

In order to focus on the change in discharge under transient climate conditions, we used the difference between simulated and observed discharge trends as an additional criterion. Good performance in the calibration period but inability to estimate the changes in observed discharge resulting from the climatic changes indicates problems under transient climate conditions. Trends were evaluated over the entire study period (1978–2013). Trend slopes were estimated by the Sen's slope estimator (Sen, 1968) and trend significance was assessed by the nonparametric Mann-Kendall test (Mann, 1945; Kendall, 1975). Lag-one serial correlation was removed by applying the trend-free prewhitening technique (Yue et al., 2002). Uncertainties of the trend slope were estimated using a bootstrapping approach. For this purpose, 1000 samples of size N were drawn, with replacement, from the record of length N years and the Sen's slope was calculated for each of the 1000 samples. Then, the standard deviation was determined. Trends and the standard deviations were first derived for each catchment and then averaged over the catchments to determine average trends and their uncertainties over a number of catchments.

2.4.2 Hypotheses for the causes of the expected mismatch between observed and simulated discharge changes

We compiled possible explanations for the expected divergence between the observed and simulated changes in discharge based on the frameworks suggested by Westra et al. (2014) and Fowler et al. (2018) and the discussion in Coron et al. (2014). The working hypotheses are grouped into (1) data problems, (2) problems related to the model calibration, and (3) problems of the model structure (see Table 2). In a first analysis, the hypotheses were evaluated based on process understanding and literature. During this process, a number of the working hypotheses were rejected or assessed unlikely a cause of the differences between the observed and simulated discharge changes. Other hypotheses were evaluated using simulations with modifications of the model (Table 3).

(1) Data problems

Discharge data can be misleading if they are influenced by abstractions or streamflow diversions. For example, a general increase in water abstractions would reduce a positive streamflow trend. However, our study includes only catchments that were classified as devoid of substantial anthropogenic influences (Viglione et al., 2013) and any existing streamflow diversions were introduced before the beginning of our study period (BMLFUW, 2015). Changes in water abstractions due to irrigation are not believed to be a major cause for the deviations between simulated and observed discharges as only about 3 % of the arable land in Austria is irrigated (FAO, 2016), the fraction of arable land is small in most of the study catchments (median 5 %, see Section 2.1) and the study catchments have only little overlap with those regions where irrigation is most relevant. These are small areas east, southeast and northwest of Vienna, where estimated average irrigation amounts of agricultural areas exceed 10 mm yr^{-1} (BMLFUW, 2011). Erroneous trends in the discharge data could be caused by systematic trending errors of the rating curve. However, it seems unlikely that the discharge data of a large number of catchments are afflicted by systematic trends in the same direction. Problems in the discharge data were thus assumed unlikely to be a relevant cause for the differences between simulated and observed discharge trends.

Inhomogeneities of the precipitation data would result in biased estimates of the precipitation trends. A problem that would affect a large number of catchments is a varying number of precipitation stations included for generating the gridded precipitation data set. The precipitation data set used by Merz2011 was based on all available stations and included ~800 stations in the end of the 1970s and ~1050 stations around the year 2000 (~~Figure 2~~Supplementary Figure S2). The effect of the changes in the number of stations on the trends in the water balance components was analysed by simulations with a precipitation data set based on all available stations (P0) and simulations with a precipitation data set based on a constant number of stations (P1). Changes in ~~climate variables that would result in variations of~~ the gauge undercatch error due to changes in climate would also affect a large number of catchments. An increase of precipitation intensity and a decrease of the snow-to-rain ratio are expected to result in a higher catch ratio, meaning that the precipitation increase is ~~stronger~~ lower than perceived by the observed data. The effect of neglecting the systematic precipitation error was estimated by simulations with a precipitation data set that is based on a constant number of stations that was corrected for the systematic gauge undercatch considering the influence of the precipitation type and daily precipitation intensity on the catch ratio (precipitation data set P2).

Similar to the precipitation data set, the air temperature data set in the baseline model was based on a variable station network, though the number of air temperature stations varies much less than the number of precipitation stations (Supplementary Figure S2~~Figure 2~~). We investigated the effect of the changes in the number of air temperature stations ~~on the trends in the water balance components~~ by simulations with ~~an~~ air temperature data sets ~~that uses~~ based on all available stations (T0) ~~compared to simulations with an air temperature data set that uses~~ or a constant number of stations (T1).

(2) Problems related to the model calibration

Problems in the model calibration relate to the problem that in principle parameter sets exist that allow good performance in the calibration and evaluation period but these parameter sets are not the ones identified during model calibration. Possible causes are, for example, a too short calibration period that results in overfitting, or processes that are relevant in the evaluation period but not activated in the calibration period. We therefore tested whether increasing the model calibration period from 5 yrs to 25 yrs reduces the bias between simulated and observed discharge trends. We furthermore investigated whether including annually aggregated discharge data into the objective function improves the model performance under contrasting climate conditions, as found in a study by Hartmann and Bárdossy (2005). Since snow related processes are important in the mountainous part of the study area, we investigated further whether including data on interpolated snow depth into the objective function has an effect on the model performance under transient climate conditions. A recent study has shown that including snow data into the objective function can improve the temporal stability of snow related parameters (Sleziak et al., 2020).

(3) Problems of the model structure

In case the problem cannot be solved by rectifying problems in the data and model calibration, problems in the model structure are likely. These include inadequate process representations and changes in the catchment that are not represented by the model.

Differences between the observed and simulated trends in streamflow may result from a misconception of changes in E_{ref}. In Merz2011 as well as in the baseline model of our study, E_{ref} is estimated using a modified Blaney-Criddle equation, which implies that interannual changes in E_{ref} resulting from changes in other climate variables than air temperature are not accounted for. To consider effects of changes in global radiation and vapour pressure, we therefore additionally applied a more physically based method for estimating E_{ref} using the Penman-Monteith equation (E1).

Further changes may result from changes in the vegetation dynamics as well as the land cover, such as a lengthening of the growing season, or increases in forest at the expense of cropland and extensive grassland, as observed in many parts of Austria (Krausmann et al., 2003; Gingrich et al., 2015). To test the possible effect of changes in vegetation dynamics on changes in the simulated trends of streamflow and evaporation, we performed additional simulations where we calculated a modified E_{ref} considering changes in surface resistance based on a satellite-based vegetation index (E2). Land cover changes from agricultural land to forest may also contribute to changes in the satellite-based vegetation index. It is therefore assumed that the simulations with E_{ref} considering changes in vegetation dynamics include also, to some extent, the effect of changes in land cover.

3 Results

3.1 Deviations between simulated and observed changes in discharge and evaporation of the baseline model

There is a clear gap between simulated and observed trends in discharge when the model calibrated in the first subperiod is applied to the entire period. On average over all catchments, the difference is 92 ± 50 mm yr^{-1} per 35 yrs over 1978−2013 or 12.3

± 6.8 % in relation to observed flow (Table 4~~Table 4~~). This is illustrated in Figure 2a that shows observed and simulated discharge for the model calibrated to 1978−1982 over the entire simulation period. Observed discharge of the 156 catchments showed only small increases over 1978–2013, with an average trend of 30 ± 94 mm yr^{-1} per 35 yrs and significant ($p \leq 0.05$) increases and decreases in 10~~1~~ % and 6~~7~~ % of the catchments. In contrast, simulated discharge on average increased by 122 ± 82 mm yr^{-1} per 35 yrs, with significant increases and decreases in 36~~7~~ % and 1 % of the catchments. Discharge trends were overestimated by the model in many catchments all over Austria (Figure 2c). Large differences between simulated and observed trends particularly occur in central Austria, southern Carinthia and western Tyrol.

The deviations in simulated and observed changes in discharge correspond to deviations in simulated and observed changes in evaporation. The dark blue line in Figure 2b shows the ~~difference between observed precipitation and discharge~~precipitation minus runoff difference, which may be interpreted as water-balance-based evaporation plus storage changes (E_{wb})~~, assuming negligible storage changes~~. The fact that E_{wb} includes storage changes and E_{sim} does not, is relevant for short time scales but less so for long-term trends, as the fluctuations tend to average out over time. For example, the large interannual variations of E_{wb} compared to E_{sim} may be explained by storage changes. Large interannual variations are also observed for the difference between precipitation and simulated runoff, which is conceptually equivalent to E_{wb} (Supplementary Figure S3). Comparing ~~temporal~~ the long-term variations in E_{wb} and E_{sim}, both E_{wb} and E_{sim} show increases, but E_{sim} increased at a much lower rate than E_{wb}. Furthermore, the trend of E_{wb} is reversed for the last two subperiods, whereas E_{sim} increased over the entire simulation period. While the average trend of E_{wb} over 1978–2013 is 131 ± 59 mm yr^{-1} per 35 yrs, with significant increases in 76~~9~~ % of the catchments, the average trend of E_{sim} is 50 ± 13 mm yr^{-1} per 35 yrs, with significant increases in 94~~7~~ % of the catchments.

In order to investigate whether the overestimation of the simulated discharge trend is related to a decrease in simulated storage that is not represented by observed storage we examined simulated changes in storage. For this, we analysed the sum of all simulated storages, i.e. soil moisture storage, upper and lower zone subsurface storage and snow water equivalent, and calculated trends of annually average values (based on hydrological years). Trends in s~~S~~imulated 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 simulate storage. Small changes in simulated storage are ~~This is~~ in agreement with no consistent large scale groundwater changes in the observations ~~(which is the largest component of the simulated storage)~~ (Blaschke et al., 2011; Neunteufel et al., 2017).

While discharge volume biases during calibration were small, with average values over all catchments of 0.005–0.03 for the different subperiods, discharge biases during evaluation were much higher, with average values of −0.13–0.18 over the study catchments (Figure 3a). Curves of average bias during evaluation over the different subperiods for models calibrated in different subperiods show an interesting pattern. Average bias values during evaluation increase from subperiod S1 to S6 by 0.15–0.18 and decrease again for the last period. The curves run almost parallel and differ by a vertical offset that ensures low bias during the

calibration period. The changes in the average bias were not caused by few catchments with very large changes, as shown by changes in the distribution of bias across all catchments (Supplementary Figure S4). NSE values during model calibration varied in the range of 0.704–0.757 on average over the catchments, showing that the model performed well in each subperiod when calibrated to it. As expected, model performance during evaluation was lower, with average values over the study catchments of 0.568–0.713 (Figure 3b). In many cases, model performance decreases with increasing distance between the calibration and the evaluation period, particularly for model evaluations in subperiods S1 and S2. ~~An increase in model performance loss with increasing distance from the calibration to evaluation periods was only observed for evaluating the models in subperiod S1 and S2.~~

The performance of the baseline model agrees well with the study by Merz2011, who found average NSE during model calibration of 0.74–0.77 and average NSE during model evaluation of 0.64–0.69, when evaluating over all subperiods except the one used for calibration (compared to 0.704–0.757 during calibration and 0.637–0.6670 during evaluation in our study). Discharge biases during calibration were slightly smaller in the present study, due to including a penalty for discharge bias in the objective function. The longer study period used in our study revealed that the trend of an increasing difference between simulated and observed discharge, when applying the model calibrated in subperiod S1 to the entire study period, was not continued during the last subperiod.

3.2 Data problems

3.2.1 Precipitation

Driving the hydrological model with a precipitation data set based on a variable number of precipitation stations may influence the estimated trend of precipitation and thus the trend of simulated discharge. In order to quantify this effect, we performed model simulations with a precipitation data set based on a constant number of stations (P1) in comparison to the baseline precipitation data set P0 that uses a variable number of stations. This reduced the gap between simulated and observed discharge from 92 ± 50 mm yr^{-1} per 35 yrs to 55 ± 47 mm yr^{-1} per 35 yrs (~~Table 4~~Table 4), i.e. a reduction by 37 ± 26 mm yr^{-1} per 35 yrs (Table **5**~~Table 5~~). The reduced gap between simulated and observed discharge is consistent with the difference in the trends in the precipitation data sets. The baseline precipitation data set P0 suggests a precipitation increase of on average 161 ± 89 mm yr^{-1} per 35 yrs, whereas the precipitation data set P1 results in an increase of 122 ± 89 mm yr^{-1} per 35 yrs (Figure 4a). Better model performance with respect to changes in streamflow volume is also reflected by smaller increases in bias during evaluation in the different subperiods (Figure 5a).

Changes in the snow-to-rain ratio and in the precipitation intensity may affect the undercatch error and thus the precipitation trend. Figure 4c–e shows that, over the study period, the snow-to-rain ratio decreased and the daily precipitation intensity increased, whereas the number of precipitation days remained relatively stable. In the precipitation data sets P0 and P1, the precipitation

undercatch error is neglected. In order to estimate the magnitude of the effect of changes in air temperature and precipitation intensity on changes of the undercatch error, we performed simulations with a precipitation data set that was corrected for undercatch accounting for daily precipitation intensity and precipitation type, which was estimated based on air temperature (precipitation data set P2). Precipitation data set P2 exhibits generally higher precipitation and, with an average trend of 120 ± 94 mm yr^{-1} per 35 yrs, a similar absolute and a lower relative precipitation increase over time compared to the precipitation data set P1 (Figure 4a). Simulations with precipitation data set P2 resulted in a gap between simulated and observed discharge trends of 48 ± 46 mm yr^{-1} per 35 yrs (Table 4~~Table 4~~), i.e. a reduction by 44 ± 28 mm yr^{-1} per 35 yrs compared to the baseline model V0 that uses precipitation data set P0 (Table **5**). Comparing model variants V2 to V0, strong reductions of the differences between simulated and observed discharge trends particularly occurred in catchments where the differences between simulated and observed discharge trends were large (Supplementary Figure S6d, Figure 2c). T~~t~~he tendency to further reduce the gap compared to simulations with the precipitation data set P1 of 7 ± 9 mm yr^{-1} per 35 yrs was not significant.

3.2.2 Air temperature

In order to investigate the possible effect of changes in the station network for air temperature data, we performed simulations with gridded air temperature data based on stations with a complete record over the study period (T1), as compared to simulations with a gridded data set based on all available air temperature series (T0). This showed virtually no differences in discharge trends between the two variants (Table 4~~Table 4~~). The small effect of varying the air temperature data set can be explained by the fact that changes in the station network were only small (Supplementary Figure S2~~Figure 2~~) and the two data sets result in very similar changes over time (Figure 4b).

3.3 Problems of the model calibration

3.3.1 Varying the length of the calibration period

In order to evaluate whether the calibration period was too short, we increased the calibration period from 5 yrs (1978–1982) to 25 yrs (1978–2002) (model variant V4). This resulted in an average discharge trend of 117 ± 82 mm yr^{-1} per 35 yrs over 1978–2013 (Table 4~~Table 4~~) and thus virtually no effect compared to the baseline model.

3.3.2 Varying the objective function

~~Similarly, c~~Changing the objective function by including annually aggregated discharge data (model variant V5) led to an average discharge trend of 119 ± 83 mm yr^{-1} per 35 yrs over 1978–2013 (Table 4~~Table 4~~) and thus no improvement in the simulation of the long-term discharge trends either.

Including a snow related criterion into the objective function (model variant V6) improved the model performance with respect to snow without deteriorating the model performance for discharge (Supplementary Table S1). The performance of the model compared to observed snow cover derived from interpolated snow depth was comparable to Parajka et al. (2007), when considering the same set of catchments. Model performance with respect to long-term trends was not improved, with an average gap between simulated and observed discharge trends of 91 ± 50 mm yr^{-1} per 35 yrs over 1978–2013 (Table 4).

3.4 Problems of the model structure

3.4.1 Calculation of E_{ref} using the Penman-Monteith equation

To estimate the effect of using a simplified versus a more physically-based equation for estimating E_{ref}, we compared simulations with E_{ref} estimated by the Blaney-Criddle method (simulation V0) to simulations with E_{ref} estimated by the Penman-Monteith method (model variant V6~~7~~). The results showed only negligible differences between the two model variants in terms of simulated discharge trends (Table 4~~Table 4~~). This is consistent with small differences between the trends in E_{ref} estimated by the two different methods, with average trends of 69 ± 13 mm yr^{-1} per 35 yrs for E0 (Blaney-Criddle) and 71 ± 17 mm yr^{-1} per 35 yrs for E1 (Penman-Monteith) (Figure 6).

3.4.2 Calculation of E_{ref} considering changes in vegetation dynamics

In order to consider changes in the vegetation dynamics, we estimated changes in surface resistance based on changes in a satellite-based vegetation index for the calculation of E_{ref} ~~(model variant V7)~~. Accounting for vegetation dynamics in the calculation of E_{ref} increased trends in E_{sim} to 84 ± 16 mm yr^{-1} per 35 yrs (model variant V8), compared to 50 ± 13 mm yr^{-1} per 35 yrs in the baseline model V0 (Table 4~~Table 4~~). This reduced the gap between simulated and observed discharge trends from 92 ± 50 mm yr^{-1} per 35 yrs to 56 ± 49 mm yr^{-1} per 35 yrs (Table 4~~Table 4~~), i.e. a reduction by 35 ± 9 mm yr^{-1}. Increased trends in E_{sim} are consistent with E_{ref} trends that increased from 69 ± 13 mm yr^{-1} per 35 yrs in the baseline model V0 to 110 ± 17 mm yr^{-1} per 35 yrs in model variant V7~~8~~ (Figure 6). Accounting for vegetation dynamics had a rather consistent effect on the discharge trends throughout the catchments (Supplementary Figure S6b and e).

In order to evaluate the effect of combining the model modifications that had a considerable effect on the gap between trends in observed and simulated discharge, we combined ~~Combining~~ the use of the precipitation data set P2 (model variant V2) and the consideration of vegetation dynamics in the calculation of E_{ref} (~~i.e.~~ model variant V8) as model variant V9.~~,~~ Compared to the baseline model, the differences in trends between simulated and observed discharge were reduced by 87 ± 31 mm yr^{-1} per 35 yrs in this model variant so that the differences largely disappeared (Table 4~~Table 4~~). Bias values in the evaluation period for variant V8~~9~~ show only little variation between subperiod S2 to S6, but some variation remains when transferring models from subperiods S1 or S7 to subperiod S2 to S6, or vice versa (Figure 5h). Bias values in the evaluation period were reduced from −0.13–0.18 in

the baseline model to −0.03–0.10 in model variant V8~~9~~. Comparing model variant V9 and the baseline V0, the differences in trends of simulated and observed discharge were reduced in most catchments, with stronger reductions in catchments that showed higher differences in trends of simulated and observed discharge in the baseline model (Supplementary Figure S6f).

4 Discussion

Our analyses suggest that problems in the precipitation data and neglecting changes in vegetation activity were the most important causes of the poor performance of the HBV model in Austrian catchments in a transient climate. Inhomogeneities in the precipitation data set due to a variable number of stations explained 37 ± 26 mm yr^{-1} per 35 yrs of the difference between simulated and observed discharge trends (or 44 ± 28 mm yr^{-1} per 35 yrs when using a precipitation data set that was additionally undercatch corrected). While the original model neglected changes in the vegetation activity and length of the growing season, considering these changes by calculating E_{ref} accounting for changes in stomata resistance based on changes in a satellite-based vegetation index reduced the gap between simulated and observed discharge trends by 35 ± 9 mm yr^{-1} per 35 yrs. Combining both modifications, using a precipitation data set based on a constant number of stations and considering vegetation dynamics for the calculation of E_{ref}, reduced the gap between simulated and observed discharge trends by 95 %.

The model structure deficiencies with respect to vegetation dynamics~~, identified as one cause for the poor performance of the HBV model in Austrian catchments,~~ are likely relevant for a large number of studies in a transient climate, including simulations in the context of climate change impact assessments. In a changing climate, changes in vegetation dynamics (such as increased growing season length) can have substantial effects on changes in the water balance. ~~While the original model neglected changes in the vegetation activity or length of the growing season, considering these changes by calculating E_{ref} accounting for changes in surface resistance based on changes in a satellite based vegetation index reduced the gap between simulated and observed discharge trends by 35 ± 9 mm y^{-1} per 35 yrs.~~ The effect of considering changes in vegetation dynamics observed in this study ~~This~~ is in agreement with other studies that demonstrate impacts of climate-induced changes in growing season length and vegetation growth on the water balance (Caldwell et al., 2016; Hwang et al., 2018; Kim et al., 2018; Gaertner et al., 2019). For example, long-term hydrologic changes in two forested catchments in the southern Appalachians could only be simulated if full vegetation dynamics were incorporated in the eco-hydrologic model (Hwang et al., 2018). Lengthening of the growing season intensified climatically driven increases in evaporation and reductions in streamflow in a mixed forest catchment in New England (Kim et al., 2018). Decreased catchment streamflow over the last 15 years was linked to increased growing season length in six northern headwater catchments (Wang et al., 2019). Increases in evapo~~transpir~~ation in the central Appalachian Mountain region were attributed to longer growing seasons, with an increase of growing season length of 1 day resulting in a moderate increase of evapo~~transpir~~ation of 0.5 mm yr^{-1} (Gaertner et al., 2019). Here, we considered changes in vegetation dynamics by using a variable surface resistance based on changes in a satellite-based vegetation index. Based on a rather simple approach, this should be seen as a first ~~order~~ estimate to demonstrate the significance of changes in vegetation dynamics on the water balance. While in this

study we assume that the simulations accounting for vegetation dynamics also partly reflect the effects of changes in land cover, an approach that allows disentangling these effects would be preferable in future work. The changes in vegetation dynamics were derived from satellite-based data, which are often not available in the context of climate change impact assessments. Future work should therefore aim at approaches that simulate the changes in vegetation dynamics in response to climatic changes that may be implemented into conceptual hydrologic models. The effect of increased atmospheric CO_2 concentrations on surface resistance was neglected in the present study. At the global scale, it is estimated that this effect may have reduced evaporation in the order of 1.6 to 2.0 mm yr^{-1} $decade^{-1}$ since the 1960s (Gedney et al., 2006; Piao et al., 2007).

In this study, we found problems in the model structure with respect to the calculation of evaporation to contribute to poor model performance in a transient climate. Model structural problems albeit in different model components were also found to cause poor performance in a transient climate in other studies. For a case study in south Australia, model performance was improved by allowing the parameter for the maximum capacity of the soil store to vary in time as a function of a linear trend, which was interpreted as increased catchment storage through an increase in farm dams in the catchment (Westra et al., 2014). For a case study in southwest Australia, introducing a nonlinearity parameter and a threshold value for the rainfall-runoff relationship enabled the simulation of dry and non-dry years with the same parameter set, which was not possible with the original model (Fowler et al., 2018). Changes in glacier volume may cause deviations between simulated and observed discharge trends if not accounted for by the model. Therefore, glacier covered catchments were excluded in our study. Model structural deficits with respect to glacier dynamics may be responsible for further deviations between simulated and observed discharge trends in the study by Merz2011, which did not exclude glacier covered catchments, although the total glacier cover of Austria is small (0.5 %; Fischer et al. (2015)).

The mismatch between simulated and observed discharge trends was partly caused by inhomogeneities in the precipitation data. Thus, the problem of the limited suitability of the hydrological model under transient conditions is less severe than previously assumed. The comparison of the precipitation data sets based on a constant and variable station network (Figure 4a) shows very well that trend analyses of gridded data based on a variable number of stations can be misleading. Particularly large effects of changes in the gauge network on estimated trends may occur if the gauged precipitation values are interpolated directly (as for the baseline precipitation data P0), in contrast to interpolation methods that make use of a two-step procedure by interpolating against a climatology (Fawcett et al., 2010). While the SPARTACUS data are currently seen as the best-suited gridded data set for trend analyses in Austria, they may however contain further inhomogeneities. Network inhomogeneities were avoided by using a constant station network and interpolating against a monthly climatology. However, inhomogeneities may be present in the series of individual stations. Homogenized series were available only for 4 % of the station data used for the SPARTACUS data set, and it is estimated that 25 % of the stations used may still be affected by inhomogeneities (Hiebl and Frei, 2017). However, while we expect changes in the precipitation trends for individual (smaller) catchments, it seems unlikely that inhomogeneities in the station data cause changes in the precipitation trends in the same direction for a large number of catchments.

Considering the precipitation undercatch error~~,~~ including effects of climate variability on the undercatch error~~,~~ had a small and not significant effect, when compared to the simulation using the same precipitation data without undercatch correction. Since high quality wind speed data were not available, wind speeds were not considered in the calculation of the undercatch error. Analyses of the available data in Austria over 1977–2014 show a slight decrease in wind speeds (on average -3.0 ± 2.5 % per decade, see Supplement S2 in Duethmann and Blöschl (2018)). Decreasing wind speeds would ~~strengthen the decreasing trend of the undercatch error~~would result in increasing catch ratios and mean that our estimate of ~~this~~ the effect of changes in the catch ratio due to climatic variability on the difference between simulated and observed discharge trends is at the lower end.

Increasing the length of the calibration period ~~from 5 to 25 yrs~~ did not reduce the gap between trends in simulated and observed discharge (Table 4~~Table 4~~). This is in agreement with several other studies that found little improvement of the observed poor performance in contrasting climate by using a longer calibration period (Luo et al., 2012; Brigode et al., 2013; Coron et al., 2014). Similarly, changes to the objective function to improve the internal consistency of the model did not lead to a better performance in a changing climate. In this study, we included snow data because of the influence of snow on the hydrology in the study region. Seibert (2003) tested whether including groundwater-level observations in the calibration reduced their problem of low model performance ~~Seibert (2003) found low performance of their model~~ for large floods, when there were no large floods in the calibration period, but this did not lead to improvements. ~~Possible problems in the parameterization were analysed by a stronger focus of the objective function on floods or by calibrating to groundwater data in addition to discharge, but these changes did not solve the problem.~~ The results are more variable with respect to changes in the objective function that put a stronger focus on interannual variability. While including annually aggregated discharge data into the objective function did not reduce the gap between trends in simulated and observed discharge in this study,~~However, other studies have shown problems in the model calibration to be one cause of poor transferability in contrasting climates (Fowler et al., 2016; Hartmann and Bárdossy, 2005).~~ Hartmann and Bárdossy (2005) found that ~~changes to the objective function, such as~~ including annually aggregated discharge data in the objective function in addition to daily discharge data~~,~~ did ~~can~~ improve the transferability of a distributed conceptual hydrological model under contrasting climate conditions in their study. A way to find out whether parameter problems might be the cause when a model shows poor performance in contrasting climates is to apply multiobjective calibration to the contrasting periods, as suggested by Fowler et al. (2018). If this is the case, efforts of finding a parameterization method that identifies parameter sets suitable for contrasting climates only from the calibration period may then be undertaken in a second step. Multiobjective calibration to the contrasting periods was applied i~~I~~n a study that used five different model structures and 86 catchments in Australia (Fowler et al., 2016). The results showed that depending on the acceptance threshold for good model performance, parameterization problems caused a decline in model performance in contrasting climate periods ~~, Fowler et al. (2016) investigated whether failures of the DSST were due to problems of model parameterization or model structure (after excluding catchments with data issues such as systematic errors in the discharge and precipitation series), using multiobjective~~

~~calibration to the contrasting periods. Depending on the acceptance threshold for good model performance, parameterizations that result in a good model performance in a transient climate were found~~ in 35 % or 55 % of the cases of DSST failure.

The present study included a large number of catchments, so we assume that our results are robust. However, it is limited to a particular hydrologic model and a particular region. It should therefore be complemented by further studies on the causes of poor (and good) performance of hydrological models in transient climate conditions. The aim is a more complete picture on in what cases what model structure components and what parameterization methods result in poor model performance in a transient climate so that these model structure components and parameterization methods can be avoided for applications where good model performance in a transient climate is relevant, as for example in climate change impact assessments. Ultimately, this will increase the robustness of hydrologic simulations in a changing climate.

5 Conclusion

In this study, we investigated why the HBV model failed to predict changes in discharge in response to observed increases in precipitation and air temperature for 156 catchments in Austria. The baseline model overestimated the observed discharge trends over 1978–2013 and on average over all catchments by 92 ± 50 mm yr^{-1} per 35 yrs, or 12.3 ± 6.8 % per 35 yrs relative to observed discharge. Simulations with variants of the model ~~showed~~ indicate that the poor performance of the HBV model in Austrian catchments in a transient climate could largely be ascribed to two problems, a model structure that neglects changes in the vegetation dynamics, and inhomogeneities in the precipitation input. Considering changes in the vegetation dynamics by calculating E_{ref} accounting for changes in surface resistance based on changes in a satellite-based vegetation index reduced the gap between simulated and observed discharge trends by 35 ± 9 mm yr^{-1} per 35 yrs. Inhomogeneities in the precipitation data set due to a variable number of stations on average explained 37 ± 26 mm yr^{-1} per 35 yrs of the difference between simulated and observed discharge trends. Extending the calibration period from 5 to 25 yrs, including annually aggregated discharge data or snow cover in the objective function, or estimating evaporation with the Penman-Monteith instead of the Blaney-Criddle approach had little influence on the simulated discharge trends. The model structure deficiencies with respect to vegetation dynamics are likely relevant for a large number of studies in a transient climate, including climate change impact studies. The precipitation data problem highlights the importance of using precipitation data based on a constant number of stations for studies on long-term dynamics. Our study emphasizes the importance of considering interrelations between changes in climate, vegetation and hydrology for hydrological modelling in a transient climate.

Data availability. The discharge data and precipitation data from HZB can be accessed through https://ehyd.gv.at/ (last access: 26 November 2019). The meteorological data from ZAMG are currently not freely available; requests should be directed to klima@zamg.ac.at. The Corine land cover map can be downloaded from https://www.eea.europa.eu/data-and-maps/data/clc-2000-vector-6 (last access: 26 November 2019). The SRTM DEM can be obtained from http://srtm.csi. cgiar.org (last access: 26

November 2019). The NDVI data can be downloaded from https://ecocast.arc.nasa.gov/data/pub/gimms/. The hydrological model simulations are available upon request from the first author.

Author contributions. DD conceived and designed the study, performed the analyses, and prepared the manuscript. GB contributed to the study design and interpretation of the results. JP contributed to the numerical analyses. All authors actively took part in the discussion of the results and revising the paper.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. We thank Mojca Sraj, David Post, Chang Liao, Yan Liu, Veit Blauhut, Amelie Herzog, Tunde Olarinoye Ruth Stephan, Taehee Hwang and an anonymous referee for their comments that helped to improve the manuscript. We gratefully acknowledge the financial support from the DFG (German Research Foundation) through a research scholarship to DD (DU 1595/1-1). We would like to thank the Central Hydrographical Bureau and the Austrian Central Institute for Meteorology and Geodynamics for providing the hydrographic and meteorological data.

References

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration - Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56, FAO, Rome, 300, D05109, 1998.

ATV-DVWK: Verdunstung in Bezug zu Landnutzung, Bewuchs und Boden, GFA-Ges. zur Förderung d. Abwassertechnik e.V., 2001.

Bergström, S., and Singh, V.: The HBV model, Computer models of watershed hydrology., 443-476, 1995.

Blaschke, A., Merz, R., Parajka, J., Salinas, J., and Blöschl, G.: Auswirkungen des Klimawandels auf das Wasserdargebot von Grund- und Oberflächenwasser, Österreichische Wasser-und Abfallwirtschaft, 63, 31-41, 2011.

Blöschl, G., and Montanari, A.: Climate change impacts-throwing the dice?, Hydrol. Process., 24, 374-381, 10.1002/hyp.7574, 2010.

BMLFUW: Irrigated areas in Austria – final report (Bewässerte Flächen in Österreich – Endbericht), in German, https://gruenerbericht.at/cm4/jdownload/download/28-studien/470-39-abschaetzung-der-bewaesserungswuerdigen-flaechen, access: 11. March 2020, 2011.

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

Böhm, R.: Heisse Luft: Reizwort Klimawandel: Fakten, Ängste, Geschäfte, Ed. Va Bene, 2008.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, J. Hydrol., 476, 410-425, 10.1016/j.jhydrol.2012.11.012, 2013.

Caldwell, P. V., Miniat, C. F., Elliott, K. J., Swank, W. T., Brantley, S. T., and Laseter, S. H.: Declining water yield from forested mountain watersheds in response to climate change and forest mesophication, Global Change Biology, 22, 2997-3012, 10.1111/gcb.13309, 2016.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resour. Res., 48, W05552, 10.1029/2011wr011721, 2012.

Coron, L., Andreassian, V., Perrin, C., Bourqui, M., and Hendrickx, F.: On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments, Hydrol. Earth Syst. Sci., 18, 727-746, 10.5194/hess-18-727-2014, 2014.

Dakhlaoui, H., Ruelland, D., Tramblay, Y., and Bargaoui, Z.: Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia, J. Hydrol., 550, 201-217, 10.1016/j.jhydrol.2017.04.032, 2017.

Duan, Q. Y., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015-1031, 1992.

Duethmann, D., and Blöschl, G.: Why has catchment evaporation increased in the past 40 years? A data-based study in Austria, Hydrol. Earth Syst. Sci., 22, 5143-5158, 10.5194/hess-22-5143-2018, 2018.

DVWK: Ermittlung der Verdunstung von Land-und Wasserflächen, 1996.

European Environment Agency: Corine Land Cover 2000 seamless vector data (Version 18.5), Kopenhagen, Denmark, 2016.

FAO: AQUASTAT Main Database, http://www.fao.org/nr/water/aquastat/data/query/index.html, access: 27.04.2020, 2016.

Fawcett, R., Trewin, B., and Barnes-Keoghan, I.: Network-derived inhomogeneity in monthly rainfall analyses over western Tasmania, IOP Conference Series: Earth and Environmental Science, 2010, 012006,

Fischer, A., Seiser, B., Waldhuber, M. S., Mitterer, C., and Abermann, J.: Tracing glacier changes in Austria from the Little Ice Age to the present using a lidar-based high-resolution glacier inventory in Austria, Cryosphere, 9, 753-766, 10.5194/tc-9-753-2015, 2015.

Forland, E. J., and Hanssen-Bauer, I.: Increased precipitation in the Norwegian Arctic: True or false?, Clim. Change, 46, 485-509, 10.1023/a:1005613304674, 2000.

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.

Gaertner, B. A., Zegre, N., Warner, T., Fernandez, R., He, Y. Q., and Merriamb, E. R.: Climate, forest growing season, and evapotranspiration changes in the central Appalachian Mountains, USA, Science of the Total Environment, 650, 1371-1381, 10.1016/j.scitotenv.2018.09.129, 2019.

Gedney, N., Cox, P. M., Betts, R. A., Boucher, O., Huntingford, C., and Stott, P. A.: Detection of a direct carbon dioxide effect in continental river runoff records, Nature, 439, 835-838, 10.1038/nature04504, 2006.

Gingrich, S., Niedertscheider, M., Kastner, T., Haberl, H., Cosor, G., Krausmann, F., Kuemmerle, T., Müller, D., Reith-Musel, A., and Jepsen, M. R.: Exploring long-term trends in land use change and aboveground human appropriation of net primary production in nine European countries, Land Use Policy, 47, 426-438, 2015.

Hartmann, G., and Bárdossy, A.: Investigation of the transferability of hydrological models and a method to improve model calibration, Adv. Geosci., 5, 83-87, 10.5194/adgeo-5-83-2005, 2005.

Hiebl, J., and Frei, C.: Daily temperature grids for Austria since 1961 - concept, creation and applicability, Theor. Appl. Climatol., 124, 161-178, 10.1007/s00704-015-1411-4, 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.

Hwang, T., Martin, K. L., Vose, J. M., Wear, D., Miles, B., Kim, Y., and Band, L. E.: Nonstationary hydrologic behavior in forested watersheds is mediated by climate-induced changes in growing season length and subsequent vegetation growth, Water Resour. Res., 54, 17, doi:10.1029/2017WR022279, 2018.

Kendall, M. G.: Rank correlation methods, 4 ed., Charles Griffin, London, 196 pp., 1975.

Kim, J. H., Hwang, T., Yang, Y., Schaaf, C. L., Boose, E., and Munger, J. W.: Warming-Induced Earlier Greenup Leads to Reduced Stream Discharge in a Temperate Mixed Forest Catchment, Journal of Geophysical Research-Biogeosciences, 123, 1960-1975, 10.1029/2018jg004438, 2018.

Klemeš: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31(1), 13-24, 1986.

Kling, H., Stanzel, P., Fuchs, M., and Nachtnebel, H.-P.: Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates, Hydrological Sciences Journal, 60, 1374-1393, 10.1080/02626667.2014.959956, 2015.

Krausmann, F., Haberl, H., Schulz, N. B., Erb, K.-H., Darge, E., and Gaube, V.: Land-use change and socio-economic metabolism in Austria—Part I: driving forces of land-use change: 1950–1995, Land use policy, 20, 1-20, 2003.

Luo, J. M., Wang, E. L., Shen, S. H., Zheng, H. X., and Zhang, Y. Q.: Effects of conditional parameterization on performance of rainfall-runoff model regarding hydrologic non-stationarity, Hydrol. Process., 26, 3953-3961, 10.1002/hyp.8420, 2012.

Magand, C., Ducharne, A., Le Moine, N., and Brigode, P.: Parameter transferability under changing climate: case study with a land surface model in the Durance watershed, France, Hydrol. Sci. J.-J. Sci. Hydrol., 60, 1408-1423, 10.1080/02626667.2014.993643, 2015.

Mann, H.: Non-parametric test against trend, Econometrica, 13, 245-259, 1945.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47, W02531, 10.1029/2010wr009505, 2011.

Nelder, J. A., and Mead, R.: A Simplex Method for Function Minimization, The Computer Journal, 7, 308-313, 10.1093/comjnl/7.4.308, 1965.

Neunteufel, R., Schmidt, B.-J., and Perfler, R.: Ressourcenverfügbarkeit und Bedarfsplanung auf Basis geänderter Rahmenbedingungen, Österreichische Wasser- und Abfallwirtschaft, 69, 214-224, 2017.

Norris, J. R., and Wild, M.: Trends in aerosol radiative effects over Europe inferred from observed cloud cover, solar "dimming" and solar "brightening", J. Geophys. Res.-Atmos., 112, 10.1029/2006jd007794, 2007.

Parajka, J., Merz, R., and Blöschl, G.: Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments, Hydrol. Process., 21, 435-446, 10.1002/hyp.6253, 2007.

Piao, S. L., Friedlingstein, P., Ciais, P., de Noblet-Ducoudre, N., Labat, D., and Zaehle, S.: Changes in climate and land use have a larger direct impact than rising CO_2 on global river runoff trends, Proceedings of the National Academy of Sciences of the United States of America, 104, 15242-15247, 10.1073/pnas.0707213104, 2007.

Richter, D.: Ergebnisse methodischer Untersuchungen zur Korrektur des systematischen Messfehlers des Hellmann-Niederschlagsmessers, Selbstverl. des Dt. Wetterdienstes Offenbach, 1995.

Schöner, W., Böhm, R., and Haslinger, K.: Klimaänderung in Österreich–hydrologisch relevante Klimaelemente, Österreichische Wasser-und Abfallwirtschaft, 63, 11-20, 2011.

Seibert, J.: Reliability of model predictions outside calibration conditions, Nord. Hydrol., 34, 477-492, 2003.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth Syst. Sci., 16, 1171-1189, 10.5194/hess-16-1171-2012, 2012.

Sellers, P. J., Los, S. O., Tucker, C. J., Justice, C. O., Dazlich, D. A., Collatz, G. J., and Randall, D. A.: A revised land surface parameterization (SiB2) for atmospheric GCMs - 2. The generation of global fields of terrestrial biophysical parameters from satellite data, J. Clim., 9, 706-737, 10.1175/1520-0442(1996)009<0706:arlspf>2.0.co;2, 1996.

Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, Journal of the American Statistical Association, 63, 1379-1389, 1968.

Sleziak, P., Szolgay, J., Hlavcova, K., Duethmann, D., Parajka, J., and Danko, M.: Factors controlling alterations in the performance of a runoff model in changing climate conditions, Journal of Hydrology and Hydromechanics, 66, 381-392, 10.2478/johh-2018-0031, 2018.

Sleziak, P., Szolgay, J., Hlavčová, K., Danko, M., and Parajka, J.: The effect of the snow weighting on the temporal stability of hydrologic model efficiency and parameters, J. Hydrol., 583, 124639, 2020.

Stephens, C. M., Marshall, L. A., Johnson, F. M., Lin, L., Band, L. E., and Ajami, H.: Is Past Variability a Suitable Proxy for Future Change? A Virtual Catchment Experiment, Water Resour. Res., 56, e2019WR026275, 10.1029/2019wr026275, 2020.

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., and Teng, J.: Climate non-stationarity - Validity of calibrated rainfall-runoff models for use in climate change studies, J. Hydrol., 394, 447-457, 10.1016/j.jhydrol.2010.09.018, 2010.

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins - Part 3: Runoff signatures in Austria, Hydrol. Earth Syst. Sci., 17, 2263-2279, 10.5194/hess-17-2263-2013, 2013.

Vormoor, K., Heistermann, M., Bronstert, A., and Lawrence, D.: Hydrological model parameter (in)stability - "crash testing" the HBV model under contrasting flood seasonality conditions, Hydrol. Sci. J.-J. Sci. Hydrol., 63, 991-1007, 10.1080/02626667.2018.1466056, 2018.

Wang, H. L., Tetzlaff, D., Buttle, J., Carey, S. K., Laudon, H., McNamara, J. P., Spence, C., and Soulsby, C.: Climate-phenology-hydrology interactions in northern high latitudes: Assessing the value of remote sensing data in catchment ecohydrological studies, Science of the Total Environment, 656, 19-28, 10.1016/j.scitotenv.2018.11.361, 2019.

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.

Yue, S., Pilon, P., Phinney, B., and Cavadias, G.: The influence of autocorrelation on the ability to detect trend in hydrological series, Hydrol. Process., 16, 1807-1829, 2002.

Tables

Table 1 A priori distribution of parameter values where p_l and p_u are the lower and upper bounds, α and β the parameters of the a priori distribution, and p_{max} the parameter value at which the a priori distribution is at its maximum. Note that the parameters T_R, T_S, C_T and B_{max} were set constant and are therefore not listed here.

Parameter	Unit	Description	p_l	p_u	p_{max}	α	β
SCF	-	Snow correction factor	1	1.5	1.03	1.1	2.5
DDF	mm ~~/~~(°C d~~ay~~)$^{-1}$	Degree-day factor	0.5	5	1.25	1.5	3.5
T_m	°C	Melt temperature	-2	2	0	2	2
FC	mm	Maximum soil moisture storage	0	600	150	1.05	1.15
LP/FC	-	Ratio of limit for E_{ref} and FC	0	1	0.94	4	1.2
B	-	Nonlinearity parameter of runoff generation	0	20	3.4	1.1	1.5
K_0	days	Very fast s~~S~~torage coefficient of additional outlet	0	2	0.5	2	4
K_1	days	Fast storage coefficient	2	30	9	2	4
K_2	days	Slow storage coefficient	30	250	105	1.05	1.05
C_p	mm ~~/~~d~~ay~~$^{-1}$	Percolation rate	0	8	2	2	4
LSUZ	mm	Storage capacity threshold	1	100	50	3	3

Table 2 Working hypotheses for potential causes of the divergence between observed and simulated discharge changes.

Working hypothesis	**Analysis or further explanation**
(1) Data problems	**→ Section 3.2**
(1.1) Problems in the discharge data	
Changes in abstractions or diversions	Catchments with anthropogenic influences were generally excluded. Reviewed comments in the hydrological yearbooks: diversions were introduced before the start of the study period. Only a small fraction of the arable land in Austria is irrigated and this does largely not overlap with the study catchments
Rating curve errors	Rating curve errors unlikely to occur in the same direction for a large number of catchments. → Unlikely to be relevant for a large number of catchments.
(1.2) Problems in the precipitation data	
Inhomogeneities in the precipitation data due to instrument changes	Introduction of heated precipitation gauges → Would result in larger precipitation increases and thus increase the gap between changes in E_{wb} and changes in E_{sim}. Since at most locations with a heated gauge, there is a manually operated gauge in addition and values of the latter are used to report daily precipitation sums, this effect is likely not relevant.
Inhomogeneities in the gridded precipitation data due to changes in the number of stations	Simulations with a precipitation data set that uses a constant number of stations (model variant V1)
Biased estimates of the precipitation trend due to changes in the catch ratio caused by changes in the snow-to-rain ratio and changes in precipitation intensities (in addition to inhomogeneities due to a variable number of stations)	Simulations with a precipitation data with a constant number of stations and correction for the systematic precipitation undercatch (considering the precipitation type and precipitation intensity (based on daily precipitation amount)) (model variant V2)
(1.3) Problems in the air temperature data	
Inhomogeneities in the gridded air temperature data due to changes in the number of stations	Simulations with a data set that uses a constant number of stations (model variant V3)
(2) Problems related to the model calibration	**→ Section 3.3**
Too short calibration period	Simulations with a 25-year calibration period (model variant V4)
Objective function insensitive to long-term discharge variations	Simulations with a modified objective function that includes annually aggregated discharge data (model variant V5)
Internal inconsistencies due to calibration only to discharge	Simulations with a modified objective function that includes a comparison against snow data (model variant V6)
(3) Problems of the model structure	**→ Section 3.4**
Effects of changes in radiation and saturation deficit not reflected by the model	Calculation of E_{ref} with the Penman-Monteith approach (model variant V~~6~~7)
Effects of changes in the vegetation dynamics and land cover not reflected by the model	Calculation of E_{ref} using a variable surface resistance based on a satellite-derived vegetation index (model variant V~~7~~8)

Table 3 Overview of model variants.

Abbreviation	Description	Input precipitation	Input air temperature	Length of cCalibration periods	Objective function	Calculation of E_{ref}
V0	Baseline model	P0	T0	5 yrs	f_1	E0
V1	Vary *P* data set	P1	T0	5 yrs	f_1	E0
V2	Include *P* undercatch correction	P2	T0	5 yrs	f_1	E0
V3	Vary air temperature data	P0	T1	5 yrs	f_1	E0
V4	Increase length of calibration period	P0	T0	25 yrs	f_1	E0
V5	~~Vary objective function~~Include annually aggregated *Q* into obj. function	P0	T0	5 yrs	f_2	E0
V6	Include snow into obj. function	P0	T0	5 yrs	f_3	E0
V6~~7~~	E_{ref} based on Penman-Monteith	P0	T0	5 yrs	f_1	E1
V7~~8~~	Modified E_{ref} dependent on NDVI	P0	T0	5 yrs	f_1	E2
V8~~9~~	Combine V2 and V8~~7~~	P2	T0	5 yrs	f_1	~~E1~~E2

Table 4 Linear trends in water balance components (mm yr^{-1} per 35 yrs) over 1978–2013 as averages over all catchments. Simulated values refer to the model calibrated in subperiod S1 1978–1982. Uncertainties relate to standard deviations of the trend slope averaged over all catchments. For trends in Q_{sim} – Q_{obs}, we first derived series of the differences Q_{sim} – Q_{obs} for each catchment and then estimated trends.

	P_{obs}	E_{ref}	Q_{obs}	E_{wb}	Q_{sim}	E_{sim}	Q_{sim} – Q_{obs}
V0 Bbaseline model	161 ± 89	69 ± 13	30 ± 94	131 ± 59	122 ± 82	50 ± 13	92 ± 50
V1 Vvary *P* data set	122 ± 89	69 ± 13	30 ± 94	92 ± 57	85 ± 80	49 ± 14	55 ± 47
V2 iInclude *P* undercatch correction	120 ± 94	69 ± 13	30 ± 94	90 ± 57	78 ± 86	57 ± 13	48 ± 46
V3 VVary air temperature data	161 ± 89	69 ± 13	30 ± 94	131 ± 59	120 ± 82	51 ± 13	90 ± 50
V4 iIncrease length of calibration period	161 ± 89	69 ± 13	30 ± 94	131 ± 59	117 ± 82	56 ± 14	87 ± 50
V5 Include annually aggregated *Q* into obj. functionvary objective function	161 ± 89	69 ± 13	30 ± 94	131 ± 59	119 ± 83	51 ± 14	89 ± 49
V6 Include snow into obj. function	161 ± 89	69 ± 13	30 ± 94	131 ± 59	122 ± 83	50 ± 14	91 ± 50
V67 E_{ref} based on Penman-Monteith	161 ± 89	71 ± 17	30 ± 94	131 ± 59	120 ± 84	51 ± 14	89 ± 49
V78 Modified E_{ref} dependent on NDVIintroduce NDVI dependent E_{ref}	161 ± 89	110 ± 17	30 ± 94	131 ± 59	87 ± 83	84 ± 16	56 ± 49
V89 combine V2 and V78	120 ± 94	110 ± 17	30 ± 94	90 ± 57	35 ± 86	101 ± 17	5 ± 46

Table 5 Working hypotheses for potential causes of the divergence between observed and simulated discharge changes that were further analysed and estimated magnitude of the effect on the gap between trends in Q_{obs} and Q_{sim} (mm yr^{-1} per 35 yrs) over 1978–2013 compared to the baseline model. This was calculated by deriving series of the differences in annual discharge of the respective model variant compared to the baseline model (e.g., $Q_{sim,V1} - Q_{sim,V0}$) for each catchment and then estimating trends. Uncertainties relate to standard deviations of the trend slope averaged over all catchments.

Working hypothesis	Model variant	Result	Magnitude of the effect (mm yr^{-1} per 35 yrs)
(1) Data problems		**→ Section 3.2**	
(1.2) Problems in the precipitation data			
Inhomogeneities in the gridded precipitation data due to changes in the number of stations	V1	Reduces the gap between changes in Q_{obs} and Q_{sim}	↓−37 ± 26
Biased estimates of the precipitation trend due to changes in the catch ratio caused by changes in the snow-to-rain ratio and changes in precipitation intensities (in addition to inhomogeneities due to a variable number of stations)	V2	Reduces the gap between changes in Q_{obs} and Q_{sim}	↓ −44 ± 28
(1.3) Problems in the air temperature data			
Inhomogeneities in the gridded air temperature data due to changes in the number of stations	V3	Little effect on simulated discharge trends	−1 ± 5
(2) Problems related to the model calibration		**→ Section 3.3**	
Too short calibration period	V4	Little effect on simulated discharge trends	↓ −4 ± 9
Objective function insensitive to long-term discharge variations	V5	Little effect on simulated discharge trends	↓ −3 ± 13
Internal inconsistencies due to calibration only to discharge	V6	Little effect on simulated discharge trends	0 ± 4
(3) Problems of the model structure		**→ Section 3.4**	
Effects of changes in radiation and saturation deficit not reflected by the model	V7	Little effect on simulated discharge trends	↓−2 ± 7
Effects of changes in the vegetation dynamics and land cover not reflected by the model	V8	Reduces the gap between changes in Q_{obs} and $Q_{sim.}$	↓ −35 ± 9

Figures

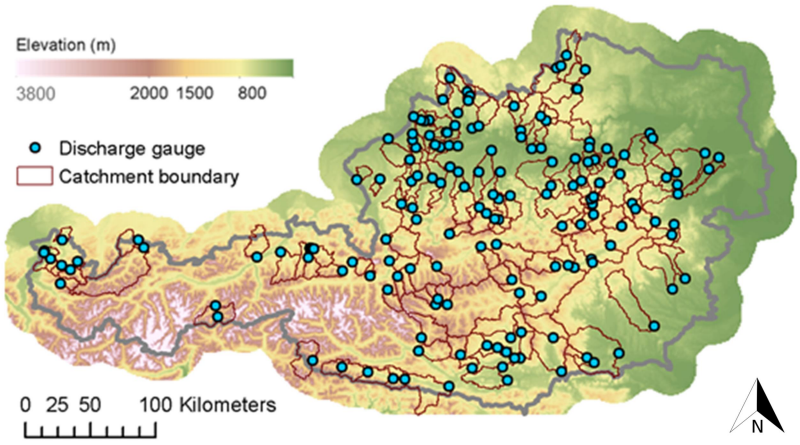

Figure 1 Distribution of the study catchments in Austria.

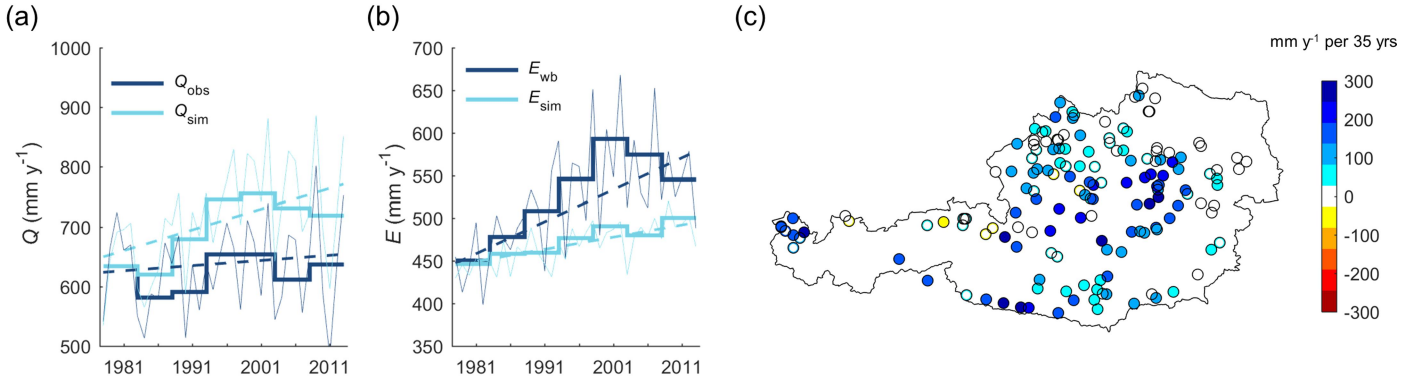

Figure 2 (a) Temporal variations in simulated discharge (Q_{sim}) and observed discharge (Q_{obs}), as averages over all 156 study catchments. (b) Temporal variations in simulated evaporation (E_{sim}) and evaporation derived from the water balance (E_{wb}), as averages over all study catchments. Note that E_{wb} includes storage changes that are particularly relevant for the interannual variations. The thick lines show subperiod annual means, the thin lines annual sums, and the dashed lines linear trends. (c) Spatial pattern of the differences of simulated ~~minus~~ and observed trends in discharge. Filled circles indicate significant trends at $p \leq 0.05$.

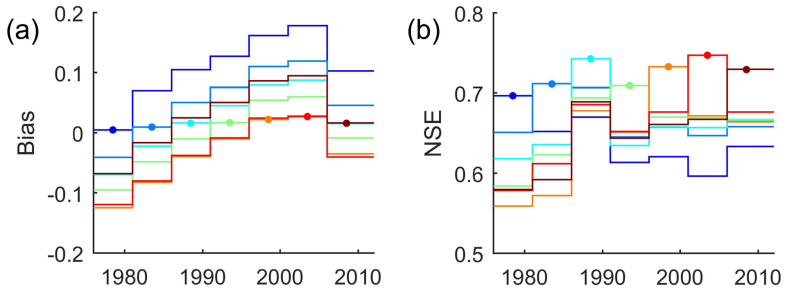

Figure 3 (a) Bias and (b) NSE for the different subperiods averaged over all study catchments for the baseline model V0. Each line refers to models calibrated in one subperiod, showing bias and NSE during calibration (marked by the filled circle) and during evaluation in the other six subperiods.

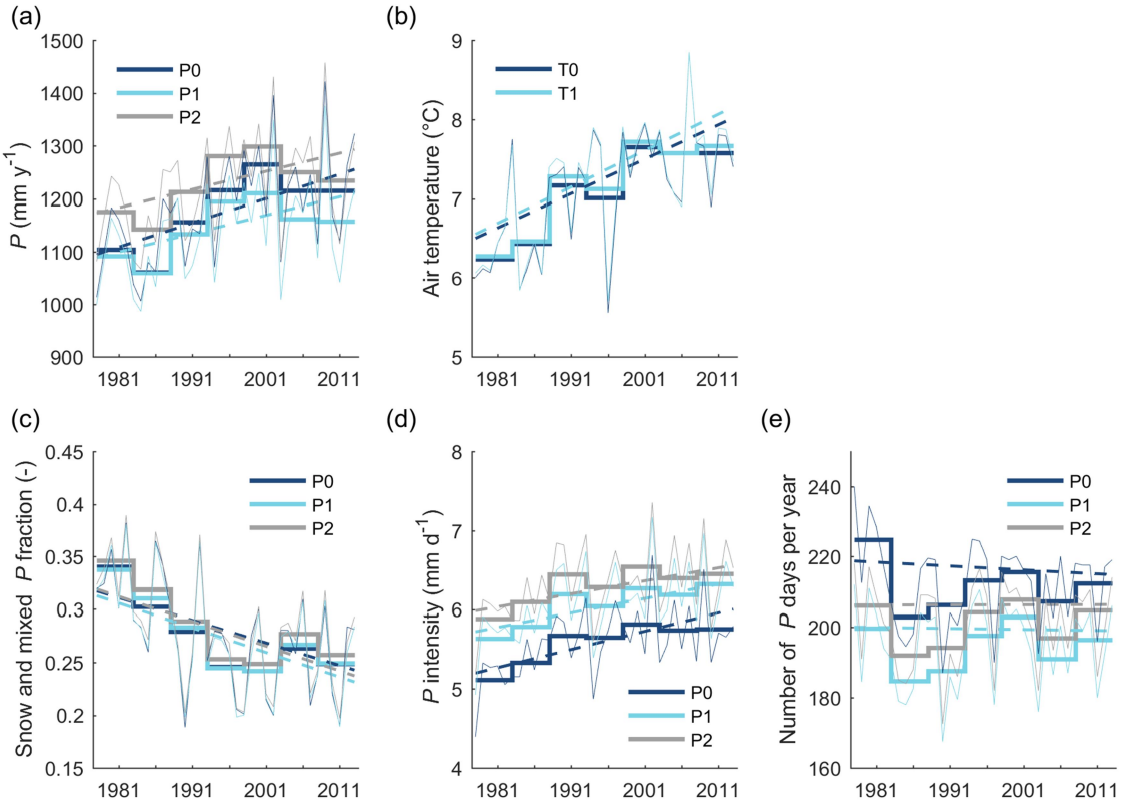

Figure 4 Temporal variations of (a) precipitation, (b) air temperature, (c) fraction of snow and mixed precipitation (estimated as precipitation on days with average daily air temperatures below 3°C), (d) precipitation intensity (precipitation day defined as day with precipitation ≥ 0.1 mm d^{-1}), (e) number of precipitation days per year; as represented by different data sets, averaged over all catchments. The thick lines show subperiod means, the thin lines annual sums, and the dashed lines linear trends, the different colours represent different data sets. Precipitation data set P0 is based on a variable number of stations over time, P1 is based on a constant number of stations, and P2 is based on a constant number of stations and includes a correction for undercatch. Air temperature data set T0 is based ~~i~~on a variable number of stations and T1 is based on a constant number of stations.

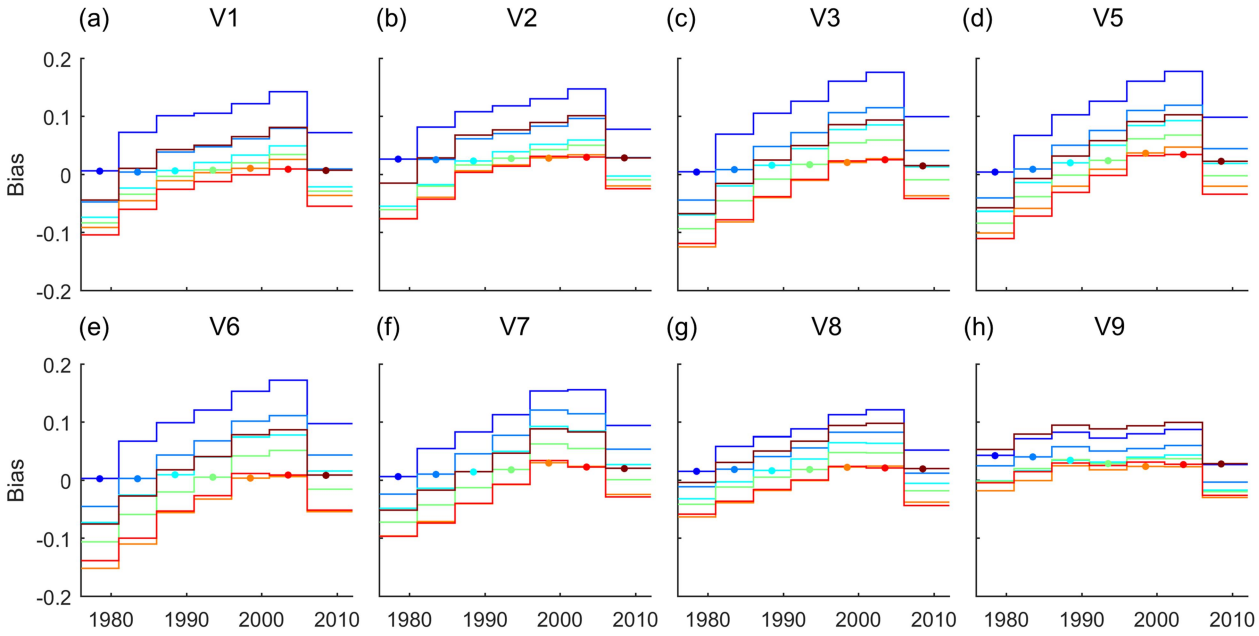

Figure 5 Bias for the different subperiods averaged over all study catchments for model variants V1–V3 and V5–V9 (model variant V4 was not calibrated for different subperiods). Figure 3a shows this for the baseline model V0. Each line refers to models calibrated in one subperiod, where the filled circle marks the calibration period, showing bias during the calibration period and during evaluation in the other six subperiods. For a description of the model variants see Table 3 and section 2.4.2.

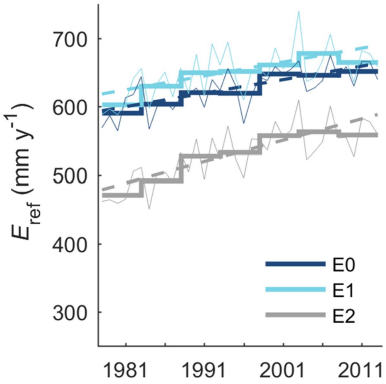

Figure 6 Temporal variations of E_{ref} as calculated by three different methods, averaged over all catchments. The thick lines show subperiod means, the thin lines annual sums, and the dashed lines linear trends, the different colors represent different data sets. Calculation of E_{ref} by: E0 Blaney-Criddle, E1 Penman-Monteith, E2 Penman-Monteith using a variable surface resistance based on changes in a satellite-based vegetation index.

Supplement

Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change?

Doris Duethmann[1, 2], Günter Blöschl[1], Juraj Parajka[1]

[1]Institute for Hydraulic and Water Resources Engineering, Vienna University of Technology, Karlsplatz 13/223, 1040 Vienna, Austria.
[2]IGB Leibniz Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 310, 12587 Berlin, Germany.

Correspondence to: Doris Duethmann (duethmann@igb-berlin.de)

Supplement S1 Influence of including a penalty for model parameters that deviate from an a priori distribution into the objective function

The objective function applied for model calibration contains a penalty for model parameters that deviate from an a priori distribution, consistent with the study by Merz2011. In order to test the possible influence of this criterion on the difference between simulated and observed discharge trends, we also performed simulations where the model was calibrated without this criterion, i.e. w_3 in Eq. 1 was set to 0. This had only small effects on changes in model performance over time (Fig. S1). When the penalty for deviating from the prior distributions was omitted from the objective function, calibrating the model in S1 and applying it to 1978–2013 resulted in a gap between simulated and observed discharge trends of 88 ± 48 mm yr^{-1} per 35 yrs and thus negligible differences to the original model with a gap of 92 ± 50 mm yr^{-1} per 35 yrs.

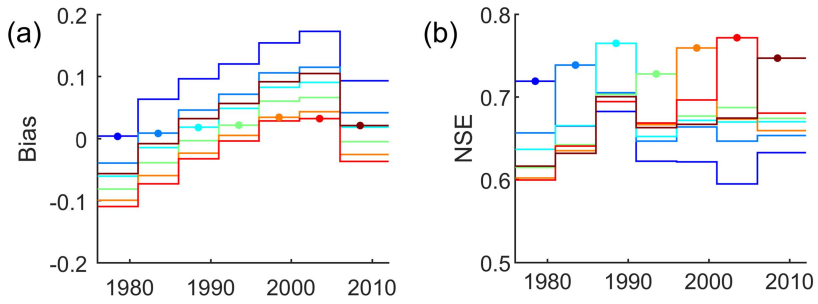

Figure S1 (a) Bias and (b) NSE for the different subperiods averaged over all study catchments when omitting the penalty for deviating from the prior distributions. Each line refers to models calibrated in one subperiod, showing bias and NSE during calibration (marked by the filled circle) and during evaluation in the other six subperiods.

Further supplementary tables and figures

Table S1 Model performance with respect to discharge and snow cover for the baseline model V0 and model variant V6 (where a criterion on snow cover was included in the objective function), as averages over all catchments. Values indicate the range over 7 calibration periods and 42 evaluation periods (using the other 6 subperiods for each calibration period). NSE: Nash-Sutcliffe efficiency for discharge, bias: volume bias for discharge, Z_s: ratio of days with poor snow cover performance (see main text) to the total number of days in the simulation period.

	NSE		Bias		Z_s	
	calibration	evaluation	calibration	evaluation	calibration	evaluation
V0 baseline model	0.70–0.75	0.56–0.71	0.005–0.03	-0.13–0.18	0.08–0.12	0.07–0.13
V6 include snow data in calibration	0.70–0.75	0.58–0.71	0.004–0.04	-0.11–0.18	0.05–0.08	0.05–0.09

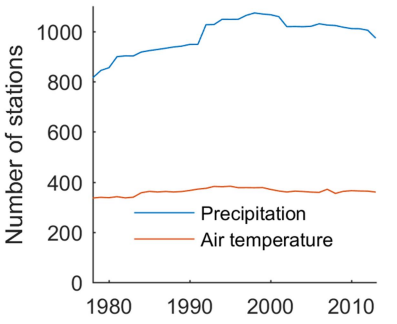

Figure S2 Number of precipitation and air temperature stations included for the interpolation of precipitation and air temperature in the data sets P0 and T0.

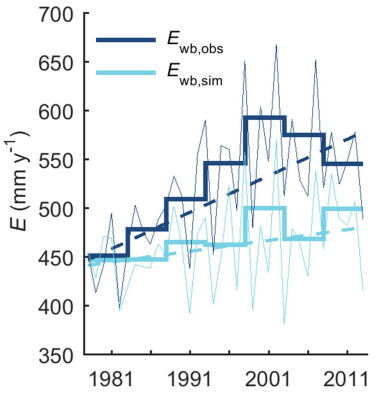

Figure S3 Temporal variations in simulated and observed water balance derived evaporation plus storage changes (calculated as precipitation minus simulated respectively observed discharge), as averages over all study catchments. The thick lines show subperiod annual means, the thin lines annual sums, and the dashed lines linear trends.

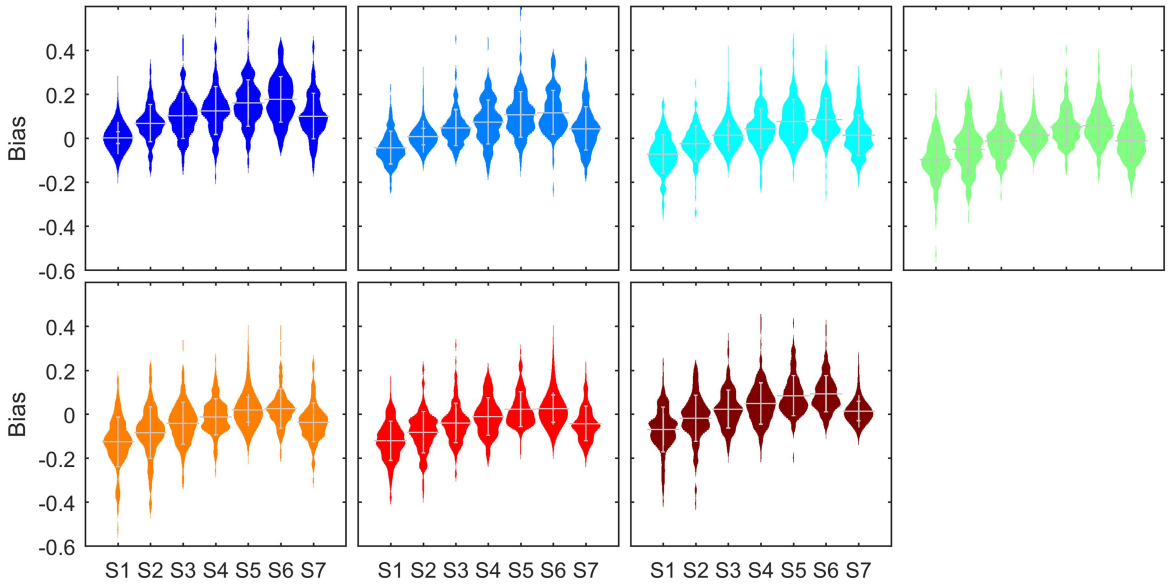

Figure S4 Violin plots showing the distribution of the bias across the 156 study catchments of the models calibrated in subperiod S1 (a), S2 (b), S3 (c), S4 (d), S5 (e), S6 (f), and S7 (g), evaluated for subperiod S1–S7 (*x*-axis in each subplot). Grey crosses represent the mean and standard deviation. The means are the same as shown in Fig. 4 (a).

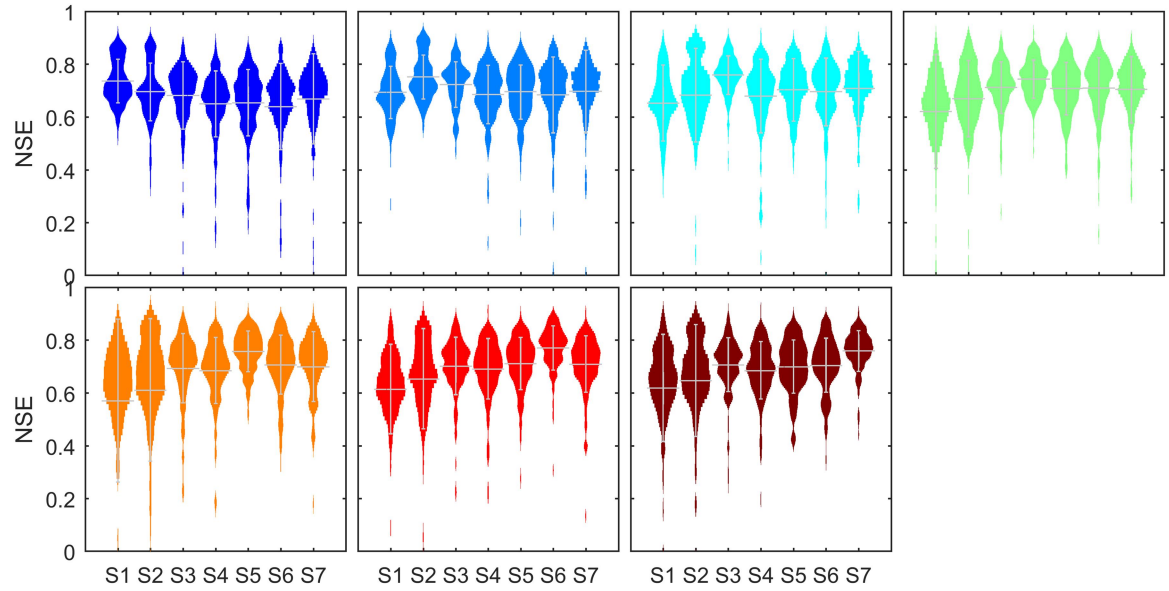

Figure S5 Violin plots showing the distribution of NSE across the 156 study catchments of the models calibrated in subperiod S1 (a), S2 (b), S3 (c), S4 (d), S5 (e), S6 (f), and S7 (g), evaluated for subperiod S1–S7 (*x*-axis in each subplot). Grey crosses represent the mean and standard deviation. The means are the same as shown in Fig. 4 (b).

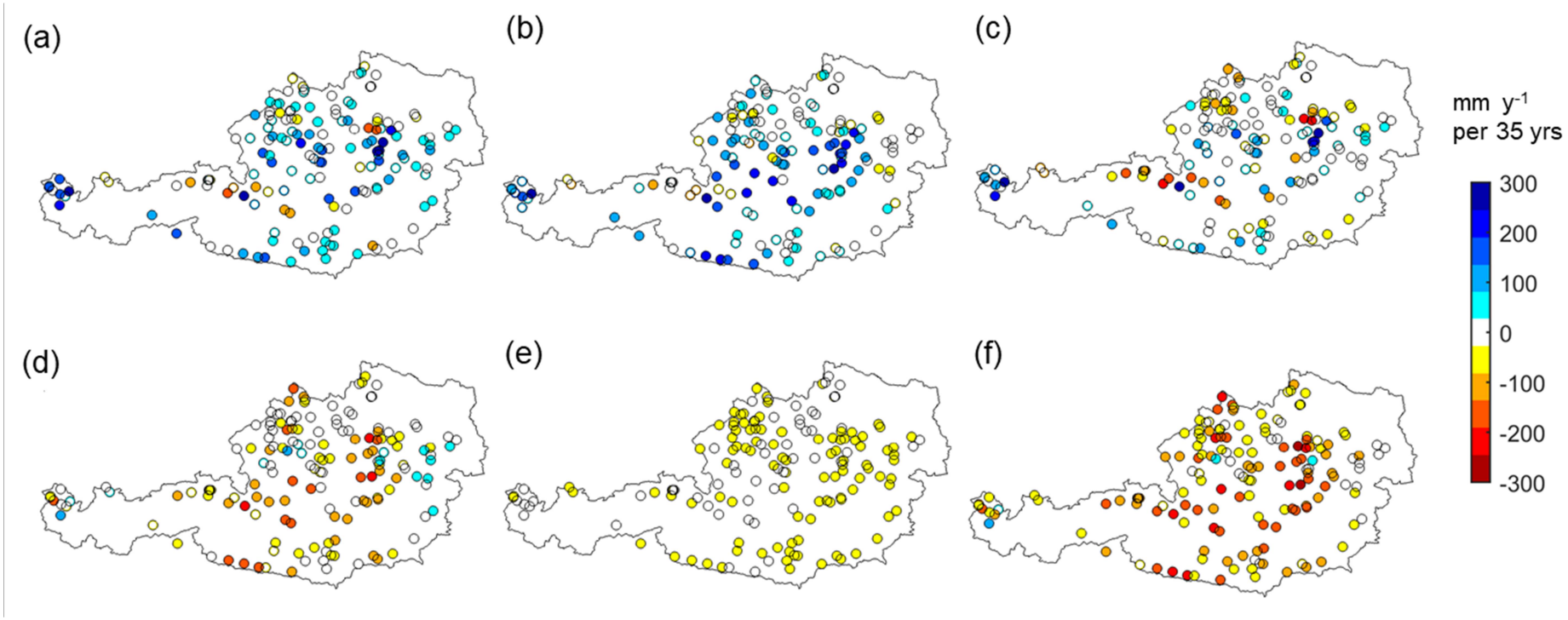

Figure S6 Spatial patterns of trends of the differences between simulated discharge for selected model variants and **(a–c)** observed discharge, or **(d–f)** simulated discharge of the baseline model V0. **(a,d)** refer to model variant V2, **(b,e)** to V7, and **(c,f)** to V8. Filled circles indicate significant trends at $p \leq 0.05$.