Revision #2:

Manuscript title: Improving the quantification of climate change hazards by hydrological models: A simple ensemble approach for considering the uncertain effect of vegetation response to climate change on potential evapotranspiration

Response to the Editor's comments and the peer reviewer's comments

Comments are written in normal letters. Our response is written in *italics.* Changes in the revised text are shown in **bold**.

Comments from the Editor

Unfortunately, one of the reviewers of the first version of the manuscript was not available for a second review, so that only the other reviewer wrote a review of the revised manuscript. This reviewer is not convinced by the revisions made according to the comments on the first version of the manuscript and suggests rejecting it for publication in HESS.

After going again through the comments, I see a main issue of both reviewers in the first round that components of the model were assigned to processes (vegetation adaptation to rising CO2 concentrations), which the reviewers found not appropriate. The reasoning is that the real processes are much more complex than what is captured in the model. They both do not argue that it is wrong to follow the approach in the paper. The point made is that the processes captured by the model are not discussed appropriately and thus are mis-described, or 'oversold'. The point made seems plausible to me. It is not only the comment on the vegetation adaptation, but there are more (smaller) comments that go in a similar direction. The replies to such comments are somewhat evasive or indicate agreement, which is, however, not really accounted for in the changes made. All single comments are maybe not so crucial, but in sum the general impression is that the point of the reviewers was not really taken up and maybe misunderstood. Often the reply is that the model cannot take up something or works in this way or the other and that is why the comments will not be further addressed in the manuscript.

One (small) example: A reviewer comments on the use of the term evapotranspiration. The comment is cited: 'Line 128: AET from soil is a function of soil PET (calculated as the difference between total PET, snow sublimation and canopy evaporation) and soil water saturation. This correct but in detail you are interested in the effect on transpiration of plants.'. The reply is: 'WGHM does not distinguish evaporation from soil and transpiration from vegetation.' I can understand that a reviewer is not satisfied with such a reply. It might be that the model does not distinguish, but in this case this might be a simplification that might matter (to my understanding as being not an expert in the topic, it would). One would not expect

that the model is changed, but that these things are discussed or at least mentioned. A model is just a model and the concepts that matter for a study should be discussed in a paper and not be treated as given and set. Again: This does not mean that it should be changed, but it is important for the interpretation of the results to have such points discussed.

For the vegetation part, both reviewers did not write that the approach is too simplified, but wanted to see that it is sufficiently discussed. Just writing that it mimics (and if one calls it to mimic or to emulate is not the point) plant adaptation is not sufficient. One should get a good understanding of what can be captured by it and what cannot be captured by it, because the related processes are not captured by the modelling approach. This does not have to be an extensive discussion.

The reviewer has three more points, to which the answer was that no changes should be done. The reviewer disagrees.

As both reviewers were in general positive in the first review round, it seems unsatisfying to not publish the paper in HESS. However, the reviewers had reasonable comments. I agree with the reviewer of the revised manuscript that many of the comments were easily discarded, often arguing with how a particular model works or what it can or cannot do. Many points could have taken up at least by adding some discussion into the manuscript. If you think that it is really not possible to address the comments of the reviewers more comprehensively, according to the suggestions of the reviewer, the manuscript cannot be published. I believe, however, that it would not be too much effort to revise the manuscript such that the comments are addressed.

We appreciate the opportunity to revise the paper based on the valuable comments and explanations of the editor and the reviewer(s). We have in the meantime understood, agreeing with the reviewer(s), that it is wrong to call our modified approach for computing Priestley-Taylor PET under future climate change (PT-MA) an approach that mimics or emulates what dynamic vegetation models (or GCMs) simulate. In the previous versions we have, to some extent, "oversold" and misrepresented the scope, meaning and applicability of the proposed PT-MA. We are glad that the editor and the reviewer persisted on changing this.

After reconsidering the applicability of the PT-MA method for climate change impact studies, we have reframed the manuscript as follows. We describe now in more detail the results of Milly and Dunne (2017) who show (somehow hidden in their publication) that the approach of approximating future PET changes as a function of the change in net radiation only (the idea we implemented in our hydrological model) does not work well for some river basins. In these basins, the standard PET computation results in an AET that is closer to the mean AET of the GCM ensemble. Also taking into account the heterogeneous effects of the vegetation responses on groundwater recharge in the study of Reinecke et al. (2021) (as computed by global hydrological and land surface models with and without simulation of vegetation

processes and the impact of CO_2), we came to the conclusion that we want to propose the following approach for climate change impacts studies with hydrological models that do not simulate vegetation processes.

To quantitatively estimate the uncertainty in hydrological responses to climate change that is due to the uncertain vegetation response to climate change, hydrological models should be run in two variants, one with their standard PET approach and one with the modified (PT-MA) approach. Considering the studies of Milly and Dunne (2017) and Reinecke et al. (2021), the second variant is expected to lead to more reliable hydrological changes in most regions, but it is not yet clear in which regions. Therefore, we suggest that these two variants approximately represent the uncertainty bounds for hydrological changes that are related to the vegetation response. The two variants thus serve to generate an improved ensemble of future hydrological changes, which, as a standard, includes model runs driven by the (bias-adjusted) output of multiple GCM or the output of multiple hydrological models.

To reflect the new framing, we have changed large parts of the text in all sections including the abstract, removed the terms "active vegetation", "emulate" and "mimic" and changed the title to:

Improving the quantification of climate change hazards by hydrological models: A simple ensemble approach for considering the uncertain effect of vegetation response to climate change on potential evapotranspiration

In relation to your specific example regarding our lack of discussion/explanation of how transpiration is handled by WaterGAP (old line 128), we added, in section 2.1, the sentence:

When computing soil AET, transpiration of plants is not distinguished from evaporation from the soil, and like typical hydrological models, vegetation responses to changing atmospheric CO_2 and climate that affect transpiration such a stomatal closure and changing leaf area are not simulated.

For the revision, we have fully addressed the points 1, 2 and 3 of the reviewer report #1. Point 1 is addressed by the reframing and comprehensive re-writing described above. To clarify point 2, we added text to the manuscript. Regarding point 3, we applied a statistical significance test to indicate the meaningfulness of the differences in PET and RWR computed with PT and PT-MA. We used the Wilcoxon signed-rank test to identify the grid cells where the difference between the two methods are insignificant, and grayed them out in Figs. 4, 5, 6, B2 and B3. Please see our responses below in the section "Response to Report #1".

We have also improved Figures 4, 5, 6, B2 and B3 by including an example to facilitate easily interpretation of the DC maps.

While we were unable to address the fourth major comment due to the scope limitations, we believe that the improvements made in the re-revised and reframed version warrant further consideration for publication. We assure you that we have

taken the reviewers' comments seriously and have made every effort to enhance the quality and validity of this work.

Thank you for your time and consideration.

Response to Report #1

Overall, I feel the authors did not address the substance of the reviewer comments. Both reviewers raised similar concerns about the analysis and discussion of the approach, to which the authors only made minor edits to the manuscript and no additional analysis. I think the manuscript is focused on an important issue (i.e., that hydrological models need to account for the impacts of vegetation changes), but it is not convincing that the approach presented here effectively addresses the issue. Below are a few examples of why the authors' responses are unsatisfactory.

Thank you for your thorough review of our manuscript and for providing valuable feedback. Our intention has never been to disregard or neglect any of the reviewer's concerns but, as pointed out already in our answer to the editor, we had not really understood the reviewer feedbacks, in particular regarding the concerns regarding "mimicking" and "emulating". We hope that this revision is now an appropriate and correct presentation of our work. We have now reframed the manuscript and substantially changed the text in all sections of the manuscripts, in particular the recommendation to hydrological modelers, and we have carried out a statistical significance test. We believe these revisions address the concerns expressed in a more satisfactory manner.

We greatly appreciate your patience and understanding as we have continued to refine our work. We genuinely value your expertise and perspective in helping us enhance the quality and impact of our research. Thank you once again for your thorough evaluation and constructive comments.

1. Both reviewers had concerns about the phrase ""mimic active vegetation". However, the word "emulating" is not an improvement over mimicking as used in the revised version. It is problematic for two reasons: (1) The approach is not "emulating" active vegetation - from my interpretation, what the approach is doing is taking into account processes that might counter the effects of long term warming. While reductions in stomatal conductance do lead to reduced transpiration that would counter an increase in evaporative demand, other vegetation responses may have the opposite effect, e.g., increasing leaf area leads to more canopy rainwater interception and more canopy evaporation. In some GCMs, the increasing canopy evaporation component can be almost as large as the decreasing transpiration component, but with a lot of regional variation depending on vegetation type and

density. The language used in this manuscript should be much more specific about what physical/vegetation processes the method is attempting to account for, rather than just saying "active vegetation response". (2) The term "emulator" is becoming commonly used to describe some artificial intelligence and machine learning methods, which are more widely used with these types of models in recent years. I would avoid using the term here, so there is no confusion.

In the revised manuscript, the words "emulate", "mimic" and "active vegetation" do not appear anymore. We have (not only) revised the discussion to clarify what process can(not) be taken into account by the PT-MA approach. Also based on a more thorough analysis of the study of Milly and Dunne (2017), we have changed our framing and our proposed way of doing climate change impact studies with hydrological models that do not simulate vegetation processes. Instead of "selling" the PT-MA approach as the simple approach to approximate the effect of vegetation response on PET, we recommend that hydrological modelers take into account the uncertainty that the vegetation response causes for potential evapotranspiration (PET) by always running, in studies on future climate change hazards, two model variants, one with the standard PET approach and one with the modified approach that was implemented for Priestley-Taylor PET with the approach presented in the paper. The reframing can be seen in the new first and last part of the abstract:

Almost no hydrological model takes into account that changes in evapotranspiration are affected by how the vegetation responds to changing CO_2 and climate. This severely limits their ability to quantify the impact of climate change on evapotranspiration and thus water resources. As the simulation of vegetation responses is both complex and very uncertain. We recommend a simple approach for considering, in climate change impact studies with hydrological models, the uncertainty that the vegetation response causes for the estimation of future potential evapotranspiration (PET). To quantify this uncertainty in a simple manner, we propose to run the hydrological model in two variants, with its standard PET approach and with a modified approach for computing PET.

(…)

While the modified approach for computing PET is likely to avoid the overestimation of future drying in many if not most regions, the vegetation response in other regions may be such that application of the standard PET leads to more likely changes in PET. As these regions cannot be identified with certainty, the proposed ensemble approach with two hydrological model variants serves to represent the uncertainty of hydrological changes due to the vegetation response to climate change that is not represented in the model.

The new perspective and recommendation are also reflected in the revised title. Instead of "Improving the quantification of climate change hazards by hydrological models: A simple approach for considering the approximate impact of active vegetation on potential evapotranspiration", it now reads:

Improving the quantification of climate change hazards by hydrological models: A simple ensemble approach for considering the uncertain effect of vegetation response to climate change on potential evapotranspiration

To obtain a consistent storyline, we substantially modified not only the abstract but also the introduction, the results, the discussion and the conclusions.

2. Both reviewers asked why no differences were shown prior to year 2000. The authors' response that the method is only useful after the reference period was not satisfactory. Why is it not useful for easier time periods? And if it is not useful for earlier time periods, why include results back to 1900 that show no difference in the plots?

There are no differences between PET-PT-MA and PET-PT before 2000 due to the choice of the reference period 1981-2000, and the temperature adjustment necessarily only starts after the reference period.

The main reason why we chose the reference period 1981-2000 was that it was the reference period MD used for evaluating the behavior of the global climate, which allow us to perform a direct comparison between the results of our implementation with MD. We modified the text in section 2.2 by adding following sentence:

We chose the period 1981-2000 as the reference period for the implementation of PT-MA in WGHM, and trend removal started in 2001. Selecting this reference period enabled a direct comparison between the results of our implementation with MD, but the PT-MA approach can be implemented with other reference periods, too.

Why is the approach not useful for historical climate change studies before the 21st century? As can be seen in Figure 2a, c, and e (most clearly in c and e related to the two global climate models HadGEM2-ES and IPSL-CM5A-LR), PET computed for the non-water-stressed grid cells used in Milly and Dunne (2016) (MD), with the standard Priestley-Taylor method starts to increase appreciably only after around 2000. Therefore, while an application of the proposed MA method would be possible for reference periods before 1981-2000, our rough approach to capture the processes that might oppose the influences of long-term warming on increasing PET would not lead to meaningful results given the much smaller CO2 and climatic changes in the 20th century as compared to what is expected in the 21st century. With our approach, we want to support the many studies on future climate change impacts on water resources that estimate changes between a reference period in the recent past and a future time period.

In our model validation section (Section 3.1 with Fig. 2) and Section 3.2 (Fig. 3) where we wanted to clarify and illustrate our approach, we showed the time series from 1901-2099 to clarify

1) how the reference period and start year for the adjustment impact adjusted PET computation method, PT-MA,

*2) that the PT-MA adjustment in our model is successful in reducing the PET increment such that it is very similar to the MD-detected PET change (i.e. that PET changes roughly with 0.8 * net radiation) and,*

3) that standard PET does not change appreciably before 2000.

We now start the description of Fig. 2 in section 3.1 with:

The temporal development of the two PET variants and Rn between 1901 and 2099 for the non-water-stressed cells and months, as computed by WGHM, does not show, for all three GCMs, appreciable trends in the 20th century for both PET and Rn (Figure 2 a, c and e). In the 21st century, the variables increase strongly. Reflecting the PT-MA method (section 2.2), PET-PT and PET-PT-MA only start to deviate from each other after the end of the selected reference period (here 1981-2000) of the climate change study. PET-PT-MA increases less strongly after 2000 than PET-PT and very similarly to Rn.

3. Both reviewers asked for some statistical significance testing or indication where the differences are meaningful. While the authors did add a grey mask indicating where the differences were below a given threshold, that is not really a significance test. The values used as thresholds might make sense in some regions, but not in others, depending on the magnitude of internal variability. I disagree with the statement in the authors' response that says: "Testing of the actual significance of the changes of e.g., renewable water resources change (Fig. 5), is conceptually difficult to do at the global scale." This type of test is done all the time in global models. You can simply test at each grid point if the differences are significant relative to the variability at that grid point.

We acknowledge the request from both reviewers for statistical significance testing or indications of meaningful differences in our analysis. While we had included a grey mask to highlight areas where the differences were below a given threshold, we understand that it is better to base the masking on a statistical significance test. We apologize for any confusion caused by our previous response, and we appreciate your feedback on this matter.

For this revision, we performed Wilcoxon signed-rank tests for paired samples to identify whether the differences of PET (and RWR) as computed by the two methods PT and PT-MA are statistically significant (Figures 4, 5, 6, B2 and B3).

The Wilcoxon signed-rank test is a non-parametric statistical test to compare the central tendencies of two paired samples. The null hypothesis (H0) is "The median difference between the values computed by two methods is zero (i.e. no significant difference)". The null hypothesis is rejected with 95% level of significance.

Accordingly, for the analysis period 2080-2099, we applied the Wilcoxon signed-ranked test for each grid cell for the annual time series of PET and RWR obtained through two different methods. In the figures 4 g-h, 5 g-h and 6 c-d (DC maps), we masked out the cells if the test could not reject the null hypothesis under 95% level of significance (i.e. when the difference is insignificant) .

Section 3.3 on PET was revised by adding the following text:

The Wilcoxon signed-rank test for pair samples was applied to understand the significance of the differences between the time series of annual PET 2080-2099 as computed either by the standard PT or the modified PT-MA approach. The null hypothesis is "The median difference between the values computed by two methods is zero (i.e., no significant difference)"; it is rejected at a 95% level of significance. For all GCM, it was found that there is a significant difference between the PET values computed by the two alternative approaches in all grid cells.

Section 3.4 on RWR was revised by adding the following text:

To assess the significance of the differences between the RWR values computed with the PT and the PT-MA variants at the grid cell level, we employed the Wilcoxon signed-rank test for the time series of annual RWR 2080-2099, similar to the approach used for the PET analysis (section 3.3). For 0.5° grid cells shown in light gray in 5 g and h and B3 e-h, differences resulting from the two alternative variants are insignificant.

4. Both reviewers were confused about the balance of net radiation terms and ET (Table 1), to which the authors' explained that the net radiation is calculated differently in the GCM and WGHM. This is an unsatisfactory answer. Please calculate each of the net radiation terms independently in the GCM and WGHM to quantify which ones are contributing to the difference and why. If you are trying to represent ("mimic") ET in the GCMs, it is important to understand how net radiation differs between GCM and WGHM, since this is an important condition for ET and PET. Furthermore, the authors' response that "GCMs do not provide net radiation as an output" is misleading. GCM output provided as part of CMIP experiments includes all the components needed for the authors to calculate net radiation.

While we understand your suggestion to calculate each net radiation term independently for all three GCMs and WGHM (derived with outputs of three GCMs), we have carefully considered this and concluded that conducting such a detailed analysis is beyond the scope of this paper. Several reasons support our decision:

1) Table 1 itself (the last two rows) shows that also the PET-EO values computed with the net radiation of the GCMs by Milly and Dunne (2016) themselves (not us) are not perfectly the same as the NWSAET values of Milly and Dunne (2016). Given this discrepancy due to the simple PET-EO approach, PET-EO and PET-PT-MA computed with the net radiation from our global hydrological model WaterGAP fit together quite nicely, at least much better than when using the standard PT approach. This is the most important conclusion we can draw from Table 1.

2) As mentioned in the paper, the net radiation in the GHM is computed based on bias-adjusted outputs of radiation components from GCMs. Hence the impact of the bias-adjustment method will be a major reason for the differences in the Rn computed by the GCs and WaterGAP.

3) Performing the suggested comparison of GCM and WGHM net radiation would involve extracting a significant amount of data and conducting complex computations. In light of 1 and 2, such an analysis would have a high cost and little benefit, as 1) analyzing the impact of bias-correction is out of the scope of this paper and 2) given the many other approximations it would not help to guide the application of the proposed new approach (of using the PT and PT-MA approach to quantify the uncertainty of computed hydrological changes due to the vegetation response.

To avoid confusion about the variables in the Table 1, we have revised the caption of Table 1 and included the following sentence:

PET-PT, PET-PT-MA and Rn are computed by WGHM using the bias-adjusted output of the listed GCMs.

We also introduce the variables shown in Table 1 at the beginning of section 3.1 more thoroughly:

Performance of the PT-MA method is analyzed based on the area-weighted average changes of PET and Rn over non-water-stressed grid cells and months (Figure 1), considering the changes between the reference period 1981-2000 and the future period 2080-2099 for RCP8.5 as only this RCP was considered in MD (Table 1). Table 1 also presents the respective values for the reference period. PET-PT is the PET as calculated by the standard WGHM, while PET-PT-MA is the result of the PT-MA method presented in section 2.2. For both variants, Rn is computed based on the bias-adjusted output of the three GCMs (section 2.1). PET-EO and NWSAET values for three GCMs were extracted from the MD study and are therefore not affected by any bias adjustment, as they had been derived by MD using the original GCM output.