Second review of Heselschwerdt et al., egusphere-2025-5896 "Large impact of extreme precipitation on projected blue-green water shares"

We appreciate the efforts made by the authors to address our and the other reviewers' comments which have improved the paper. After a few more refinements and expansions the paper is ready for publication in our opinion:

We thank the reviewers for their careful re-evaluation of the manuscript and for their constructive feedback. We are pleased that the revisions have improved the paper and we have addressed the remaining comments in detail below (blue).

Regarding changes made in response to main comment (2) from our previous review: The motivation across lines 13-22 motivates the use of evapotranspiration referring to e. g. land-atmosphere interactions/fluxes or the relevance of partitioning for water availability and ecosystems; while we agree that evapotranspiration does not isolate vegetation-mediated water use, it is never motivated why only vegetation mediated water use is the focus of the analysis. Furthermore, the applied threshold to exclude dry regions (where vegetation functioning and transpiration would be negligible) is very low with 18.25 mm/year, while e.g. the IPCC labels regions as arid when the annual mean precipitation is lower than 300mm/year.

We thank the reviewers for both points. (i) We have revised the Introduction (p. 2, l. 33–41) to motivate the focus on vegetation-mediated water use, explaining that transpiration is the flux through which plants couple water loss to carbon uptake and water-use efficiency, with implications for vegetation productivity, plant water-use responses, evaporative cooling, and land–atmosphere exchange. (ii) We have revised the historical screening mask in Sect. 2.3 (pp. 4–5, l. 108–123) and now retain only grid cells with mean precipitation above $0.822\,\mathrm{mm\,day^{-1}}$ ($\approx 300\,\mathrm{mm\,yr^{-1}}$) and with both mean runoff and mean transpiration above $0.05\,\mathrm{mm\,day^{-1}}$. The precipitation threshold follows the order of magnitude used in the IPCC description of arid zones; the runoff threshold follows Schleussner et al. (2016). The same low-flux value is applied to transpiration because both fluxes enter the BGWS metric symmetrically. A LAI-based criterion is now redundant under the revised precipitation, runoff, and transpiration criteria, as adding the LAI cutoff of $0.3\,\mathrm{m^2\,m^{-2}}$ used in the previous manuscript version excludes no additional grid cells.

Regarding changes made in response to main comment (4) from our previous review: We appreciate the inclusion of the suggested additional predictors. However, the reporting of the results of the tests with respect to including RX1day and crop/tree fraction focuses mostly on model performance changes, while changes in the predictor relevance ranking and the ranking position of these predictors would actually be more informative, particularly because model performance does not change much for different sets of predictors.

We thank the reviewers for this helpful suggestion. We now report the predictor rankings and the ranking positions of the tested predictors, not only model performance, for all sensitivity tests in a dedicated Supplement table (Table S3) and discuss them in Sect. S2 (pp. S2–S3, l. 829–857). The key results are: (i) replacing ΔRX5day with ΔRX1day leaves extreme precipitation as the leading predictor in both regimes, with ΔRX1day ranking first, indicating that the importance of extreme precipitation is robust to the choice of metric; and (ii) adding ΔTreeFrac and ΔCropFrac in the reduced 9-model ensemble changes some secondary ranking positions, but does not overturn the broader process interpretation. We also clarify that some ranking shifts likely reflect the correlated structure among vegetation-related predictors. ΔLAI and ΔTreeFrac are strongly correlated in the reduced ensemble (Pearson's $r = 0.70$), indicating that part of the vegetation-related signal may be shared between them and redistributed in the Elastic Net regression. We therefore interpret the land-cover sensitivity test as supporting the relevance of vegetation-structure changes, while retaining the 12-model fixed predictor set in the main analysis because tree and crop fraction changes are unavailable for the full ensemble and do not alter the overall process interpretation.

Regarding changes made in response to main comment (5) from our previous review: We appreciate the introduction of the masking in the maps in the main figures. However, it remains unclear whether the masking is applied for individual models and how you deal with different resulting masks across individual models. Also, the masking needs to be mentioned in the figure captions, together with information on the fact that the masking considers both low model agreement with each other and with reference datasets. On another aspect, we appreciate the new Figure 1 in the rebuttal. At the same time we disagree with your interpretation and handling of these results. This new figure does show a substantial change in the ranking of the most relevant predictors for the green water regime, and substantially better attribution model performance. For these reasons we feel that this should replace (or at least be added to) main Figure 3 to illustrate the relevance of VPD changes for changes in green water fluxes. Further, this result makes sense from a theoretical perspective because vegetation is projected to be exposed to unprecedented levels of VPD in many regions of the world in the future which is expected to affect its hydraulic functioning. Also, the term partial agreement used in Figures S1 and S2 is not defined.

We thank the reviewers and have clarified the masking approach in Sect. 2.3 (p. 5, l. 121–123). We use a single common mask derived from the ensemble-mean historical climatology, applied identically across all historical fields, future changes, all 12 ESMs, and the reference datasets. Per-model masking would yield a different analysis domain for each model and limit cross-model comparability, so we do not compute per-model masks. We have also revised all relevant figure captions to state explicitly which mask and which uncertainty indicator (inter-model sign agreement or reference-dataset agreement) is used in each figure.

We thank the reviewers for highlighting the relevance of the high-confidence subset shown in the previous rebuttal. We have now made this analysis explicit in the revised manuscript by adding both a map of the high-confidence grid cells by historical regime and the corresponding predictor rankings to the Supplement (Fig. S1 and Table S3), and we discuss the result directly in Sect. 3.3 (p. 16, l. 397–404). We agree that atmospheric dryness becomes more important in the high-confidence green water regime subset: $\Delta\mathrm{VPD}_{\mathrm{seas}}$ and ΔVPD become the two leading predictors, with ΔLAI third, and predictive skill increases to $R^2 = 0.85$ compared with $R^2 = 0.74$ in the full-domain green-regime attribution. We now interpret this as evidence that seasonal atmospheric demand is an important additional control in regions where models robustly agree on the sign of ΔBGWS. At the same time, the high-confidence subset retains only about 25 % of the analysed grid cells and is spatially clustered (Fig. S1). The full-domain attribution therefore remains the main Fig. 3 because it provides the regime-wide attribution across all retained grid cells, whereas the high-confidence subset is interpreted as a regime-conditional finding for these specific climate zones rather than as a replacement for the regime-wide attribution. This distinction is now made explicit in both Sect. 3.3 and Sect. S2.

We have also added a definition of "partial agreement" to the captions of Figs. S1 and S2 (now Figs. S2 and S3). Here we mean grid cells where the 12-model ensemble mean agrees with only one reference dataset.

Other specific comments:

Abstract: Results of Figure 2 which are key findings of this study are not yet mentioned in the abstract.

We thank the reviewers for pointing this out. We have revised the Abstract to include the key result from Fig. 2 (p. 1, l. 7–8). The Abstract now states: "Here, we show that projected BGWS changes are spatially heterogeneous rather than dominated by a uniform global shift."

Section 3.3: It would be useful (and honest) to highlight the uncertainty of these results across individual models in the text, as shown by the dots representing individual model results. This is even more important given the differences of the importance ranking between Figure 3 in the manuscript and Figure 1 in the rebuttal.

We thank the reviewers for this helpful comment. We have revised Sect. 3.3 (p. 14, l. 353–354; p. 16, l. 405–409) to highlight the model-to-model variability in predictor importance more explicitly. We now clarify that the ensemble-mean ranking in Fig. 3 should be interpreted as the common large-scale signal across the 12-model ensemble field, while the individual ESM points illustrate relevant model-to-model variability in predictor importance.

line 205: Recent?

We thank the reviewer for noting this. The word in l. 205, (now l. 216–217), is "represent" and the apparent break as "rep-resent" is caused by automatic hyphenation.

We do not wish to remain anonymous - joint review by Rene Orth and Josephin Kroll.

References

Schleussner, C.-F., Lissner, T. K., Fischer, E. M., Wohland, J., Perrette, M., Golly, A., Rogelj, J., Childers, K., Schewe, J., Frieler, K., Mengel, M., Hare, W., and Schaeffer, M.: Differential climate impacts for policy-relevant limits to global warming: the case of 1.5 °C and 2 °C, Earth System Dynamics, 7, 327–351, https://doi.org/10.5194/esd-7-327-2016, 2016.