We would like to sincerely thank the anonymous reviewer for the detailed and positive evaluation of our manuscript. The suggestions made by the reviewer will clearly help to improve the manuscript.
In the following, we provide point-by-point answers to the reviewer's comments. For the sake of clarity, we first repeat the reviewer's comments (in italic) and then provide our response.

Comment 1:
The authors chose to develop the runoff model from precipitation and temperature, while GS15 found for Europe that this simple model is outperformed by one considering full atmospheric forcing - which is available through GSWP3. This choice must be justified in the methods section. Further, GS15 showed that including land surface parameters (soil texture and slope) does not improve model performance. I can only assume that this is the reason why land surface parameters were not included in this study. This choice needs to be clearly mentioned in the manuscript and it should at least be discussed whether this assumption is also valid on global scale.
Reply 1:
We would like to thank the referee for raising the question regarding the optimal set of predictor variables used as input for the machine learning based global runoff estimator. As pointed out by the referee this choice is motivated by the findings of GS15 who showed that including additional atmospheric drivers and land-parameters resulted only in marginal gains in model performance.
Generally speaking, adding additional surface parameters as predictors (e.g. soil texture and slope) might improve the model skill. However, the search for an optimal extended predictors set would go clearly beyond the guidelines of ESSD, which emphasize the need to focus on the production and presentation of novel datasets of global relevance using already established methods (and not to present methodological or process oriented research). Instead, the scope of the presented paper is to apply a setup that was already successfully to a new data source at the global scale. We therefore refrained from conducting an extensive search for additional relevant predictor variables, although this would admittedly be an interesting topic for an independent research project. Finally, the reasonable performance of the GRUN model indicates that the focus on precipitation and temperature is justifiable for this global scale analysis.
The argument for limiting the predictors to P and T will be made clear in the revised manuscript in Section 4.1 ("Model Setup").

Comment 2:
GRUN runoff estimates are compared to those from an ensemble of global hydrological models. The ensemble actually consists of global hydrology and water resources models, land surface models and dynamic vegetation models. Neither type was specifically developed or is trained/calibrated to accurately predict runoff at grid cell level - and this is without question where the strength of GRUN lies. This needs to be discussed in a revised manuscript.
Reply 2:
Thank you for pointing out this fact. We will discuss this point in Section 5.3 ("Benchmarking against global hydrological models") of the revised manuscript.

Comment 3:
GHMs have mainly been validated against discharge observations from catchments >10000 km2, acknowledging the uncertainty in runoff estimates at grid cell level originating from model structure and parameters, climate input, and land surface properties. To make for a fair comparison, I challenge the authors to repeat the validation experiment in section 5.3 for the GRDC discharge records.
Reply 3:
Thank you for this comment. We will include this comparison for all ISIMIP experiments that are driven by GSWP3 atmospheric forcing in the supporting information of the revised manuscript. The overall results of the evaluation did not qualitatively change.

Comment 4:
The authors argue that GRUN estimates do not include human interference on runoff generation, thus that differences between the GRUN reconstruction and in-situ observations could potentially be used to identify streamflow stations which have a hydrological regime very different from the naturalized flow (P14, L9-10). I strongly disagree given that the screening procedures used are capable of identifying break points in the discharges records caused by the installment of (large) water infrastructure, e.g. dams and diversions, while they likely miss more gradual changes. Consequently, the discharge time series used for training the model are a blend of near natural catchments and those impacted by human activity, and it is unclear to which degree human impact is actually implicitly represented in the reconstruction. This needs to be clearly stated and discussed in the revised manuscript.
Reply 4:
Thank you for this comment. We agree that the change point detection methods are mostly sensitive to the installment of large infrastructure of water abstraction/storage and not to e.g. the impact of gradual land use change. We therefore agree that GRUN is likely not entirely free from effects of human activity as mentioned by the reviewer. It is, however, also important to remember that the temporal variability of GRUN is exclusively driven by the considered atmospheric forcing data. It is therefore our evaluation that GRUN is relatively close to near-natural conditions and clearly differs from flow conditions that are heavily impacted by human activities (see e.g. Fig. 5 and associated discussion).
We will discuss all these limitations in more detail in the revised Section 5.5 ("Limitations of GRUN").

Comment 5:
The extrapolative power of GRUN is somewhat overstated (in particular P14 L7-16). The developed RF model is (in the best case) capable of predicting short-term changes in monthly runoff/water availability as a function of (changes in) precipitation and temperature with all other boundary conditions constant. Changes in water availability originate from a variety of drivers, e.g. water abstraction/diversion, land use/cover change, reservoir management, and in many cases climate may not be the most important one.
Reply 5:
Thank you for pointing out that some of the discussion was formulated a bit optimistic. We will relax the wording in the revised document, also putting more emphasis and highlighting caveats and limitations of the GRUN data product.

Comment 6:
The authors evaluate the performance of GRUN in a cross-validation exercise. The corresponding maps displaying six performance metrics show clear regional differences in model performance, while numbers are only provided on global level (Table 1). In order to provide a more detailed picture of model performance for potential users of the data-set, I recommend to report numbers (Q5,Q50,Q95) at sub-global level, at least for the CV-SPACE experiment. This could be climate regions or the SREX regions already used by the authors.
Reply 6:
Thank you for this suggestion. In the revised supplementary material, we will provide two tables summarizing the skill distribution (Q25, Q50, Q75) for SREX regions and Koppen Geiger climate zones based on the results of the CV-SPACE experiment. We will also provide boxplots showing the skill distribution for various Koppen Geiger climate zone and SREX regions (Figures S4 and S5).

Comment 7:
Minor and technical comments …
Reply 7:
Thank you very much for identifying all these issues. We corrected all these points in the revised version of the paper. The resolution of the maps has been increased and the graphics will be provided as vector format for the final production of the paper.