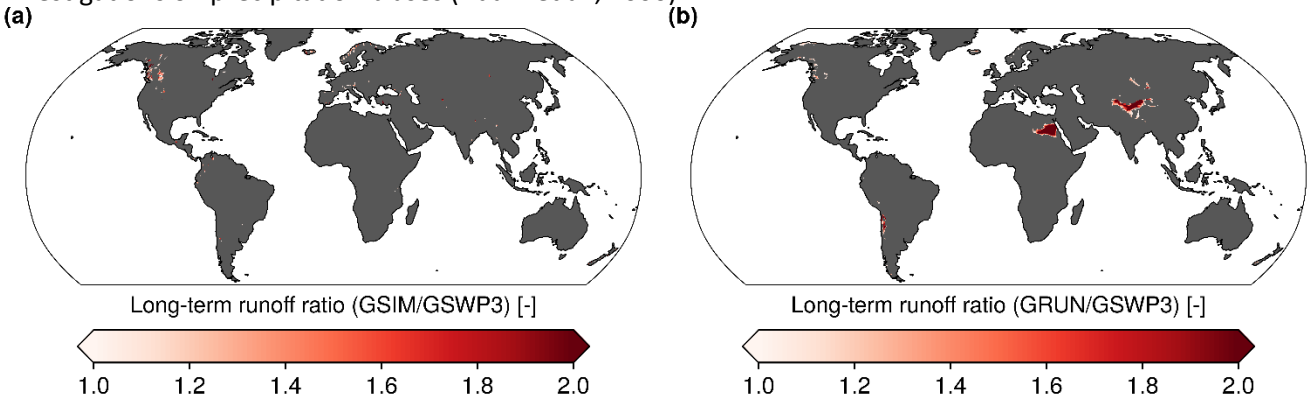
We would like to thank the anonymous reviewer for the open and overall positive evaluation of our manuscript. In the following, we provide point-by-point answers to the reviewer's comments. For the sake of clarity, we first repeat the reviewer's comments (in italic) and then provide our response.

Comment 1:
The area of basins selected for fitting the random forest model is well below the size of a grid cell. That means that the discharge in these basins might be driven by sub-grid precipitation (and climate) that is not reflected at the grid-scale.
Reply 1:
We agree with the reviewer that the discharge at any catchment represents the average runoff response of smaller subdomains, all affected by local variability in weather, climate and topography. However, the availability of global weather observations also constitutes a limiting factor in a global study, i.e. the effective resolution of the considered atmospheric data may be even coarser in regions with limited data availability. In addition we would like to point out that, we model monthly averages that have much less sub-grid variability compared to hourly or daily time scales. We also note that the model is not trained on individual catchment responses but on the mean response over a given grid cell (averaged over the catchments falling within that grid cell, see section 3.2). It is therefore our evaluation that although sub-grid variability may contribute to estimation uncertainty, this is less critical than other uncertainties (e.g. related to the quality to the atmospheric data) in the context of our investigation.

Comment 2:
As a sanity check, one should compare the observed discharge against the gridded precipitation. Perhaps, even filter out the basins where the annual discharge is higher than annual precipitation? If not, the results should show the distribution of discharge over precipitations in basins with observed discharge.
Reply 2:
Many thanks for suggesting to check the physical consistency of the runoff data and the precipitation product used to create the estimates. In the vast majority of cases, the modeled annual discharge is lower than the annual precipitation. However, while such comparisons are principally informative and help to gauge the limitations of the resulting data-product, care needs to be taken when interpreting the results. For example, both the considered streamflow observations as well as the employed precipitation data may contain biases, which might be implicitly corrected by the statistical model. As a result, there may be instances where long term mean runoff exceeds precipitation: see e.g. Fig. 1 in Gudmundsson et al. (2012) for such a comparison and the associated discussion related to underestimation of precipitation in regions with complex topography. The figure below shows the grid cell locations with a runoff ratio (Q/P) larger than one for a) the training dataset and b) the final GRUN product. In many mountainous regions, GRUN seems to account for possible bias in either the precipitation forcing or streamflow observations, which is qualitatively in line with previous investigations on precipitation biases (Adam et al., 2006).



In the revised Section 5.5 (“Limitations of GRUN”) we now point the readers to possible inconsistent water balance estimates in mountainous regions caused by bias in the precipitation or streamflow data used. Finally, we benchmarked GRUN against ISIMIP2a river flow simulations in large river basins. The results will be included in the revised paper and reveal that GRUN performs better in term of bias than any of the considered GHMs.

Comment 3:
The study would benefit by including a figure showing/comparing the distribution of data values in whole and training dataset. It is known that random forest cannot predict outside the range of data values used in training.
Reply 3:
Thank you for the suggestion. We will provide cumulative distribution functions of runoff and standard anomalies, for different Koppen-Geiger climate zones, for both CV-SREX and CV-SPACE cross-validation experiments (Fig. S6 and S7).

Comment 4:
There is no explanation on why the predictors included precipitation and temperature only, especially considering that the co-authors have produced a long-time series of terrestrial water storage variations at the global scale.
Reply 4:
Thank you for the interesting suggestion to also include the century long terrestrial water storage (TWS) reconstruction (GRACE-REC) (Humphrey and Gudmundsson, 2019) as forcing. Note, however, that the TWS reconstruction is based on the same meteorological data as GRUN. Consequently, although the TWS reconstruction is based on a different method, our evaluation is that it does not contain any additional information that the Random Forest algorithm would be able to capture. We acknowledge that TWS is of course essential for runoff generation, and this is implicitly accounted in GRUN by including antecedent monthly precipitation and temperature in the statistical model.
The choice of the predictors will be better explained in the revised Section 4.1 (“Model setup”).

Comment 5:
Several other factors such as vegetation properties, topographic indices, soil properties, and so on contain useful information on runoff generation mechanisms. Previously, these variables have been shown to include information on baseflow characteristics globally (see H. Beck et al, 2013 (WRR), and 2015 (JHM)).
Reply 5:
We recognize that such factors are important when estimating streamflow characteristics from sub-daily up to weekly time scales. However, in this paper, we focus on the monthly amount of water draining from grid cells of 50 km spatial resolution. In this context Gudmundsson and Seneviratne (2015) demonstrated that the performance is already high when considering P and T as drivers only and that inclusion of selected land properties did not improve the accuracy of the estimate. One reason which could explain why land properties do not improve our model predictions is that the random forest algorithm can synthesize information about the average climatic conditions governing e.g. vegetation properties from the input precipitation and temperature data, such that the margin for model improvement is already small (also see response to Comment 1 of Reviewer 2).

Comment 6:
The validation of the predicted runoff against GRDC should include the comparison of mean flow in addition to the time series (and metrics based on it).
Reply 6:

In the revised manuscript, we will provide spatial maps of relBIAS (under/over estimation of the mean flow), rSD, R2, NSE, $R2_{anom}$, $R2_{clim}$ for large river basins from GRDC (Fig. S8). Figures S9 to S12 will also compare river discharge estimates based on GRUN against equivalent estimates from ISIMIP2a GHMs simulations.

Comment 7:
The mean discharge and runoff over the basin should be equal if there is no long-term change in river storage. If the mean runoff from GRUN is similar to GRDC observed discharge, one can infer that the poorer performance in the time series is due to exclusion of river routing.
Reply 7:
Yes, as stated in the paper, we did not account for river routing when evaluating GRUN using GRDC river discharge observations. Disagreement in the time series is only visible for very large river basins where routing should be taken into account (e.g. Amazon) or in basins where water abstractions are very high (i.e. Nile).

Comment 8:
Is it possible that the poorer performance in the humid basin is related to lack of data values similar to those in Amazon and Congo in the training dataset?
Reply 8:
We do not observe a poor performance in humid basin. Actually, GRUN shows slightly lower performance in arid regions. We recognize that additional streamflow data in these regions would likely increase the accuracy of GRUN, which is one of the reasons underlying our concluding remark of the paper that calls for increased mobilization of discharge data around the globe.

Comment 9:
In the comparison of global volumes, the GRUN also falls in the lower range. I am curious if this underprediction is coming from the lower values in the tropics which dominate the contribution to global runoff. This can be checked by having a grid-to-grid comparison (something like hexbins in Figure 3) against previous products or model simulations
Reply 9:
In figure 8c we provide the spatial difference between the long-term runoff of GRUN and the multi-model ensemble mean (MMM) of ISIMIP2a hydrological simulations. We identified a hotspot region in Bangladesh were GRUN estimates are consistently lower than the hydrological model simulations. This deviation is also visible in Fig. 8b in the latitudinal band between 20 and 30° N. GRUN falls in the lower range of ISIMIP2a especially because of this region and a tendency for lower runoff rates (compared to MMM) in the subtropics (see Figure 8b). In the tropics, the GRUN reconstruction compares similarly to MMM.

Comment 10:
If the underprediction is clear, it makes me wonder if predicting the runoff ratio (runoff to precipitation) would be beneficial.
Reply 10:
The scope of study is the reconstruction of monthly runoff rates, which are closely related to the observable streamflow, using an already established method, which is in the spirit of ESSD. Focusing on the runoff ratio instead would imply that new methods have to be developed which goes beyond the scope of ESSD. Still, we thank the reviewer for this suggestion which would be worth investigating in an independent piece of research.

Comment 11:
The GRDC discharge observations are likely to be affected by anthropogenic uses, as the manuscript rightly points out several times. Therefore, wouldn't it make sense to evaluate the GRUN product against previous mean total runoff from Beck et al., 2015 which uses fewer observation stations but more predictors (also based on neural network)?

Reply 11:
In figure 3c we report the agreement between observed and predicted long-term mean runoff of 5544 grid cells. In this study, R2 is slightly higher (0.92, Fig. 3) than in Beck et al., 2015 (0.88, Table 5).
We carefully thought to compare QMEAN from Beck et al., 2015 against GRUN. However, QMEAN in Beck et al., 2015 is obtained by establishing a regression between QMEAN (estimated for every catchment over different time periods and number of timesteps) and (time invariant) multiple predictors. Therefore, QMEAN in Beck et al., 2015 is not representative of a specific time period and thus we opted to avoid such comparison.

Comment 12:
Wouldn't it make sense to evaluate the GRUN product against mean ET from, say, LandFlux-EVAL, part of which is coming from satellite observations?
Reply 12:
In this paper, we evaluated GRUN against observed runoff and streamflow measurements. Generally, we expect discharge observations to be more accurate than global-scale ET estimators, which is the reason why we did not compare mean runoff against mean ET. Note also, that a comparison similar to that suggested by the reviewer was conducted for the European case by Gudmundsson and Seneviratne (2015)(see Fig. 11). This analysis showed that data-driven runoff reconstructions are generally consistent with ET estimates, but also highlighted the large uncertainties associated with global ET products.

Reference

Adam, J. C., Clark, E. A., Lettenmaier, D. P. and Wood, E. F.: Correction of global precipitation products for orographic effects, J. Clim., 19(1), 15–38, doi:10.1175/JCLI3604.1, 2006.

Beck, H. E., de Roo, A. and van Dijk, A. I. J. M.: Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments, J. Hydrometeorol., 16(4), 1478–1501, doi:10.1175/jhm-d-14-0155.1, 2015.

Gudmundsson, L. and Seneviratne, S. I.: Towards observation based gridded runoff estimates for Europe, Hydrol. Earth Syst. Sci., 19(6), 2859–2879, doi:10.5194/hess-19-2859-2015, 2015.

Gudmundsson, L., Wagener, T., Tallaksen, L. M. and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resour. Res., 48(11), 1–20, doi:10.1029/2011WR010911, 2012.

Humphrey, V. and Gudmundsson, L.: GRACE-REC: a reconstruction of climate-driven water storage changes over the last century, Earth Syst. Sci. Data Discuss., (February), 1–41, doi:10.5194/essd-2019-25, 2019.