We would like to thank Anonymous Referee #3 (AR3) to write this review. Below, we will reply to the points made by AR3, with the comments from AR3 in black, our response in blue.

The paper "Seasonal discharge response to temperature-driven changes in evaporation and snow processes" aims at testing the hypotheses that both seasonal changes in snowmelt and enhanced evaporation can exacerbate low flows, and that changes will increase with temperature under realistic warming. I think this study fits very well with the scope of ESD journal and it addresses an important and timely topic. However, the authors should first address some major issues before this manuscript can be considered for possible publication.

We would like to thank the reviewer for the time and the valuable suggestions. We are happy to read that the reviewer considers the topic we address timely and important. We agree with the issues raised by the reviewer and believe that addressing them will improve the quality of the manuscript.

My first concern is related to the structure and readability of the paper. I found the description of the method really poor, fragmented, and the results section is a mix of both findings and methods. Overall, I believe that this paper structure makes the manuscript confusing and difficult to read. Why not add a "case study" and "Experimental setup" sections before describing the results? Also, the authors described figure 2.a and then they jumped to figure 3, while the remaining part of figure 2 is described only after. A solution could be to split figure 2 into different ones. Try to be more consistent.

We agree with the point, and it was also brought up by the two other reviewers. We will extend the methods section, and add a better description on the experimental setup. We agree to split figure 2 into two separate figures, and will use another figure to visualize the forcing-swapping (see below).

My second concern regards the swapped method introduced in this study to understand how the individual forcing variables affect the hydrological cycle. I found this approach quite atypical, and an adequate justification should be included for why such an approach is employed. I am not deeply familiar with this swapped approach, but there is no reasoning provided for why such an approach should be preferred over other statistical approaches.

We hope that the new description of our approach (see below) removes your concerns. If not, please let us know how to improve this.

Another serious issue is the model structure. Several major hydraulic works and flood control measures were constructed over the years in the Rhine basin, strongly modifying its hydrological cycle and flood responses. How are those structures included in your distributed efficient hydrological model? This can be a major issue as major hydraulic works may have a higher influence than the forcing variables analyzed in this paper, thus compromising the findings and conclusions of the study.

It is correct that the Rhine model does contain management structures. Despite this, we show that the model is able to correctly simulate the discharge across several subbasins. This performance will likely reduce when looking only at hydrological extremes, which we are not investigating in this study. In this study, we investigate the hydrological response to changes in temperature, but how river management will change under changing condition is also a large unknown.

Linking to the structural issue of the manuscript, the authors can clearly mention in the new "Experimental setup" section that the swapped and changing-temperature approaches are meant to answer the two hypotheses of this study (see introduction). Moreover, to further improve the readability of the paper, the authors could better connect the two hypotheses of the introduction with the results summarized in the conclusions. Right now everything is there, but it takes quite some time to grasp the main take-home message.

This is a good point, and we will link the results back to the hypotheses in the conclusion section.

The authors mentioned in the conclusions that "Here we selected a resolution of 4×4 km, so we can use the ERA5 forcing data without downscaling methods (adding uncertainty and potential errors)". However, if ERA5 has a resolution of 0.25degree, how it is possible not to downscale the dataset to adapt it to a higher resolution of 4km?

This is indeed described in an ambiguous way. We did use bilinear downscaling, but argue that the scale difference is still small enough to not have to used more advanced downscaling methods. We will clarify this in the next version of the manuscript.

". . . yet these results can be interpreted for the many different basins around the globe depending on both rain- and snowfall". This sentence should be rephrased as you do not know what can occur over other basins with totally different characteristics. The results of this study cannot be generalized to other studies without proper large-scale validations. The same applies to other generalizations introduced in the conclusions. Include study limitations in the conclusion section

We agree that this may be misleading. We will clearly state the limitations and adapt our conclusions in the next version of the manuscript.

1 Methods - Forcing swap

In the first experiment, we aim to understand how each forcing variable can explain the resulting changes in discharge, and their relative importance. To perform this, we setup the experiment according to the conceptual overview presented in Fig. 1. In order to investigate how temperature influences evapotranspiration and snow processes separately, we perform model runs in which the total temperature change is splitted into temperature effects on evapotranspiration ("Changed T_{evap}") and snow processes ("Changed T_{snow}"). In addition, another run is performed with only changes in P ("Changed P"), so that these individual runs can be compared to a run where all changes in forcing are enabled (2010s forcing). The resulting simulated discharge is compared to the 1980s run, to determine the discharge change. In this way, we can evaluate the relative impact of each forcing variable on the discharge.

We sum the discharge changes of the three forcing swapped runs, to obtain $\mathrm{Sum}\Delta$:

$$\mathrm{Sum}\Delta = \Delta \mathrm{Q_P} + \Delta \mathrm{Q_{T\ snow}} + \Delta \mathrm{Q_{T\ evap}} \tag{1}$$

where $\Delta \mathrm{Q_x}$ represents the discharge difference of the forcing swapped simulations. We can use this $\mathrm{Sum}\Delta$ to study how well it explains the 2010s run, by comparing it to $\Delta \mathrm{Q_{2010s}}$. We hypothesise that when $\mathrm{Sum}\Delta$ is equal to $\Delta \mathrm{Q_{2010s}}$, the effect of the forcing is additive, and together explain all differents. We will refer to this as the direct effects. In the case of a discrepancy between $\mathrm{Sum}\Delta$ and $\Delta \mathrm{Q_{2010s}}$, this can be attributed to interaction between the three forcing components. We will refer to this as indirect effects. We define $\mathrm{Sum}\Delta$ to have explanatory value when it has the same sign as $\Delta \mathrm{Q_{2010s}}$. We calculate the contribution of the direct effects (ϕ) using the following equation:

$$\phi = \begin{cases} \frac{\min(\mathrm{Sum}\Delta, \Delta \mathrm{Q_{all}})}{\max(\mathrm{Sum}\Delta, \Delta \mathrm{Q_{all}})}, & \text{if } \mathrm{sign}(\mathrm{Sum}\Delta) = \mathrm{sign}(\Delta \mathrm{Q_{all}}) \\ 0, & \text{if } \mathrm{sign}(\mathrm{Sum}\Delta) \neq \mathrm{sign}(\Delta \mathrm{Q_{all}}) \end{cases} \tag{2}$$

This value can then be used to calculate the relative (direct) contribution of each forcing variable, using the following equation:

$$\phi_\mathrm{x} = \frac{\mathrm{abs}(\Delta \mathrm{Q_x})}{(\mathrm{abs}(\Delta \mathrm{Q_P}) + \mathrm{abs}(\Delta \mathrm{Q_{T\ snow}}) + \mathrm{abs}(\Delta \mathrm{Q_{T\ evap}}))} \cdot \phi \tag{3}$$

where $\Delta \mathrm{Q_x}$ should be replaced by $\Delta \mathrm{Q_P}$, $\Delta \mathrm{Q_{T\ snow}}$, or $\Delta \mathrm{Q_{T\ evap}}$.

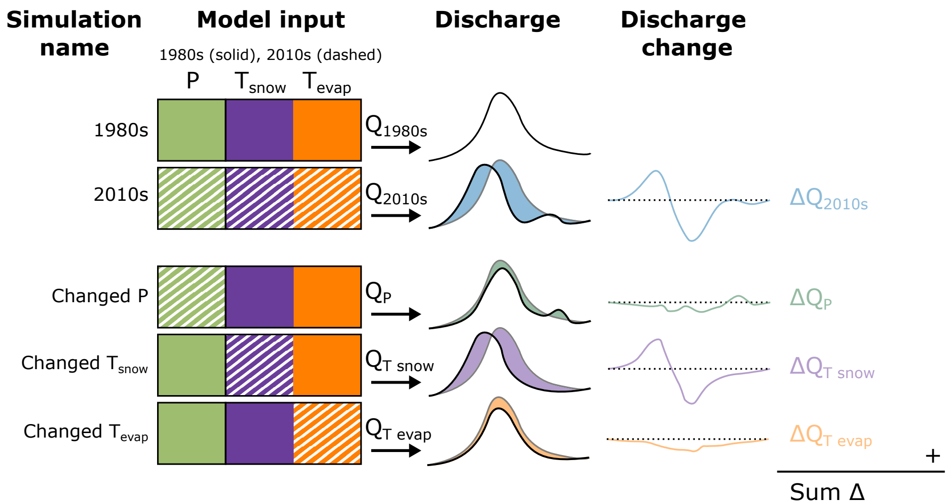

Figure 1. Conceptualization of the forcing swap experiment, showing the different simulations (rows) and steps in the analysis (columns). The different forcing variables are visualized as colored blocks, where the solid and dashed boxes indicate forcing data from the 1980s and the 2010s, respectively.