We thank the reviewer for constructive comments and provide our answers hereunder.

RC3

This paper introduces modelling climate change impacts to river temperatures in Switzerland. To do this, the authors conducted a multi fidelity modelling method which uses statistical pattern recognition to estimate river water temperatures under climate change and thereby close the aforementioned spatial gap by determining, in an automated manner and on a country-wide scale, how future river water temperatures are likely going to change.

The authors frequently refer to their method as novel. I suggest to remove all occurrences of this claim of novelty. Simply describe the model used. Some would argue that the discipline of stream temperature modelling has advanced beyond the use of air temperature and discharge alone for predicting river temperature, regardless of whether it is focused on current versus future climate. While it may be practical for a nationwide attempt, and in that case also 'efficient', it is not necessarily "novel".
The word novel occurred once on line 93, it has been removed from the manuscript.

The reason physically-based models require a lot of data is because they attempt to represent mechanisms and therefore attribute causality of rising river temperatures. River temperatures are a function of many processes beyond simply river discharge and air temperature, as has been discussed in recent literature. The limitation of the "efficient" model approach is that many, many physical drivers of river warming are completely ignored. In predictions of stream temperature, simplifying the "*more complex processes into purely empirical parameters"* often involves using lumped parameters and lumped heat exchange coefficients which ignore aspects of climate change, especially with respect to the shortwave and longwave radiation balance, and increases in atmospheric emissivity which is driving the air temperature warming. There is not a single mention of any of this. The authors simplify the controls to the energy balance as being based on discharge and air temperature, which is not complete, nor does it use best-available-science. If the authors make simplifications in processes, and vary the number of parameters used across their different simulations in order to get a nation-wide dataset for Switzerland, they need to be very clear about this approach and also be upfront about the many, many limitations of their results.

Physical drivers are not ignored in this kind of study, they are included indirectly through parameterization. One can correctly argue that being included under a parameter constitute being frozen in time. Since we use air temperature as input from Ch2018, we capture part of the changes over time in the surface heat budget relevant for water temperature. A more complete heat budget was included for snow and glacier melt in Hydro-CH2018, which provides the discharge to this study.
Apart from the effect of air temperature on water temperature, the models additionally resolve the effect of river discharge, depth, thermal signals from tributaries, inverse stratification in lakes during winter, and seasonal cycles.
The manuscript has been updated in both the introduction, method and discussion sections to make it clearer for the reader the limitations and advantages of our approach.

Line 60 to 107 in section 1

A common challenge for model-based studies is the question of the optimal model to use. In surface hydrological applications, models can broadly be split into two major groups: process-based and statistical/stochastic models (Benyahya et al., 2007). Process-based models are based on physical equations and can resolve many hydrological processes in a physically robust manner, from the local to the catchment scale. However, albeit physically more robust, process-based models generally require a significant amount of input data and computational resources for the simulation of hydrological processes on the catchment scale, therefore limiting their applicability for climate change analyses on national scales. Statistical/stochastic models, as opposed to process-based models, are data driven, that is, are based on empirical relationships between input and output data. While they are physically less robust, their advantage lies in their relative simplicity and limited data requirements, sacrificing detail for increased repeatability and spatial coverage. However, in order to build on the efficiency of statistics whilst preserving a clear physical basis, as a compromise between the two major model groups, a sub-group of semi-empirical models, which employs physically meaningful equations but simplifies the more complex processes into purely empirical parameters, was developed (Piccolroaz et al., 2013). These semi-empirical models are ideally suited for hydrological climate change projections, as they provide much more robust projections compared to purely statistical approaches but simultaneously allow for a more comprehensive analysis than process-based models by enabling multi-model climate change ensemble analyses (La Fuente et al., 2022; Meehl et al., 2007).

The study of climate change includes the investigation of physical processes on global, regional and local scales. As scales change so too does the required level of detail needed to resolve the different water cycle components that are relevant on the respective scale. An ideally suited approach to address this challenge in hydrological modeling is a multi-fidelity model framework, which combines multiple computational models of varying complexity in an automated selection framework that ensures robust predictions while limiting the computation to only the necessary level of detail (Fernández-Godino, 2023). The use of process dependent fidelity ensures proper representation of physical processes on regional to local scales while keeping computational costs to a minimum. Multi-fidelity modeling is especially useful when acquiring high-accuracy data is costly and/or computationally intensive, as is the case for climate change impact assessment on the hydrological cycle.

Given the past and future changes to Swiss river water temperatures and considering both the high sensitivity of aquatic species to river water temperatures and the increasing demand for river water by agriculture, industry and society as a whole, it is critical to obtain a robust spatial and temporal understanding of the temperature increases that are expected for the many different rivers and streams of Switzerland. Here, we developed an efficient multi-fidelity modeling method guided by statistical pattern recognition to estimate river water temperatures under climate change and thereby close the aforementioned spatial gap by determining, in an automated manner and on a national scale, how future river water temperatures are likely going to change. Compared to previous projections of climate warming in Swiss rivers (Michel et al., 2022), the simplified multi-fidelity modeling approach not only enabled to cover the national scale (+90%) but also further thermal regimes (here 5, previously 2) and based on 22 GCM-RCM chains (previously 7). By grouping catchments together via statistical pattern recognition, we were able to classify rivers (including spring-fed rivers) into 5 different thermal regimes, improving model results by allowing for optimal model selection at each station and enabling regime-specific analyses. The effect on warming by changing river discharge was

investigate through a hysteresis analysis. Additionally, we introduce the *extreme event severity* index as an analytic tool to evaluate the change in thermal extreme amplitude.

Line 261 to 273 in section 2,5 Surface water temperature model setup

Both models include up to eight parameters (a_1 to a_8) which are fitted towards measured data. Apart from the effect of air temperature on water temperature, the models additionally resolve the effect of river depth, discharge, thermal signals from tributaries, inverse stratification in lakes during winter, and seasonal cycles. Model complexity, i.e. how many processes are directly being resolved by the models or indirectly included through parameter estimation, can be varied by removal of one or more of the additional processes listed above, resulting in the use of 8, 7, 6, 5, 4 or 3 parameters. Depending on local conditions, model performance can be improved by the removal of processes which play a minor or insignificant role for water temperature. Where this simplification with removal of parameters was done (Table B2), removed processes plays a minor role for the simulation of water temperature as evident from decreased model performance while being included. For additional information about *air2stream* and *air2water* see Appendix A and Piccolroaz et al. (2013) and Toffolon & Piccolroaz (2015).

Lines 302 to 318, Section 2.6 Trend correction

Empirical models generally predict less warming in the future compared to physically based models, the primary reason being underrepresentation of the thermal catchment memory, including snow and ice (Leach & Moore, 2019). To quantify how good the models air2stream and air2water, which both lack deterministic considerations of snow and ice melt, are able to recreate past trends, we compared trends from river water temperature measurements and corresponding modeled temperature trends between 1990 and 2019. On an annual basis, this comparison was possible for 25 out of 82 river stations, consisting of 9 Downstream Lake, 7 Regulated, 7 Swiss Plateau, 2 Alpine, and 0 Spring thermal regime river stations. Stations were selected with a 30 years of continuous data requirement in air and water temperature and river discharge. Only statistically significant trends ($p < 0.05$) were considered.

Both air2stream and air2water underestimate the annual temperature trend during the reference period on average by 0.14 and 0.11 °C per decade, respectively. For air2stream, the annual trend bias is smallest for the Swiss Plateau thermal regime (0.09 °C per decade) and largest in the Alpine thermal regime (0.17 °C per decade). Seasonally, the trend bias is largest from June to August and September to November, whereas, especially for air2water, the bias is small from December to February and March to May.

Line 580 to 596

4 Discussion

4.1 Multi-fidelity modeling approach

The use of semi-empirical models by definition means that some of the physical processes affecting heating is simplified under parameterization and some are directly resolved. The models air2stream and air2water resolve the effect of river depth, discharge, thermal signals from tributaries, inverse stratification in lakes during winter, and seasonal cycles. Parts of the heat balance (e.g. short and longwave radiation) is thus not allowed to change as climate

change in our study. However indirectly we consider heat budget changes by using high quality air temperature and discharge projections as input. Glacier retreat is included in the hydrological models providing discharge projections to this study (eg. Muelchi et al., 2021), however for temperature this effect is only indirectly considered in air2stream and air2water through reduced water availability in summer. The effect of high altitude warming as snow and ice recede is not included. Therefore as the cooling caused by melt water recedes, it is expected that warming in high altitude rivers is larger than projected in this study. Yet the lower fidelity water temperature model approach using high-fidelity climate/hydrological model outputs as input enable the important principle of multi-model ensemble, comparison and analysis that is required for robust climate change impact assessments (Duan et al., 2019).

Lines 122-125: It is unclear how many years of actual data were used. This must be clarified. In one sentence, they say at least 1 year, in another sentence they say *"data should preferably cover 30 years"*. Authors need to specify which simulations used which timespan of datasets, as this is a fundamental influence on the accuracy of the predictions you are reporting in your Results section.

The duration of datasets used for calibration and validation are given in Table B2 and are described in section 2.5. The following section has been moved from section 2.5. to 2.1.
"Temporally overlapping, daily averaged near-surface air temperature and river discharge measurements spanning the 30-year reference period of 1990 to 2020 were used as calibration data, while for validation the data from 1980 to 1990 were used (Table B2 in Appendix). By choosing to use the most recent data for calibration rather than validation ensures that recent local climate conditions are carried into future projections (Shen et al., 2022). For the few cases where no forcing data for calibration did exist between 1990 to 2020 (Table B2), validation was deprioritized and calibration performed for the 1980-1990 data."

Lines 151-153: The authors state, "*For monitoring stations at which historic river discharge data or future river discharge projections weren't available, only future near-surface air temperature projections were used to simulate water temperature.*" This is a major limitation. For how many stations did the authors predict river temperature only from air temperature alone? And how do you correct for the fact that come used discharge and some didn't use discharge, but you are presenting the results of those two different simulation approaches as being equal in your Results section?
Lines 154-156: Many studies have demonstrated that the resolution of the climate model data will influence your results. Here the authors state, "*Where climate projections were available at multiple different spatial resolutions (i.e. 0.11° and 0.44°), only one model, as indicated in Table 1, was included in the analysis, following the approach of Muelchi et al., 2021.*"
These two items above both will affect the model results, potentially significantly. Sometimes the authors use air temperature and discharge to predict river temperature. Sometimes the authors use only air temperature to predict river temperature (many authors have shown this is not sufficient). Sometimes the authors used 0.11° spatial resolution and sometimes they used 0.44° resolution. How are the results defensible and comparable?

We agree with the reviewer that discharge is an important parameter for modeling water temperature and should be used wherever applicable. Here at 47 out of 82 stations we could use river discharge. 35 stations were modeled without discharge (Table B1).
Our study combines a wide variety of datasets (measured and modelled) with varying degree of data availability and accuracy. In the multi-fidelity modeling approach, we do not rank the inputs from climate models or measurements. Nor do we select “representative” model runs or climate scenarios.
Instead, the simplicity of this method enables us to use a wide range of climate models, flow models and water temperature models. Through the use of ensembles and combined analysis inconsistencies and biases included in all data and models are smooth out. This follows recommendations from recent climate model downscaling in Switzerland: “To account for the inherent climate model and greenhouse gas scenario uncertainty, we also advise users to employ a maximum number of CH2018 simulations (CH2018 project team).”
A trend correction was preformed to correct for seen discrepancies between our models and measurements during the reference period. The correction needed was smallest for the *air2water* model (18 stations) compared to *air2stream* (17 stations). The *air2water* model which works completely without discharge outperformed *air2stream* downstream of lakes, this indicates that despite lacking the input of discharge we could model the impact of climate change satisfactionally without river discharge see section 2.6 above.

Lines 165-172: Again, the deviation across methods raise concerns for presenting comparable results. This study employs large datasets which require some level of computational proficiency, but it appears they did not employ spatial interpolation methods of weather data across elevation or across distance. It is very common (and not difficult) to employ spatial interpolation methods of time-series weather data to a particular river location, in order to produce more accurate results at a specific distance along a river. The authors state: “*Meteorological stations were subsequently paired with hydrological stations such that (a) the horizontal distance between river and meteorological stations was minimal (criterion "DIS"), (b) the meteorological station was representative of the conditions in the upstream drainage area (criterion "DRA"), and (c) the elevation difference didn't exceed a reasonable threshold of 200 m (criterion "ELE"). Where possible, all three criteria were adhered to. For situations where the closest meteorological station was either not fulfilling DRA or ELE, the DIS criterion was evaluated only for stations which fulfilled both DRA and ELE.”* While this explanation is, in theory, reproducible, I am not sure that adjusting the criteria on a station-by-station basis is defensible. Authors need to address this.

Both the *air2stream* and *air2water* models use representative atmospheric forcing for a drainage area above a certain point to model water temperature. Simulations are thus conducted towards all relevant heat transfers taking place upstream of this point as captured by water temperature measurements obtained at the point. Thus, for these models the exact location of atmospheric forcing in the drainage area is of minor importance. It is far more important to have representative meteorological conditions, hence the selection criteria’s above. Anny remaining inconsistency between the actual dataset used as input and how local atmospheric conditions affect water temperature in the drainage area, are compensated for in the calibration of the two models with up to 8 parameters. By preforming spatial interpolation of meteorological data and climate model results unknown biases are created, especially in

settings with pronounced relief (the Alps), bias which increase with the distance from each station.
In Ch2018, regional climate models were downscaled with quantile mapping towards measurements at local stations. Naturally, the quality of this downscaling improved towards the meteorological stations with minimal climate model bias right next to each station. Thus, by selecting to use the downscaled climate model data delivered at the location of the meteorological stations, climate model bias was minimized.
For processes such as stream discharge, spatial and temporal distribution of precipitation and snow/glacier melt is more important compared to heat budget processes for water temperature modeling. Spatial dependency was considered in Hydro-Ch2018, those discharge projections we use here.

Lines 322-324: What do the authors mean by "shape-preserving interpolation" across multiple days without data, and where is this interpolation method presented in this paper? Authors state: "*Before adjusting the water temperature model output from 1990 to 2099, Bcs was combined into a continuous dataset by filling in the 3- to 5-day gap in between each season with shape-preserving interpolation.*"
Now reads on lines 348to 352:
„Before adjusting the water temperature model output from 1990 to 2099, the seasonal Bc_s was combined into a continuous dataset *Bc*. To avoid a sharp shift in *Bc* between each season, a 3- to 5-day gap in between each season was smoothed with shape-preserving interpolation (Piecewise cubic Hermite interpolation, PCHIP; MATLAB© R2022a)."

Line 439: "*Considering only the far future"* ◊ what do the authors mean by "far future". Please clarify.
Far-future period (2070 to 2099), is defined in the manuscript.

The authors' most significant result is summarized by "*Climate change impact was heterogeneous between stations, yet common patterns were found within thermal regimes".* It is concerning to present results when each result was achieved through a subtle deviation from the methods, the spatial resolution of inputs, the handling of missing days of data, and even using different model inputs. In some simulations the only model input is air temperature. How can results and hysteresis loops be viewed as comparable across simulations by the reader, when the methods employed to get there were modified, changed, required deviation of some methods, used a different number of parameters in 'air2water'/'air2stream' (i.e. Line 695 "*adapting their parametrization complexity to the required level")*, or were slightly different methods across simulations?

The reviewer is correct that the methods differed for each station. However, this was an intended and needed but not a random process. The process, which is known as multi-fidelity modelling, selects for each station the best possible model according to the available data (in this context meaning the model with optimal model complexity as warranted by the data). It would of course be desirable to have an identical data basis for all stations, but this is the real-world, and in the real-world, this is absolutely never the case. Hence, in order to project river water temperatures for real-world measurement stations, one is left with the choice to either use the lowest complexity model for all stations, as warranted by the station with the poorest data basis for projection, which would lead to comparable but underwhelming projections, or

one can choose the optimal model complexity as warranted by the data availability of each individual station, producing, for all stations, projections with the highest fidelity. We chose, in agreement with the multi-fidelity modelling approach and philosophy, to compare the projections of all stations based on their highest fidelity model and data basis. This is the most appropriate approach to compare and judge projections for real-world stations. For more precise viewpoints we referee to our previous answers in this review.