Response to referee comment Anonymous Referee #1

We would like to thank Anonymous Referee #1 for the constructive comments concerning our manuscript. These comments will help to improve and clarify our paper. Our response to the comments can be found below (original comments in black, our response in blue).

The manuscript aims at investigating the hydrological response (in terms of evapotranspiration and runoff) to seasonal extremes of temperature and precipitation in five Swiss mesoscale catchments with varying elevation and glacierization. To assess the influence of spatial model resolution, the authors run the SPHY model in both high (500 m) and low (40 km, corresponding to a lumped simulation with a single grid cell) resolution and find that in the the distributed simulations model responses to climatic extremes can be much more pronounced. The topic fits the scope of HESS and is of scientific interest. The article is well written and both the methodology and the results are generally well explained and presented. I would therefore recommend publication of this article after considering a few, mostly minor, comments.

Thank you very much for the recommendation for publication. We will take the comments into account, and respond to those below.

P1 L4: It would be helpful if "hyperresolution" could be specified quantitatively (500x500 m) already in the abstract. Possibly the corresponding value in km could also be added for the coarse resolution.

Thanks for this suggestion. We will incorporate this in the abstract.

P1 L23f: Do you intend to say here that it is assumed that the coarse-scale anomaly matches the average fine-scale anomaly, or that the signs of the coarse-scale and the average fine-scale anomaly match?

We did mean to say that the sign of the coarse-scale anomaly is representative for range of fine-scale anomalies. We will modify this sentence to clarify this statement.

P4 L11: In which spatial resolution are the meteorological data sets originally provided?

The meteorological data is at a resolution of roughly 2x2km. We used bilinear interpolation to downscale this to the required resolution for the high resolution model. To resample the forcing data to the 40x40km pixel, we averaged all cells within this pixel range.

P4 L16: When resampling to the coarse scale, did you average only over the basin interiors or did you consider the entire data within the 40x40 km grid cell?

We averaged over the entire data present in the 40x40km grid cell. This does include cells outside the catchment. This method had our preference, since it better resembles the methods used by global studies.

P4 L19f: A calibration period of only two years seems very short, especially considering that you calibrated on monthly values. Is there a reason why you did not choose a longer calibration period, or why did you choose exactly these two years? In any case I would recommend to include a table with some performance measures (e.g. NSE, PBIAS, . . .) for the calibration and validation periods and both resolutions. Also, a plot showing observed and simulated discharge (either as a time series e.g.

for a period encompassing the four extreme events, or average monthly discharge over a given period) for all basins could be helpful.

We chose for a calibration period of 2 years in order to keep the uncalibrated period as long as possible. These years were selected since they contained a relatively wet (1999) and an average year (2000). By calibrating on those two years, we ensure that both the average situation and a more extreme situation are simulated correctly. We agree that Fig. 4 does not provide sufficient information to show the performance of the model. We will add the monthly simulated discharge values with Kling-Gupta efficiencies for the calibration period, and the monthly average discharge values as validation.

P9 L13f: "This might be related to the relative coarse monthly calibration time step": possibly also a combination of the coarse calibration time step, the short calibration period, and the choice of optimization function. Minimizing the sum of squares between observed and simulated runoff favors calibrating the model towards the periods with high runoff volumes, whereas larger relative (but smaller absolute) errors in low-flow periods tend to be suppressed. In basins with a strong runoff seasonality (as in the snow- and glacier melt dominated Alpine basins) this could be especially prevalent and might be one reason for the low DJF performance. Including a plot with the average discharge seasonalities (see my comment above) might also be insightful here.

This could also be a cause of the low winter performance of SPHY. We looked again at the simulated discharge values, and think it is also related to a model error. During winter periods, the model fails to simulate a constant baseflow, which is visible in the observed discharge. During these periods, the model assumes that most of the precipitation is falling and stored as snow, and is therefore not directly contributing to the discharge. We will add the average discharge seasonalities (see above).

P10 L6: Parts of the sentence are missing here.

We will fix this sentence in the revised version of our manuscript. We wanted to say that processes generating discharge do not seem to be consistently driven by either precipitation or temperature, but rather by a combination of both.

P10 L28-31: In the case of summer 2003, the areas with positive anomalies are probably only the glacierized areas?

Most of the cells with a positive response are indeed cells containing glaciers. There are a few cells however showing a positive anomaly which are not covered with glaciers. This is also visible in Fig. 7, where a number of cells falling in the “other” land cover class (sparse/bare vegetation) are showing a positive anomaly. This indicates that those cells are mostly dependent on the temperature, meaning that these cells where (partially) covered with snow.

P16 L1-4: As you describe the results for the Rhone basin here, maybe consider showing this basin also in Fig. 8.

Since the response of the Rhone basin is very similar to the response of the Reuss basin, we decided that it does not add any information and chose not to depict this graph. We noted the high DWD value for the Rhone basin for this particular period, and explain the cause of this high value.

Fig. 1: Please consider also showing the land cover distribution of the basins here (this could be helpful for, e.g., better interpreting Fig. 5).

This is the reason why we used a satellite image as background for this figure. We agree that this could be improved, and we try to incorporate this in Fig. 1.

Fig. 6: Looking only at the figure it is easy to overlook that not the absolute quantities but their standard deviations are shown. Maybe you could make this more clear by e.g. labeling the y-axes with "sd(ETa)" or similar.

We understand the potential confusion, and we will modify the axis label to emphasize that the standard deviation values are shown.

Fig. 8: Please label the basins here also with their names instead of the IDs.

An old version of this figure accidently made it into this manuscript. The latest version of this figure, with the basin names, will be in the revised version

P1 L12 and L23f: course -> coarse (3x)

Thank you for mentioning this typo, we will correct this.

P3 L32: runs on a fixed daily time step and a user defined spatial resolution

Thank you, we will correct this in the revised version.

P7 L16f and P14 L3-5: extend -> extent (4x)

Thanks again for spotting this mistake, we will change this.