Reviewer 2

A representative simulation of ROS melt events is important for improving hydrological modeling practice in snow dominated region. It is valuable to look into the future impact of ROS melt events under climate change. This is exactly what this work intends to address. However, the current manuscript is not yet ready for publication, due to two points :

> Response: We thank the reviewer for the thoughtful feedback which we have used to improve the quality and clarity of the manuscript.

1. This work utilized a calibrated SWAT ROS model to simulate the hydrological process using CMIP5 climate projections. All analyses are based on the assumption that this calibrated model is representative. However, as described "The SWAT ROS model for the Great Lakes Basin simulated historic streamflow at the daily time step with an NSE of 0.38 (with 29% of stations greater than 0.5) and a dr of 0.62 (Myers et al., 2021b). The model simulated historic snowpack SWE at the daily time step with an MAE of 26 mm", the model cannot be well considered well-calibrated with a low NSE of 0.38 for discharge simulation. Moreover, 26 mm MAE for daily SWE is a considerable high bias in comparison to the SWE value of the study area (e.g. Figure 4). The median SWE value of many months is around 50 mm or lower. GCM climate projections are highly uncertain already. A hydrological model with high bias will make the combination much worse. As a consequence, it is not reasonable to trust the analyses of this work about future climate change impact, even the analysis strategy is comprehensive. Therefore, the authors should implement the climate change investigation based on a reasonably well-calibrated SWAT ROS model. Moreover, detailed information about the rationality of the calibrated SWAT model is necessary but missing. Such information should be properly added to this paper or its supplementary material for its readers. The authors simply cited the paper that developed and evaluated the SWAT ROS model (reference below). But it is not open-access.

Myers, D. T., Ficklin, D. L., and Robeson, S. M.: Incorporating rain-on-snow into the SWAT model results in more accurate simulations of hydrologic extremes, J. Hydrol., 603, 126972, https://doi.org/10.1016/J.JHYDROL.2021.126972, 2021b.

> Response: Our multi-station calibration of streamflow and snowpack with a single global parameter set across the hydrologically diverse Great Lakes Basin enabled us to represent and compare streamflow and snowpack processes more comprehensively and objectively, as we describe further in the text below. This approach allowed us to verify that any spatial variation in ROS was actually due to variation in climate forcings, rather than artifacts of model parameters being overfitted to individual watersheds, and facilitated the communication of model performance in different hydrologic systems. Thus, as the evaluation statistics (e.g., NSE) are arbitrary when comparing different modeling approaches, we believe that the benefits of this approach outweigh the sacrifice in NSE in comparison with more regional or station-selective calibration options, which could have higher NSE values but be overfit to individual systems and produce more uncertainty

whether climate forcings or different model parameters would be causing the spatial ROS variation.

We attempted several new calibrations of the Great Lakes Basin SWAT ROS model aimed to improve model performance, including expanding parameter ranges within reason and removing the least sensitive parameters from the calibration. However, we were unable to improve model performance for simulating streamflow and snowpack beyond that of the Myers et al. 2021b model, which had been heavily experimented on with different calibration strategies and evaluations during that study. We believe that the accuracy of our model is actually good considering the uncertainties of all the spatially aggregated datasets that go into it (climate, snowpack, soils, etc.), and especially when considering the use of our daily time step, which incorporates high temporal scrutiny in our evaluations over a large geographic area. Previous work by Kalin et al. (2010) has stated that arbitrary interpretations of performance metrics for models at small temporal scales should be relaxed compared to what would be expected for models at coarse (e.g., monthly) time steps, for instance that an NSE between 0.3 and 0.5 could fit criteria for satisfactory model performance.

We also expanded our description of model evaluations to include more synopses from the Myers et al. 2021b study, and a figure which depicts the geographic dispersion of stations that calibrated well and those that did not calibrate as well. Figure 2 (below) now shows how stations that perform well for streamflow and snowpack simulation are dispersed throughout the Great Lakes Basin. We apologize that the Myers et al. 2021b paper is not open access. We are happy to share this paper, through the editor if you prefer. In addition, we provide more background for our choice of the Myers et al. 2021b model to represent spatial variability in ROS across the Great Lakes Basin using a single global parameter set and including evaluation stations that perform well along with those that do not perform as well. We provide the updated text below, to be added at the location of page 4, line 84 in the preprint.

"*The SWAT ROS model for the Great Lakes Basin simulated historic streamflow at the daily time step with an average NSE of 0.38 (with 29% of stations greater than 0.5, 48% of stations having NSE > 0.4, and a maximum NSE of 0.71) and a d_r of 0.62 (Myers et al., 2021b). The model simulated historic snowpack SWE at the daily time step with an MAE of 26 mm (Figure 2a-c). Calibrated parameters for this model can be found in Table S1 in Myers et al. (2021b). We also investigated seasonal model performance for only days when ROS melt was occurring, and found that the SWAT ROS model we use had an MAE of 8.6 mm, 9.4 mm, and 5.8 mm, respectively for simulating melt on those days in the winter, spring, and fall.*

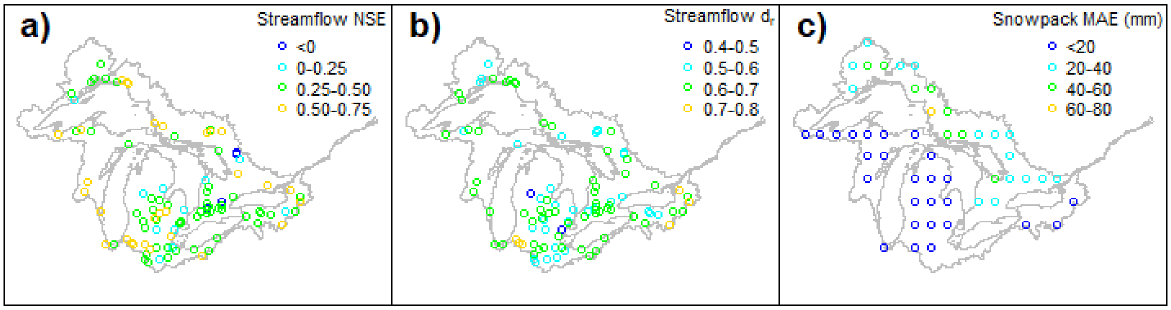

Figure 2. Evaluations for simulating historic streamflow Nash Sutcliffe Efficiency (NSE), streamflow revised Index of Agreement (d_r), and snowpack mean absolute error (MAE) at the daily time step from Myers et al., (2021b).

The multi-station evaluation we used with a global parameter set (from Myers et al., 2021b) enabled us to represent streamflow and snow processes more comprehensively and comparatively across the hydrologically diverse basin. Thus, any spatial differences in ROS would represent actual differences due to spatial variation in climate forcings, rather than artifacts of regionally calibrated parameter sets which could simulate processes differently. Alternatively, a regional calibration may have improved model performance for some stations, but at the expense of no longer being able to objectively evaluate spatial variation in ROS across the basin. We included stations that performed well (NSE > 0.5) along with stations that did not perform as well in our evaluations across the basin. Along with providing transparency about model performance in different hydrologic systems, this allowed us to verify that our model was performing as well as possible for simulating many watersheds and hydrological processes in a comparative way, and openly communicate instances where the model did not perform as well (Figure 2). Alternatively, we could have used fewer stations and evaluated the model to a more geographically limited system, which would have resulted in better performance for simulating those stations, but at the expense of misrepresenting the diversity of hydrological responses (e.g., snowpack) across the basin. We believe that the benefits of our approach, to be representative of variable hydrological responses and use a global parameter set, in providing confidence to our interpretations of spatial ROS variation outweigh the sacrifice of model performance for some stations."

For background from the Myers et al. 2021b publication, the SWAT ROS model was an improvement over the traditional SWAT model (not including ROS) which had an average daily streamflow NSE of -0.05 and average snowpack NSE of 48 mm). Also, the SWAT ROS model we used in this study was more accurate for simulating historic snowpack SWE at all 50 calibration stations than the traditional SWAT model. The traditional SWAT had larger snowmelt simulation errors of 10.4 mm, 10.2 mm, and 6.4 mm, respectively, for ROS days in winter, spring, and fall. In Myers et al. (2021b), the SWAT ROS model for the Great Lakes Basin was further evaluated in comparison with the traditional SWAT model for streamflow and snowpack simulation over 504 randomly generated parameter sets. The SWAT ROS model we used in this study improved performance (in comparison to the traditional SWAT) for simulating daily streamflow in

99.4% of the potential parameter sets, and for simulating daily snowpack in 100% of the sets. This demonstrated the nearly universal benefits for model performance of including ROS in simulations of Great Lakes Basin hydrology as we did.

In response to other reviewer comments, we are also adding a section 4.2, which discusses our historic ROS estimates in comparison with other studies, and is described in our public comment to Reviewer 1. In this section, we also objectively evaluate the historic ROS melt amount and ROS frequency estimates of our GCM ensemble against estimates from historic observations. This showed that our SWAT ROS model with GCM forcings was reasonably simulating ROS melt in the Great Lakes Basin in comparison with historic evidence.

Kalin, Latif, et al. "Predicting water quality in unmonitored watersheds using artificial neural networks." Journal of Environmental Quality 39.4 (2010): 1429-1440.

2. Future climate projects have large uncertainty. When evaluating climate change impacts, it is more reasonable to discuss the trend or relative changes rather than absolute quantities. The authors should shorten such contents and keep the necessary ones only. Besides, Figure 2 shows different behaviors of climate driving force during different future periods. It would be interesting to investigate the corresponding hydrological signatures of different future periods. Although, as described in section 2.4, the analyses of future period include mid-21st century and late-21st century. Throughout the paper, the result and analysis of late-21st century is almost none. Please complete such missing parts.

Response: We now primarily focus on relative changes in our analyses, while some absolute changes are also mentioned to provide context for the magnitude of the changes. An example to be added at the location of page 8, line 157 of the preprint is below.

"*For instance, the area-weighted median (value of a ranked set where half the total area is ranked lower; Willmott et al., 2007) amount of January ROS melt among subbasins increases 59% from 3.2 mm historically to 5.1 mm by mid-21st century (Figure 5a). Similarly, area-weighted median February ROS melt rises 50% from 7.0 mm historically to 10.5 mm in the mid-21st century. In the spring, the amount of ROS melt decreases due to the reduction in snowpack.*"

For consistency and clarity in our findings, we decided to remove the isolated references to late-21st century from the paper, and focus on mid-21st century changes only. We focus on mid-21st century because it is the period that the models (and RCPs) generally agree about climate changes, and because we expect that it will be the most meaningful focus for water resource managers in the Great Lakes Basin. Thus, we updated the preprint's Figure 2 (now Figure 3) to end after mid-21st century and clarified our text to alleviate any confusion about future time period.

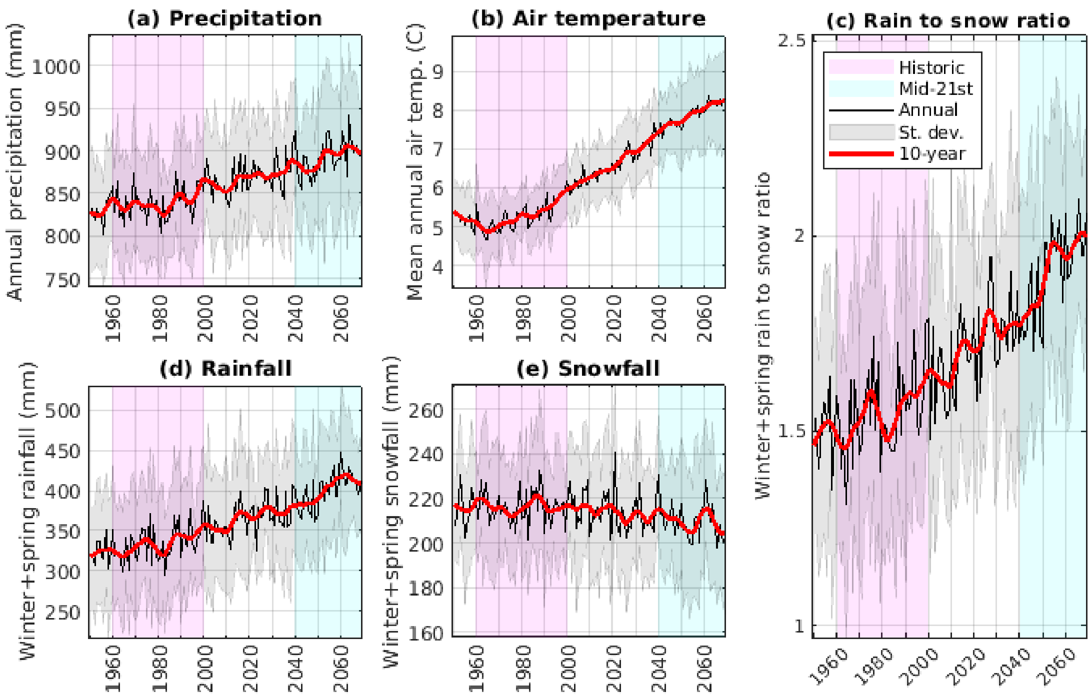

Figure 3. Basinwide ensemble-average a) annual total precipitation, b) annual air temperature, c) winter+spring rain to snow ratio, d) winter+spring rainfall, and e) winter+spring snowfall from the climate input data, with 10-year averages (red lines), based on the RCP 4.5 pathway. Shading indicates historic 1960-1999 (red) and mid-21st century 2040-2069 (blue) periods, as well as ensemble standard deviations (grey).