Reviewer 1

General comments:

This study is a novel investigation that is of interest to the professional community and in-line with the aims and scope of the journal. The topic is appropriately introduced with justification provided for the specific objectives. While some additional details on the statistical testing could be added (see below), the methodological approach appears logical and reproduceable. The results are organized around specific themes with figures that enhance understanding and are aligned with the final conclusions. Prior to supporting acceptance and publication, there are a small number of outstanding concerns with the manuscript that are addressed below as specific comments.

> Response: We thank the reviewer for the thorough and helpful review that has improved the quality of the manuscript.

Specific comments:

The proportions of historical ROS melt [to total melt] is larger here than a variety of previous findings for the region. For instance, Welty and Zeng (2021) find extreme ROS occurrence is approximately 24% for the Great Lakes basin, similar to the value the authors give on line 34 at over 25% of extreme ablation events being ROS. Looking at all ROS events, not just extreme, the maximum value to date I am aware of for this region is found in Suriano (2022). This notes between 30-50% of ablation is ROS in the eastern lakes, compared to less than 20% in the extreme northern/western regions. While the results here have a similar spatial pattern to Suriano (2022), with more ROS in the eastern lakes and less to the north and west, the magnitudes are rather different. Given one of the primary results of this study is the detection of large decreases in ROS events under the RCP4.5 scenario relative to historical period, it is warranted to provide further discussion on the robustness of the historical model values relative to observations. This appears absent from the manuscript currently and should be incorporated into the discussion section of the revision.

> Response: For our study, we used the definition of Jeong and Sushama (2018) to define an ROS event, as this definition was being used by them to project future climate impacts using RCP's across North America, and was based on the ROS definitions of studies before them. Thus, we defined an ROS event as >1 mm rainfall on >1 mm SWE and snowmelt occurring, so our results would be directly comparable with theirs. We now include an additional Discussion section (4.2) and figure that discusses the comparability of findings of historic ROS melt with other studies, and objectively evaluates our models against historic observed data, to be added at the location of page 19, line 351 of the preprint.
>
> "*Previous work by Jeong and Sushama (2018), whose definition of ROS we adopted, has found comparable estimates of historic frequencies of ROS events as we did,*

approximately 10-20 ROS days per year in the Great Lakes Basin. Also, Jeong and Sushama (2018) report an historic average annual amount of ROS runoff of approximately 100 mm or greater throughout the Basin, which is of a similar magnitude to our historic estimates. Jeong and Sushama evaluated their models using historic observations and found that spatial patterns in ROS were captured reasonably well, though errors could arise from uncertainties in the data driving their models rather than problems. Nonetheless, other studies have used different definitions of ROS events and/or have reported variable findings for the Great Lakes Basin. For instance, Welty and Zeng (2021) defined an ROS occurrence as air temperature >0 °C and precipitation >5 mm during 2-day extreme snowmelt events (e.g. >50 mm and the top 10 events over a 30 year historic period), which includes far fewer ROS events that we did. Additionally, Suriano (2022) defines an ROS event as a snow depth decrease of at least 1 cm with average daily temperature >0 °C, at least 0.01 cm precipitation, and no more than 2.54 cm snowfall (by depth) the previous day, over a 1960-2009 historic period. With this definition, Suriano (2022) reports a historic frequency of approximately 5 to 15 ROS events per year in the Great Lakes Basin.

To objectively verify the robustness of our historic estimates, we identified ROS amounts and frequencies in observed data using the same approach and definition as our GCM-forced SWAT model. The historic climate observations were from Maurer et al. (2007), used in Myers et al. (2021b), and our historic SWE observations were from Myers et al. (2021b), which had been estimated off the daily gridded North American snow depth dataset (Mote et al., 2018), both in a matching 1° latitude/longitude grid with 50 evaluation points over the Great Lakes Basin. We found that for historic annual estimates of ROS melt, the mean among the gridded evaluation points for our GCM ensemble was 120 mm, while the mean calculated from observations was 118 mm, which was not a significant difference ($p=0.90$). For individual evaluation points, the estimates of annual ROS melt were positively related with an MAE of 33 mm (Fig. S2a). This suggests that our GCM ensemble was reasonably estimating historic ROS melt amounts in the Basin. We also found that the historic observations estimated a mean average annual ROS frequency of 20 days across the evaluation points, which was greater than the mean of 12 days estimated by our GCM ensemble for the points over our historic 1960-1999 period ($p<0.001$). This was because our ROS definition included historic observed events that were the result of natural stochasticity in snowpack SWE amounts (i.e., sporadic daily increases or decreases in the SWE data, rather than "clean" modeled melt). Thus, our definition overestimated the frequency of ROS days when applied to historic observations, compared with our modeled ROS frequency, due to the additional stochastic small melt events identified by the criteria, with an MAE of 8 days (Fig. S2b). However, when ROS amounts are accumulated over the season, this issue is remedied (Fig. S2a)." *The smoothing produced by the observed data being aggregated to a 1 degree grid could also affect the comparisons between the historic observed and modeled estimates.*

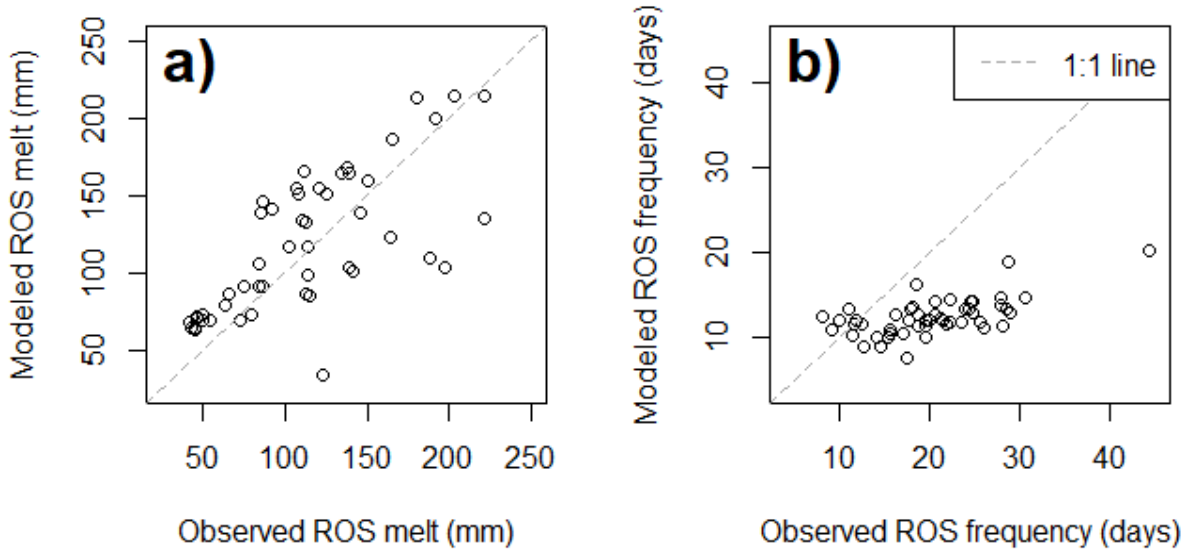

Figure S2. For the 50 gridded climate and snowpack evaluation points in the Great Lakes Basin, a) Comparison of historic (1960-1999) mean annual ROS melt amounts calculated for observed data with those modeled by our ensemble of climate projections, and b) The same comparison for the mean annual frequency of ROS events.

References:

Jeong, D. Il and Sushama, L.: Rain-on-snow events over North America based on two Canadian regional climate models, Clim. Dyn., 50, 303–316, https://doi.org/10.1007/s00382-017-3609-x, 2018

Suriano, Z. J.: North American rain-on-snow ablation climatology, Clim. Res., 87, 133–145, https://doi.org/10.3354/CR01687, 2022.

Welty, Josh, and Xubin Zeng. "Characteristics and causes of extreme snowmelt over the conterminous United States." Bulletin of the American Meteorological Society 102.8 (2021): E1526-E1542.

The authors acknowledge on line 126 the threshold used for statistical significance for their correlation tests. However, it is unclear if any significance testing was conducted for the rest of the study. Was any sort of t-test or difference of means testing conducted for the results comparing the historical period to the mid-century period? If not, this should be considered by the authors to aid in differentiating meaningful changes from ones still within the noise.

Response: We now have included significance testing for our comparisons of ROS and climate between the historic and mid-21st century periods. We also now describe the

approach for this in the methods, and provide instances where significance testing is used below. However, throughout our revision we keep effect size as the focus, rather than statistical significance, following the guidance of previous work (Wasserstein et al., 2019; Ziliak & McCloskey, 2008)

"*For comparisons between time periods, significance was tested by comparing annual area-weighted ensemble-average values for the Great Lakes Basin between the historic (1960-1999, n=40 years) and mid-21st century (2040-2069, n=30 years) periods using two-tailed unpaired t-tests.*" (to be added at page 6, line 126 of the preprint)

"*Spatially averaged annual precipitation increases 6.3% from 839±63 mm (mean and standard deviation of GCM ensemble) during 1960-1999 to 892±77 mm by the mid-21st century (p<0.001), while spatially averaged annual air temperatures increase from 5.2±0.7 °C during the 1960-1999 period to 7.9±1.0 °C (a 2.7 °C increase, p<0.001).*" (to be updated at page 6, line 132 of the preprint)

"*Further, our model shows that winter+spring rain to snow ratios over the basin (calculated by dividing the total winter+spring rainfall by total winter+spring snowfall) increase from around 1.5 historically to 1.9 by mid-century (p<0.001), which means that proportionally more rainfall could contribute to the declines in snowmelt and snowpack SWE*." (to be updated at page 6, line 141 of the preprint)

"*Overall, the ensemble average amount of annual snowmelt during ROS events, at the major river basin scale using RCP 4.5 models, changes by -42% to +1%, with a basinwide area-weighted average of -22% (p<0.001).*" (to be updated at page 12, line 197 of the preprint)

References:

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". The American Statistician, 73(sup1), 1-19.

Ziliak, S., & McCloskey, D. N. (2008). The cult of statistical significance: How the standard error costs us jobs, justice, and lives. University of Michigan Press.