Dear authors,

thank you for sumbitting your revised manuscript and responses to the reviewer reports. While all comments by referee #1 have been cleared, the report of referee #2 indicates that there is still need for clarification, especially regarding important aspects of model performance and scope. I invite you to respond to these comments and provide a manuscript with corresponding revisions.

Kind regards,
Daniel Viviroli

> Response: We thank the editor for all the help with the review and suggestions for improving the manuscript. Below are our responses to the reviewer comments.

Reviewer 1

> Response: We thank the reviewer for the helpful comments on the manuscript and the positive feedback.

Reviewer 2

The manuscript has been improved according to the comments in round 1. Especially, the climate projection sections read more reasonable now. But (there is a but), the model rationationty part, as indicated in my first comment of last round, is still not convincing. This is the basis for whatever outcomes of the future projection investigations in this work. Because ROS is such an interesting research topic and this work could be beneficial for the community, I dig into the methodology paper (Myers et al., 2021b) that this manuscript is based on. To make it clear, let me zoom into the specific points:

> Response: We thank the reviewer for providing helpful comments to improve those sections, and for providing constructive feedback for what still needs improvement. We hope our responses can satisfactorily address the remaining concerns.

1. Low average NSE for discharge simulation. “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% greater than 0.4, and a maximum NSE of 0.71)”. It is not acceptable, in particular, the author wants to address the effects on extreme high water yield (as defined by the authors: “Finally, the SWAT model outputs for water yield represent the area-averaged water export through the outlet in mm.”). with such a definition, streamflow is more or less equivalent to or contributes a lot to the water yield of this work. Mathematically, NSE is highly affected by peak flow which is highly relevant to extreme high water yield. Low NSE

largely indicates a bad fitting of the peaks. Thus, it is not reasonable to discuss the implications on extreme water yield, if the model cannot represent them properly.

Response: Following the reviewer's feedback and further research, we removed the analysis regarding extreme winter and spring water yields from the results, as well as its mention in the methods and conclusions sections. We now introduce it as a potential avenue for further research. We also elaborated on how the model performance guides the scope of investigations to only those most reliable and representative.

"*Stations with streamflow NSE > 0.50 at the daily time step were spread throughout the Great Lakes Basin, suggesting that the model was representative of the spatial tendencies in climate forcings and hydrological responses (Figure 2a). Stations that performed well with d_r > 0.60 were also distributed across the Basin, which is important because d_r is not as influenced by extremes as NSE (Willmott et al., 2012) and can be a more interpretable indicator of overall model performance (Willmott et al., 2015). Further, the presence of stations not performing as well by NSE could be at least partly explained by the diversity of spatially-distributed hydrological behaviors of the Basin having been simplified in the model's multi-site and multi-objective calibration (Zhang et al., 2008), as well as uncertainties in gridded climate forcings (Maurer et al., 2010; Muche et al., 2020; Stern et al., 2022) and snowpack calibration data (Mote et al., 2018; Hill et al., 2019). Thus, we chose to focus on average amounts of ROS melt over different time periods for this study, without focusing on specific events such as extreme (e.g., 0.95 quantile) water yields. Those extremes may not be represented as reliably in future climate projections given that many of the stations had lower NSE values than we deemed adequate for that purpose, largely because NSE values are sensitive to extremes (Legates and McCabe, 1999; Willmott et al., 2009). Further, projections of climate change impacts on hydrologic extremes are best analyzed using models that focus on extreme flows specifically (Willems et al., 2014).*" (page 4, lines 88-100 in the revision with changes tracked)

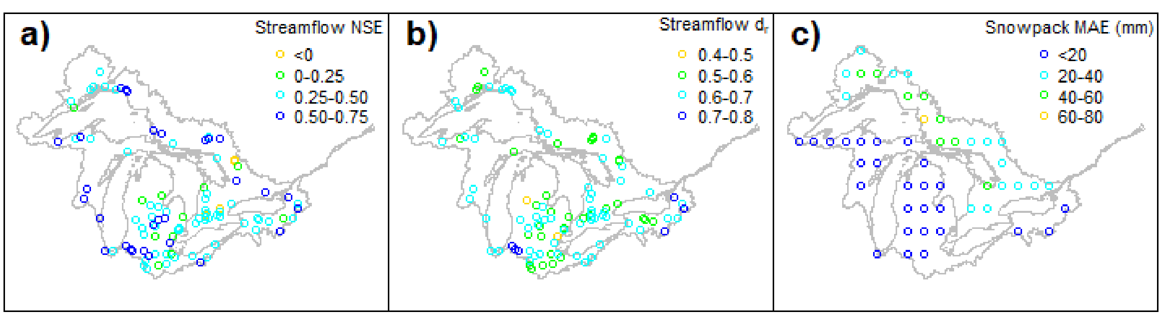

Figure 2. Evaluation statistics for simulating historic: a) streamflow using Nash Sutcliffe Efficiency (NSE), b) streamflow using revised Index of Agreement (d_r), and c) snowpack using mean absolute error (MAE) at the daily time step. Adapted from Myers et al., (2021b).

"*Investigations projecting the response of extreme water yields to changing ROS conditions in future climates are an additional avenue for future research with meaningful implications for water resources management.*" (page 22, lines 43-45)

References:

Hill, D. F., Burakowski, E. A., Crumley, R. L., Keon, J., Michelle Hu, J., Arendt, A. A., Wikstrom Jones, K., and Wolken, G. J.: Converting snow depth to snow water equivalent using climatological variables, 13, 1767–1784, https://doi.org/10.5194/tc-13-1767-2019, 2019.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Maurer, E. P., Hidalgo, H. G., Das, T., Dettinger, M. D., & Cayan, D. R. (2010). The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. Hydrology and Earth System Sciences, 14(6), 1125-1138.

Mote, T. L., Estilow, T. W., Henderson, G. R., Leathers, D. J., Robinson, D. A., and Suriano, Z. J.: Daily gridded north American snow, temperature, and precipitation, 1959-2009, version 1, Boulder, Color. USA NSIDC Natl. Snow Ice Data Center. N5028PQ3. https//nsidc.org/data/G10021/versions/1 (accessed 30 June, 2021), 2018.

Muche, M. E., Sinnathamby, S., Parmar, R., Knightes, C. D., Johnston, J. M., Wolfe, K., ... & Smith, D. (2020). Comparison and evaluation of gridded precipitation datasets in a Kansas agricultural watershed using SWAT. JAWRA Journal of the American Water Resources Association, 56(3), 486-506.

Stern, M. A., Flint, L. E., Flint, A. L., Boynton, R. M., Stewart, J. A. E., Wright, J. W., & Thorne, J. H. (2022). Selecting the optimal fine-scale historical climate data for assessing current and future hydrological conditions. Journal of Hydrometeorology, 23(3), 293-308.

Willems, P., Mora, D., Vansteenkiste, T., Taye, M. T., and Van Steenbergen, N.: Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes–Part 2: Intercomparison of models and calibration approaches, J. Hydrol., 510, 591–609, 2014.

Willmott, C. J., Matsuura, K., and Robeson, S. M.: Ambiguities inherent in sums-of-squares-based error statistics, Atmos. Environ., 43, 749–752, https://doi.org/10.1016/J.ATMOSENV.2008.10.005, 2009.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: Short Communication A refined index of model performance, Int. J. Climatol., 33, 1053–1056, https://doi.org/10.1002/joc.2419, 2012.

Willmott, C. J., Robeson, S. M., Matsuura, K., and Ficklin, D. L.: Assessment of three dimensionless measures of model performance, Environ. Model. Softw., 73, 167–174, https://doi.org/10.1016/J.ENVSOFT.2015.08.012, 2015.

Zhang, X., Srinivasan, R., and Van Liew, M.: Multi-site calibration of the SWAT model for hydrologic modeling, Trans. ASABE, 51, 2039–2049, 2008.

Thoughts: After checking the methodology paper (Myers et al., 2021b), it seems that the model simulated somehow descent streamflow in winter/spring period but performed bad in summer period. If ROS is more important for winter/spring period (I am not a snow scientist. This statement should be double checked), perhaps calculating the metrics and proving the reasonable representation of streamflow in these seasons is a breakthrough point.

> Response: We appreciate the reviewer for the constructive and helpful idea. We explored the impact of using only winter and spring data in our evaluations. However, the evaluation statistics were similar to before, with a streamflow NSE of 0.35, streamflow d_r of 0.64, and snowpack MAE of 37 mm. Thus, we decided to stick with the evaluations as reported in the Myers et al., 2021b study.

Minor suggestion: higher temporal resolution does not necessarily mean using a lower criterion for evaluating the model performance. And the spatial side should not be ignored. In this work, the spatial resolution is coarse. I suggest to remove the statement “for instance that an NSE between 0.3 and 0.5 could fit criteria for satisfactory model performance in some contexts”.

> Response: We removed the statement as suggested.

2. High MAE for daily SWE. The authors added some descriptions of snow melt performance as supporting references of reasonable snow simulations. However, acceptable MAE snowmelt simulations do not mean reasonable performance of SWE outputs. It may indicate the melt module works, and something has to be improved in the snow accumulation module. The authors should clarify this point: is 26 mm MAE for daily SWE is an acceptable bias in the study region or not? Particularly, as shown in Figure 5, the median SWE value of many winter months is around 50 mm or lower. As a reader, I will interpreter 26 mm MAE as a large error.

Response: We now add more clarification about our snowpack SWE errors and how they were distributed across the Basin, with larger absolute errors in stations that had larger historic snowpack amounts, and errors less than the mean of 26 mm MAE in stations with smaller historic snowpack amounts.

"*Daily snowpack SWE error across the Basin ranged from <20 mm MAE throughout the southwest subbasins to approximately 40-70 mm MAE in the northeast (Figure 2c). This spatial variation in MAE scaled with the average observed daily snowpack SWE during winter and spring across the Basin, which ranged from approximately 50-100 mm in the southwest subbasins to over 150 mm in the northeast, described in Section 3.4. Thus, subbasins with lower amounts of observed SWE would also have smaller errors than the average of 26 mm. We find this measure of absolute error to be acceptable, particularly considering the errors inherent to the gridded snowpack data we compare against (e.g., spatial averaging and simplification of accumulation and ablation processes during conversion from snow depth; Myers et al., 2021b; Ensor and Robeson, 2008; Hill et al., 2019). Mean absolute errors in modeling the gridded snowpack SWE have been found to vary temporally as well, for instance being 12.7 mm MAE on January 1, 45.1 mm MAE on February 1, 26.8 mm MAE on March 1, and 9.6 mm MAE for April 1, 1978, across the Great Lakes Basin in comparison with station measurements, and proportional to the amount of snowpack on the ground during those days (Myers et al., 2021b).*" (page 4, line 101 to page 5, line 112)

References:

Ensor, L. A. and Robeson, S. M.: Statistical characteristics of daily precipitation: Comparisons of gridded and point datasets, J. Appl. Meteorol. Climatol., 47, 2468–2476, https://doi.org/10.1175/2008JAMC1757.1, 2008.

Environment Canada and USEPA: The Great Lakes: An Environmental Atlas and Resource Book, https://doi.org/10.1038/184665a0, 1995.

Hill, D. F., Burakowski, E. A., Crumley, R. L., Keon, J., Michelle Hu, J., Arendt, A. A., Wikstrom Jones, K., and Wolken, G. J.: Converting snow depth to snow water equivalent using climatological variables, 13, 1767–1784, https://doi.org/10.5194/tc-13-1767-2019, 2019.

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.

3. The authors added some information about the modified ROS-SWAT in this manuscript. "simulates ROS melt based on a function of air temperature, precipitation, wind, saturated vapor pressure, and atmospheric pressure". In your current descriptions, it means wind, saturated vapor pressure, and atmospheric pressure datasets are required. It could indicate your method cannot

work in many snow process dominated remote areas due to data scarcity. But I found the answers in your supplementary source code snom.f, those variables are simulated internally based on elevation (atmospheric pressure), air temperature (saturated vapor pressure) and fixed value of 0.15 (for wind variable UADJ). To avoid misleading, please clarify briefly what exactly is needed as inputs for your module in the manuscript.

> Response: We added the following sentence to clarify that no additional data inputs are required for the SWAT ROS model.
>
> "*For the SWAT ROS model, air temperature and precipitation are based on existing SWAT model inputs, wind effects on ROS are simulated using an average function for ROS melt from turbulent energy transfer, saturation vapor pressure is based on air temperature, and atmospheric pressure is based on elevation, so no additional data inputs are required.*" (page 3, lines 70-73)

Many thanks to the editors for sharing me the methodology literature: 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.