This work provides insights and framework on how to quantify input data quantity, quality, and their impacts on ML predictions for water quality and quantity, with potential to guide future modeling efforts. Here are some main suggestions to improve the manuscript: (1) Expand introduction to include recent work on data synergy and the integration of process-based models and ML. (2). The water quality model seems to exclude nutrient inputs, which weakens the conclusions. (3). Consider simplifying some of the results figures. Not all models, basins, or test cases need to be visualized explicitly at the same time. (4) Consider exploring how the findings can be applied to improve better modeling processes in Discussion.

Detailed comments:

1. Introduction: There have been emerging studies on the "data synergy" effect of data-driven approaches and combining process-based models with ML. However, this manuscript lacks a comprehensive and evaluative literature review. Please consider including some up-to-date work, such as:

- Kratzert, Frederik, Daniel Klotz, Sepp Hochreiter, and Grey S. Nearing. 2021. "A Note on Leveraging Synergy in Multiple Meteorological Data Sets with Deep Learning for Rainfall–runoff Modeling." Hydrology and Earth System Sciences 25 (5): 2685–2703.
- Razavi, Saman, David M. Hannah, Amin Elshorbagy, Sujay Kumar, Lucy Marshall, Dimitri P. Solomatine, Amin Dezfuli, Mojtaba Sadegh, and James Famiglietti. 2022. "Coevolution of Machine Learning and Process-based Modelling to Revolutionize Earth and Environmental Sciences: A Perspective." Hydrological Processes 36 (6). https://doi.org/10.1002/hyp.14596.
- Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. "Deep Learning and Process Understanding for Data-Driven Earth System Science." Nature 566 (7743): 195–204.

2. Line 80: Can you discuss the rationale for choosing RF, SVM, and ANN as the machine learning models for streamflow predictions to test your hypothesis? Recent studies have suggested that Long Short-Term Memory (LSTM) networks are the state-of-the-art machine learning models for time-series river flow predictions, outperforming other approaches.

3. Table 1: In the current model setup, for both water quantity and water quality predictions, the baseline inputs (WDO) include only climate variables. However, in reality, nutrient inputs (e.g., fertilizers, human waste, and manure) are essential for predicting N and P, regardless of whether using machine learning or process-based models. Do you think it is fair to test TN and TP predictions when WDO includes only climate variables? If nutrient inputs were incorporated, adding loads from SWAT might have less impact. How

much of the conclusions from this study can be applied to water quality ML studies where nutrient inputs are typically included as predictors?

4. Line 255: More information about the SWAT model is needed.

- The manuscript states that SWAT can incorporate management practices, and two of the watersheds are heavily farmed and urbanized. Were inputs and parameters representing agricultural and human processes that significantly impact water quality (TN, TP, and SSD) included and calibrated? Please provide the inputs, parameters, and calibrated values.
- It is also unclear how many SWAT models were developed. Was a single SWAT model used per basin, or was a separate model created for each target variable within a basin?
- Related to (2), can you clarify the calibration process? Was a weighted multi-objective calibration approach applied, or were parameters calibrated for flow first, followed by calibration for water quality?

5. Figure 4: 3D plots are often hard to interpret. Consider using 2D plots for model performance comparison.

6. Figure 5: Could you adjust the figure size to make the duration curves less flat? It's difficult to distinguish the differences. What is the scale of the Y-axis? It appears to be on a log scale

7. Figure 6: All basins and target variables show a similar pattern, with the key takeaway being that ME increases with the amount of data, regardless of data correlation. You might consider leaving one subfigure and moving the rest to the supplemental section, or consolidating them into a single figure. For example, as the conceptual figure suggests, the X-axis could represent different data, the Y-axis ME, and different lines could indicate target variables. Adding uncertainty bands could also capture watershed variance. Just some suggestions to consider

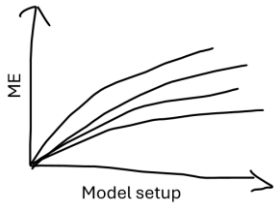

Figure 7: This figure now has too many dimensions—different basins, target variables, model inputs, and ML models—which makes it quite confusing. Could you simplify it to highlight key finds, perhaps by leveraging the suggestions for Figure 6

8. Figure 10: How well are the regressions? Can you report R2 and p-value?

9. Discussion: Several points discussed are common knowledge, making them less novel and somewhat irrelevant. For example: (1) ML model accuracy depends on study watersheds, target variables, and ML model types; and (2) water quality is generally harder to predict than water quantity. The key findings of this work is its quantitative evaluation of data quality and quantity and their relationship to model performance. Could you expand on how these findings can guide future modeling efforts, such as optimizing input selection, implementing quality control measures, or integrating insights with process-based models

10. Maybe I missed it. Did you talk about uncertainty and limitations of your work?