This manuscript presents a comprehensive large-sample study evaluating the impact of LSTM-based data integration (DI-LSTM) on streamflow simulation across hundreds of basins in the Western U.S., using both streamflow (Q) and snow water equivalent (SWE) as auxiliary inputs. The study is motivated by the operational challenges of hydrological forecasting in arid and snow-dominated regions and aims to improve short- and long-term forecasting using deep learning techniques. The authors highlight the advantages of DI over traditional data assimilation (DA) and provide an extensive experimental comparison across multiple timescales and input configurations. My detailed comments are as follows:

Major Comments

1. The manuscript title references "implications for forecasting in the Western U.S.," yet the experimental setup focuses solely on hindcasting using future observations (i.e., perfect knowledge of lagged Q or SWE). It would be better if the authors could clarify what specific implications for real-world forecasting are supported by their results, and how the proposed DI-LSTM might be adapted for settings where future information is unavailable or uncertain.

2. There is a risk that DI-LSTM overfits to future data, especially when lagged target variables (Q or SWE) are incorporated directly from observed time series. It would be better if the authors could clarify:
    - Whether the lagged variables are drawn from observations or predicted recursively;
    - How these variables are embedded into the model;
    - And whether any form of future leakage occurs during training or evaluation.
    - It would also be helpful if the authors could provide a clear schematic of the DI-LSTM architecture to illustrate how lagged information is integrated into the model.

Minor Comments

1. Line 138: The typographic dash in DI-LSTM in the formula appears to be a mathematical minus sign. Please correct this to ensure clarity.

2. The choice of using a 10-day lag for Q and a 6-month lag for SWE is not clearly justified. It would be better if the authors could explain the rationale behind these specific durations, either based on hydrological reasoning or exploratory experiments.

3. It would be better if the authors could discuss more thoroughly the phenomenon shown in Figure 10(a), particularly the performance degradation at 4–7 day lags in some snow-dominated basins.

4. Sensitivity to Random Initialization and Training Variability. It would be better if the authors could report how diverse the six randomly seeded training runs are. This would help clarify whether the models are sensitive to random initialization or the stochastic training process. Reporting variability across seeds would improve the robustness and reproducibility of the findings.

5. While Table C2 provides hyperparameters for model training, it would be better if the authors could briefly justify their selection or indicate whether any tuning or sensitivity analysis was performed. This would help assess the robustness of the model configuration and whether the selected architecture is optimal across diverse basin types.