**Response to Journal Reviewer Comments**

**Manuscript:**Characterizing the influence of remotely sensed wetland and lake water storage on discharge using LSTM models

**Authors:** Melanie K. Vanderhoof, William Keenan, Wayana Dolan, Heather E. Golden, Charles R. Lane, Jay R. Christensen, Kylen Solvik, Adnan Rajib

**Editor:**

**Comment:** Apart of the reviewers comments I noticed that the literature review is missing entirely previous recent studies from the Hydrological Sciences Journal that could highlight the relevance of the article with the journal scope.

**Response:** Hydrological Sciences Journal publishes lots of great and relevant research focused on using deep learning and LSTMs for hydrological and discharge-related applications. We have added references to three LSTM-focused manuscripts published in Hydrological Sciences Journal.

**Reviewer: 1**

Recommendation: Subject to major revision. If revised paper is resubmitted, it needs to be reconsidered and re-reviewed

**Comments:**

**Summary Comment:** The manuscript addresses an important topic at the intersection of hydrologic modeling and remote sensing by exploring how surface water storage data can enhance streamflow prediction using LSTM models. While the study presents interesting findings, several aspects related to methodology, interpretation, and clarity would benefit from further elaboration or refinement. The following comments can be considered to enhance the clarity and impact of the work:

**Response:** We appreciate your thoughtful comments which have improved the manuscript. In response we have (1) clarified the novelty of the analysis. No hydrological modeling or LSTM analyses, that we are aware of, has previously ingested and trained a multi-watershed model on landscape scale remotely sensed time series of SWstorage. We also added, (2) a paragraph discussing reservoir LSTMs to the Introduction, (3) two approaches to characterize LSTM model uncertainty and two corresponding new figures to present these results in the Appendix, (3) a correlation analysis between time series of surface water extent and storage in each watershed, and (4) a new analysis quantifying the relative contribution of each variable group to the NSE, and a corresponding new figure in the Results section, showing the relative contribution of surface water storage to model NSE. Detailed responses to all of your comments are provided below.

1) While the final sentence of abstract and conclusions suggest implications for wetland restoration, it is not clearly articulated in the manuscript "how improved discharge modeling translates into actionable insights for restoration planning or prioritization?".

**Response:** We erased this sentence from the abstract and instead now better clarify the novelty of the analysis. In the Introduction we also added references to text on restoration and prioritization.

2) In the Introduction, the phrase “we developed multiple Long Short-Term Memory (LSTM) models...” may be somewhat misleading. Since the LSTM architecture is a pre-existing model, a more precise verb such as “trained,” “configured,” or “simulated” might be preferable.

**Response:** We have revised this term from “developed” to “trained” throughout the manuscript.

3) It would strengthen the literature review to acknowledge previous studies that have demonstrated the benefits of incorporating surface water storage, such as reservoirs, into LSTM-based streamflow prediction models (<https://doi.org/10.1016/j.jhydrol.2021.126455>).

**Response:** We have added a paragraph to the Introduction, including the recommended reference, to discuss LSTM efforts that incorporated data from reservoirs. As part of that added text, we now note that LSTMs have been used to improve estimates of reservoir outflows, and several studies have shown that incorporating time series of reservoir storage improves these estimates. However, to our knowledge no prior LSTM studies have incorporated landscape-scale time series of surface water. This gap is particularly important for non-stream connected surface water which may show very different temporal lags relative to a reservoir (Lane *et al*. 2018).

4) It would be helpful to clarify whether a single LSTM model was trained across all watersheds jointly (i.e., a regional model), or if individual models were trained separately for each watershed.

**Response:** We now clarify at the start of section 2.4.1 that each LSTM was trained across all watersheds jointly.

5) The manuscript would benefit from a more explicit discussion of model uncertainty, particularly given the stochastic nature of LSTM training (e.g., weight initialization, dropout, and training dynamics). It is not currently clear whether uncertainty was quantified or accounted for (e.g., through multiple model runs, ensembles, or confidence intervals around performance metrics). This is especially important in light of Figure 4, where the performance of all four LSTM configurations appears quite similar across many watersheds. The reported differences in performance may fall within the natural variability of LSTM outputs, making it difficult to confidently attribute observed improvements to the inclusion of surface water storage data (SWstorage). Without addressing this, the specific contribution of SWstorage remains somewhat ambiguous.

**Response:** We acknowledge that characterizing model uncertainty is an important component of building confidence in model results. We now characterize model uncertainty using two different approaches. To help quantify LSTM uncertainty related to within model development, we evaluated changes in median NSE with increasing epochs, or how each model performance changed with each iteration of training samples used to update model parameters, across all watersheds and by region. In addition, because of the stochastic nature of deep learning models, each trained LSTM model will produce slightly different predictions of discharge each time the model is trained. To characterize model uncertainty attributable to model generation, or uncertainty in prediction interval, we generated an ensemble of model runs. Uncertainty in model-based prediction for each LSTM model was characterized using the R-factor, which reflects the average distance of uncertainty, or the average difference between the 5th percentile prediction and the 95th percentile prediction at each timestep, divided by the standard deviation of the observed discharge data. In the Results section we have added text to explain the uncertainty results. We modified Figure 4 to include a between model comparison of changes in median NSE with model epoch. We also added two new figures to the Appendix that show the changes in median NSE with model epochs by ecoregion, where we show that our major findings are stable and do not depend on the selected epoch. An R-factor of <1 is generally recognized as strong. Our median R-factors for each of the four LSTM models ranged from 0.07 to 0.16 supporting confidence in our results.

6) The authors convert surface water extent into surface water storage using a DEM-based approach, and appropriately acknowledge that this introduces additional uncertainty. However, it is not clear whether using surface water extent directly as an input variable was considered or tested. Given that surface water extent often captures hydrologically relevant dynamics and may offer similar predictive value, the authors are encouraged to experimentally compare the performance of models using extent versus storage. Such an analysis would clarify whether the volume conversion provides meaningful benefits or if comparable improvements in discharge prediction can be achieved with less processing and uncertainty. This comparison would also help isolate the true contribution of the surface water signal itself.

**Response:** Because surface water extent is more widely available, relative to surface water storage, we understand the interest in exploring the tradeoffs of these two related variables. However, comparing surface water extent to surface storage time series was not a primary objective of our analysis. Instead, it was important for us to include surface water as a volume, not extent, in order to more accurately represent “real-world” changes in the surface water balance and integrate these changes with changes in discharge volume. In responding to the comment, we were concerned that adding additional LSTMs trained on surface water extent would confuse and distract from the current analysis. Therefore, to evaluate the influence of the storage conversion on the time series dynamics, the gap-filled surface water extent time series were correlated with the surface water storage time series, per watershed, at a 2-week timestep (2016-2023) using Pearson correlation. These Methods have been added to the Appendix text. We now also briefly discuss the results of this in the Discussion section. In correlating surface water extent to storage time series, we found a median±standard deviation Pearson correlation of *R*=0.88±0.12 with correlation values ranging from *R*=0.49 to *R*=0.98 across watersheds, suggesting that while generally extent was highly correlated with storage, the conversion induced greater changes in time series dynamics in some watersheds.

7) Given the inherent long memory capabilities of LSTM models, it's possible that they may implicitly learn and represent the influence of storage dynamics from meteorological inputs and temporal patterns, even without explicitly including SWstorage. This raises a valuable point for discussion: to what extent is the added value of SWstorage due to new information versus redundancy with what the LSTM might infer on its own? The authors are encouraged to discuss this possibility in the Discussion section, and, if feasible, support the hypothesis through interpretable machine learning techniques, such as SHAP (SHapley Additive exPlanations), Integrated Gradients, or feature importance analyses. Doing so would help clarify the specific contribution of SWstorage and enhance the manuscript’s relevance by offering insights into how LSTM models utilize different inputs. Interpretability efforts like this would go beyond performance metrics and contribute to advancing our understanding of hydrologic modeling with deep learning.

**Response:** We recognize that because the focus of the manuscript is on identifying the relative contribution of surface water storage variables to LSTM model performance, that providing a clearer understanding of relative variable importance would be useful. Our concern with using methods like SHAP and Integrated Gradients is that their values are often biased, particularly when features are correlated, making it difficult to accurately interpret. In addition, while we have used the SHAP package in previous publications, the format in which the NeuralHydrology version of an LSTM is outputted is not easily compatible with pytorch variable importance packages. Instead, we developed an approach to use the per watershed comparisons of model NSE to quantify the contribution (0 to 100%) of specific variable groups (MET, SW, SW after CC, CC, CC after SW). This approach allowed us to stay consist with our approach of evaluating variable groups, instead of individual variables, but also provided a continuous importance measure. We have added the variable importance approach and equations used to the Methods section. To the Results section we have added a new Table 4 to present the results by region. We also eliminated the former groupings (e.g., SW a consistent contributor, SW improves upon CC), and replaced with a new 3-panel figure showing (a) the contribution of meteorological (MET) forcings to the NSE relative to the best performing model for each individual watershed, (b) the contribution of surface water (SW) to the NSE in the MET+SW model, and (c) the contribution of SW to NSE within the ALL model.

Additional Questions:

What is the main contribution of this manuscript?: Neither an original contribution to hydrological theory or methodology, nor a valuable contribution to factual information about the hydrology of a region.

**Response:** The main contribution of this manuscript is to improve our understanding of where and when accounting for surface water storage dynamics (i.e., wetlands and lakes) is important for accurate predictions of discharge. Most wetlands globally have been lost as these features have been and continue to be converted to agriculture or development. Understanding the role of surface water bodies on watershed hydrology is critical so that we can start to understand the potential unintended impact these wetland losses may have on river discharge. We have revised the Introduction to clarify the manuscripts novelty. We now state that: “But to our knowledge no prior LSTM studies have incorporated landscape-scale time series of surface water. This gap is particularly important for non-floodplain surface water which show very different temporal lags relative to a reservoir (Lane et al. 2018).” And in the last Introduction paragraph we now clarify that “Process-based hydrologic modeling efforts to date on this topic have been limited to single, small, wetland-dense watersheds, and no hydrological modeling or LSTM analyses, that we are aware of, have ingested and trained a multi-watershed model on landscape scale remotely sensed time series of SWstorage.”

**Associate Editor:** Dr Ankit  Agarwal

**Summary Comment:** This study explores the impact of incorporating remotely sensed surface water storage (SWstorage) into LSTM models for predicting daily discharge across 72 U.S. watersheds. Using Sentinel-1 and Sentinel-2 data, four model configurations are tested, revealing that adding SWstorage improves model performance in over 80% of catchments, especially in wetland-rich regions like the Prairie Potholes. The enhancement is most notable during wetting and low-flow periods. Results highlight the value of dynamic surface water data in hydrologic modeling, offering improved accuracy for water resource forecasting and management, particularly in regions dominated by non-floodplain wetlands with strong seasonal storage dynamics.

Indeed, the manuscript has potential, following the listed major comments.

**Response:** We appreciate your thoughtful comments and have addressed each one in detail below.

**Comment:** The abstract effectively presents the motivation but lacks a clear articulation of why using LSTM for SWstorage is a novel approach (Page 4, Lines 10–12). Suggest clarifying how this differs from traditional modeling methods or prior LSTM use (e.g., including a line on what has not been done yet with LSTMs and SWstorage).

**Response:** We have revised both the Introduction and Abstract to better communicate and clarify the novelty of our analysis. The Abstract now clarifies that, “This effort represents the first exploration to ingest and train a multi-watershed LSTM on landscape scale remotely sensed time series of surface water storage.”

Page 5, Lines 9–17: The manuscript mentions limitations of prior studies but does not convincingly link them to the specific research gap this paper addresses. Recommend sharpening the novelty by explicitly stating, for example: "No prior LSTM-based study has incorporated time-varying remote sensing-based SWstorage across diverse hydrologic regimes..."

**Response:** We have revised the Introduction to clarify and sharpen the novelty, we now state that: “But to our knowledge no prior LSTM studies have incorporated landscape-scale time series of surface water. This gap is particularly important for non-stream connected surface water which may show very different temporal lags relative to a reservoir (Lane et al. 2018).”

And in the last Introduction paragraph we now clarify that “Process-based hydrologic modeling efforts to date on this topic have been limited to single, small, wetland-dense watersheds, and no hydrological modeling or LSTM analyses, that we are aware of, have ingested and trained a multi-watershed model on landscape scale time series of SWstorage.”

Page 20, Lines 5–25: The explanation of LSTM model logic is unnecessarily mathematical for the intended readership. Suggest moving some of the equations to supplementary material and focusing instead on the implications of the architecture for hydrologic memory and discharge lag.

**Response:** We moved the model equations and mathematical explanation from the Methods section to Appendix A.

Page 21, Lines 12–28: While the choice of NSE as the primary metric is justified, the exclusion of RMSE or other error metrics for residual characterization should be discussed. Consider adding a brief rationale for not using RMSE or including it as a secondary metric.

**Response:** As rates of discharge were so seasonally variable, and the RMSE calculates, essentially, the average difference (m3/sec) between predicted and observed discharge, we found that the RMSE was disproportionately representing average errors during high discharge days. However, we recognize that using multiple error metrics, particularly to characterize residuals, is important. In Table 3 we present the mean percent error (MPE) by region for all flows and Q2, and average residual values for Q70. We found the mean percent error to better represent the full range of discharge conditions and also to be more easily interpretable. In former Table 4, now Table 5, we also present the absolute percent bias (PBIAS) by region for wetting, high flow, drying, low flow and annual periods. To directly address this comment we also added a sentence to the Methods justifying our choice to use metrics other than RMSE.

Sections 4.3 & 4.4: While uncertainty is addressed, the discussion underplays the spatial bias in the training data. Notably, under-sampling in mountainous and forested regions (Page 10, Lines 2–5) likely impacts generalizability. Recommend stronger qualification of claims made in conclusions (Page 39, Lines 5–10).

**Response:** We revised the Conclusion section to more clearly articulate the geographic limitations of our findings.

Page 38, Line 25: “...adequate, at least for watersheds where M3 NSE exceeded M2 NSE.” This needs elaboration on whether proxies truly replace dynamic SWstorage or only serve as correlates under limited conditions.

**Response:** We have revised this paragraph to clarify that, “However, while the inclusion of catchment characteristics can help an LSTM model watershed-based differences in discharge, dynamic datasets provide important data on temporal variability in watershed condition.”

Figure 2: The flowchart is visually dense. Recommend highlighting the key stages (e.g., SWstorage derivation and model types) in shaded boxes or with color-coded legends for better reader navigation.

**Response:** To improve flowchart clarity we have revised the figure to add shaded boxes to grouped steps, like conversion to storage, consolidation of classified images, etc. The shading is coordinated to match the legend but also organize and highlight sections of the flowchart.

Figure 6: Excellent conceptual figure, but panel labels (a–f) are too small, and their legends require clearer links to figure captions. Consider separating the map and SW image panels to avoid clutter.

**Response:** We revised this figure to remove one of the map panels, replaced by a separate, new figure. The revised figure shows the best model per watershed and examples of the surface water dynamics. We have increased the size and clarity of all text including labels and legends and updated and clarified the caption.

Table 1: This table merges too many heterogeneous variable types. Recommend separating into three subtables: meteorological inputs, SWstorage metrics, and catchment characteristics.

**Response:** We believe that listing all of the variables in a single table better facilitates understanding the differences in variables groups included in each LSTM model. However, we also acknowledge that the table contains different variable types. To improve table interpretability, we added a data frequency column to the table.

Table 5: Statistically significant improvements are bolded but hard to interpret due to inconsistent use of p-values and percentages. Add explicit column labels such as “Improvement Significance” and asterisks for clarity (p<0.05, \*p<0.01).

**Response:** We have revised the former Table 5 to improve the clarity of the significance.

**Minor Suggestions**

Page 4, Line 14: “...we developed four Long Short-Term Memory (LSTM) models, that differed...” → remove comma after “models.”

**Response:** Change made as recommended.

Abstract: “Decreases in low flow (Q70) daily residuals averaged 47.6%...” → suggest rephrasing: “Residuals during low flow (Q70) events decreased by an average of 47.6%...”

**Response:** Because of abstract word limits, we revised this to, “ Residuals during low flow (Q70) events decreased by 47.6% when adding storage to meteorological data.”

Page 8, Line 18: “...was selected to coincide with the availability...” → could be clarified to “was chosen to match the data availability from Sentinel-1 and Sentinel-2 missions.”
**Response:** Change made as recommended.

Page 17, Line 7: “...assigned a nominal depth of 10 cm and 5 cm, respectively.” → specify clearly which value corresponds to open and vegetated water for reader clarity.
**Response:** We revised this sentence to improve clarity, as “pixels were assigned a nominal depth of open water = 10 cm and vegetated water = 5 cm.”

 Page 24, Line 11: “...median improvement in NSE of 9.3%.” → provide baseline NSE value here to contextualize significance.
**Response:** We added the baseline NSE values here to improve clarity.

Page 28, Line 6: “...seasonal to relatively permanent surface water extent...” → difficult phrasing. Consider: “seasonally persistent to permanent surface water extent.”

**Response:** Change made as recommended.