Dear Dr. Onyutha,

We sincerely thank you for handling our manuscript and the two reviewers for their careful evaluations and constructive comments. Their suggestions have helped us improve the clarity, rigor, and presentation of our study. In this revision, we have carefully considered each comment, revised the manuscript accordingly, and provided detailed point-by-point responses.

We would also like to briefly clarify the central objective of this study: to establish a computationally efficient coupling framework that retains the essential effects of three-dimensional groundwater dynamics on land-surface processes. The framework accommodates application-dependent accuracy requirements, and its evaluation therefore considers both groundwater-state errors and their implications for the coupled system. Our additional analysis shows that groundwater-state errors do not necessarily translate into proportionally increasing surface-flux errors, supporting the importance of assessing surrogate performance in relation to its intended coupling objectives.

We appreciate the reviewers' suggestions concerning long-term stability, generalizability, and the mechanisms underlying error growth. We have addressed these issues through the analyses and discussion presented in the revised manuscript. Where further investigation would require dedicated experimental designs to distinguish the contributions of different factors, we have explained the scope and limitations of the present analysis in our responses. We hope that these revisions and clarifications adequately address the remaining concerns while preserving the focus and coherence of the study.

Sincerely,

Chen Yang

On behalf of all Authors

Reviewer #1

The authors have addressed several previous concerns. Excluding the first 720 h of WY2021 from the main continuous evaluation improves the separation between validation and evaluation. The repeated timing experiments provide a more transparent basis for the reported computational acceleration. The expanded discussion is also useful.

Nevertheless, I still have a few questions, which are listed below. I think the paper can be considered for publication after addressing theses issues.

We thank the reviewer for the positive assessment of our revised manuscript. We address the remaining comments point by point below. All line numbers refer to the revised manuscript with tracked changes.

1. The abstract, Figure 3, and the discussion require hybrid errors to remain below both the PFFD discrepancy and a user-defined threshold. However, the manuscript does not specify the numerical threshold used, the mathematical definition of error, or the spatial and temporal aggregation over which acceptance is evaluated.

No numerical user-defined threshold was imposed in this demonstration. We used the PFFD–CoLM/ParFlow discrepancy as the benchmark, assessing domain-averaged latent heat flux over the 8040-h evaluation period with KGE and mean bias (Figure 7). For a specific application, users could set a stricter tolerance for this measure or for other quantities, such as WTD error, using the spatial and temporal aggregation appropriate to that application. We have clarified this distinction in the Results and Discussion (Lines 504–509): “In this demonstration, no additional numerical threshold was imposed beyond the PFFD-based benchmark, although users may specify a more restrictive error tolerance according to their application requirements. For other applications, users could set tolerances for WTD error statistics or domain-averaged latent heat flux error, with the metric and its spatial and temporal aggregation chosen according to the intended use.” We also revised the corresponding wording in the abstract, Figure 3 caption, and the discussion preceding Figure 7 (Line 490).

2.Figure 3 contains an internal inconsistency. Its lower schematic shows errors exceeding the blue user threshold while still being described as acceptable. The schematic and mathematical criterion should be reconciled.

We have revised Figure 3 and its caption to ensure that errors shown as acceptable remain within the blue user-defined threshold, consistent with the mathematical acceptance criterion.

3.In Figure 8a, the horizontal coordinate is reference WTD. In Figure 8b, it is absolute WTD error. The visible dashed line at 1 m occurs in panel b and therefore indicates an error magnitude. Thus, it cannot identify a transition between shallow and deep groundwater in Figure 8b, as stated in the caption of the figure.

We added Figure 8c to show latent heat errors as a function of physics-based WTD, so the shallow and deep groundwater conditions can be compared directly. We also revised the figure caption accordingly.

Reviewer #2

General comments

The revised manuscript is substantially clearer, and the authors have addressed all my previous comments. The revisions have improved the presentation of the methodology and results and have strengthened the overall argument of the manuscript. Separating the Results and Discussion could further improve the manuscript's organization, although the current structure is already clearer than the previous version.

I have only a few remaining questions regarding the interpretation of the results and the consistency of model performance across simulation periods. These issues appear addressable through clarification and additional analysis. I therefore recommend minor revision before acceptance. All comments below are based on the revised manuscript.

We thank the reviewer for recognizing that the previous comments have been addressed and that the manuscript is clearer. We have carefully considered the remaining questions raised in this round and respond to them point by point below. Unless otherwise specified, all line numbers refer to the revised manuscript with tracked changes.

Detailed comments

3.2 Results and analysis

The Results and Discussion could be presented separately to distinguish model performance from the interpretation of underlying mechanisms. However, the current structure is clearer than the previous version.

Separating the Results and Discussion could be an alternative way to organize the manuscript. After considering this option, we feel that the current subsection-based structure already provides a clear and coherent presentation of the results and their interpretations. Further separation may offer limited additional benefit, particularly given the close connections between the results and their corresponding discussions. We have therefore retained the current structure.

Lines 461–462: The two patches with larger errors appear to exhibit more pronounced slope characteristics than those with smaller errors. Does this indicate limitations in the hybrid model's ability to reproduce terrain-related spatial heterogeneity, particularly where topography strongly influences hydrological responses?

Not necessarily. As discussed in Lines 460–468 of the previous clean version, the larger errors in these two patches may reflect the combined influences of accumulated errors, wet–dry transitions, and local terrain and subsurface properties. Terrain slope is therefore considered one of several potential contributing factors, rather than evidence of a systematic limitation of the hybrid model in representing terrain-related spatial heterogeneity.

Lines 641–643; Figures 6, 9, 11, and S12: Figure 11 shows that the separately initialized prediction for Segment 9 reproduces river-related spatial patterns in soil moisture and WTD relatively well despite pronounced variability in the forcing data. However, these features become less distinct during the middle and late stages of continuous simulation (Figs. 6 and 9), while some remain visible during the early stage, when precipitation variability is less pronounced.

The authors should clarify whether the better performance in Segment 9 primarily reflects independent initialization and reduced error accumulation or the highly variable forcing. The KGE values of runoff in Fig. S12 also deteriorate substantially across the early, middle, and late stages (0.435, −4.56, and −544.638, respectively). The authors should verify and explain this deterioration, including whether it results from accumulated errors or other sources of instability. Under what conditions can the hybrid model reliably reproduce the fine-scale hydrological features captured by the physics-based model?

The better performance in Segment 9 primarily reflects its independent initialization and reduced prior error accumulation. The continuous and segmented predictions cover the same period and therefore experience identical forcing; their principal difference is the accumulated-error history. The purpose and conclusion of this experiment are already stated in Lines 672–675 and 712–715 (or Lines 608–611 and 648–651 of the previous clean version).

We also note that the KGE values cited in the comment refer to surface-water storage rather than runoff. The late-stage KGE in Figure S12 is −44.638, rather than −544.638. This deterioration is consistent with the progressive divergence of the predicted pressure fields. Streamflow is evaluated separately at individual grid cells and is more sensitive to local surface pressure head errors because of the nonlinear relationship between pressure head and discharge. Consequently, reproducing the broad water table depth pattern does not

necessarily ensure accurate streamflow at a particular location, as already explained in Text S4. The hybrid model reproduces fine-scale hydrological features most reliably when initialized close to the target period and operated before substantial autoregressive errors accumulate as discussed in section 3.4.

Lines 564–568; Figure 8: Figure 8 shows that absolute WTD errors vary with physics-based WTD (Fig. 8a) and are associated with absolute latent heat errors (Fig. 8b). However, Fig. 8b does not directly compare latent heat errors under shallow and deep groundwater conditions. The authors should compare latent heat errors between these regimes or revise Fig. 8b to distinguish samples with physics-based WTD below and above the stated threshold. If the authors attribute these differences specifically to hydraulic connectivity, additional supporting evidence is needed.

We revised Figure 8 by replacing the original scatter plots with binned boxplots and adding panel c, which shows latent heat errors across shallow to deep physics-based WTD conditions. The revised figure shows that latent heat errors increase with WTD error over an intermediate range, but their growth becomes less pronounced at larger WTD errors. We interpret this pattern in light of the well-established critical water table depth theory of groundwater–land surface interactions e.g., (Condon, Atchley, & Maxwell, 2020; Fan, Miguez-Macho, Jobbágy, Jackson, & Otero-Casal, 2017; Ferguson & Maxwell, 2011; Kollet & Maxwell, 2008; Maxwell & Condon, 2016; Maxwell & Kollet, 2008; Yang et al., 2020; Yang, Maxwell, McDonnell, Yang, & Tijerina-Kreuzer, 2023). We have also clarified this interpretation in the revised discussion (Lines 550–624).

Lines 594–595 and 614–616; Table S1: The authors report strong correlations between precipitation variability and upper-tail WTD errors across 12 independently initialized 720 h segments, particularly for the time-averaged 75th-percentile WTD error. However, Segment 9, which has the highest precipitation variability, is reproduced relatively well when initialized separately. Although these findings are not necessarily contradictory, the relationship between precipitation variability and error amplification remains unclear.

These findings are consistent. Although the independently initialized Segment 9 performs well in terms of the median WTD error and overall spatial patterns, Figure 10 shows a pronounced increase in the 75th- and 90th-percentile errors following the transition from wet to dry conditions. This is consistent with Table S1, which links precipitation variability primarily to upper-tail WTD errors rather than to overall model deterioration. Reinitialization limits the accumulated background error, while within-segment wet–dry transitions can still amplify localized upper-tail errors.

The authors should better connect Table S1 with the temporal evolution of errors discussed in the main text. Comparing precipitation variability and WTD errors across the early, middle, and late stages of continuous simulation would help clarify whether the reported association is consistent with the progressive deterioration. The contrasting behavior of Segments 1 and 9 and other segments should also be discussed to distinguish the potential influence of precipitation variability from accumulated autoregressive errors.

The 12-segment analysis in Table S1 was designed to facilitate this comparison under more comparable conditions. The early, middle, and late stages of the continuous simulation differ in duration and accumulated-error history, making their precipitation–error relationships difficult to compare directly. By using 12 equally long, independently initialized segments, we can compare multiple precipitation characteristics and WTD error metrics without the carryover of autoregressive errors from preceding segments. This design also allows us to examine the association across a range of conditions, rather than drawing conclusions from any individual segment or pair of segments. The results show stronger statistical associations with upper-tail WTD errors than with spatial mean or median errors, and we discuss these findings alongside the temporal evolution of errors in the main text (Lines 406–421).

Further comparisons within the continuous simulation could provide additional descriptive information, but would not, by themselves, separate the influence of precipitation variability from accumulated autoregressive errors. We therefore use the segment-level statistics to complement the discussion of progressive deterioration, considering precipitation variability alongside wet–dry transitions and local physical conditions (Lines 469–472). This approach keeps the analysis focused on the development and practical implications of surrogate errors within the hybrid framework, while avoiding attribution of error growth to any single factor.

Lines 624–626 and 648–651: The authors attribute the deterioration of continuous simulations to accumulated autoregressive errors and show that shorter prediction windows mitigate error amplification. However, this raises questions about the model's applicability to long-term hydrological simulations, which often require continuous multi-year predictions. The authors should clarify how the hybrid model could be better applied in such cases and whether its computational advantages would be maintained when short prediction windows require repeated initialization.

We thank the reviewer for raising this important and broad question. Long-term error drift is a field-wide challenge for autoregressive surrogate models in hydrology and meteorology and is unlikely to be resolved by any single model or study. We would therefore like to

share four perspectives on how this issue is addressed within the present framework and how it may be further mitigated.

First, the objective of our framework is not to reproduce every groundwater state from CoLM/ParFlow exactly over an unrestricted prediction horizon (Lines 107–109 and 479–483). Rather, it is to retain the essential effects of three-dimensional groundwater dynamics on land-surface processes at substantially lower computational cost. Some predictive drift is therefore anticipated and explicitly accommodated in the error-aware strategy proposed in Figure 3. The hybrid solution is considered useful when its application-relevant error remains within the user-defined tolerance when specified and its deviation from the fully coupled reference remains smaller than the PFFD–CoLM/ParFlow discrepancy. The results show that the hybrid system remains substantially closer to CoLM/ParFlow than PFFD in representing groundwater–land-surface feedbacks. Thus, its utility does not require indefinitely exact reproduction of the groundwater state.

Second, even under a stricter evaluation based directly on agreement with CoLM/ParFlow, the hybrid model exhibits low errors during the 720 h segmented predictions and good overall performance over the near-annual 8040 h continuous evaluation. For example, the median WTD error remains within 0.5 m and the domain-averaged latent heat flux achieves a KGE of 0.959. Prediction horizons of this length can already provide substantial practical value for event-oriented hydrological assessment, emergency preparedness, and scenario-based decision support when suitable meteorological forecasts or scenario forcings are available. Many such applications do not require uninterrupted multi-year surrogate integration.

Third, the effective prediction horizon and transferability are not necessarily fixed properties of the present model. In our related work, precipitation-based data augmentation was found to reduce long-horizon error growth. Separately, zero-shot experiments demonstrated that a groundwater surrogate trained in one basin could exhibit a degree of transferability to another basin without additional training or fine-tuning (Huang et al., 2026; preprint: 10.31223/X5QV3P). These findings suggest two complementary opportunities: improving prediction stability through more diverse training data and targeted augmentation, and potentially extending surrogate applications to new domains by leveraging their inherent transferability.

Fourth, future groundwater prediction systems may increasingly develop toward continuously operating physics-based simulation and data-assimilation frameworks, analogous to modern numerical weather-prediction systems. Such systems would continuously integrate observations, update physical states, and accumulate reusable training and initialization data. The physics-based model would remain an essential

component, but its updated states could support the initialization or periodic reinitialization of many downstream surrogate applications. Its computational cost would therefore form part of a shared operational infrastructure rather than being incurred separately for every surrogate prediction, while the hybrid model would retain its advantage for rapid, application-specific simulations.

Taken together, these considerations show that long-term surrogate use should be viewed not only as a question of extending an uninterrupted autoregressive rollout, but also as a combination of application-dependent error tolerance, improved training strategies, model transferability, and integration with evolving operational modeling and data-assimilation systems. These considerations are consistent with the discussion of limitations and future directions already provided in Lines 719–721 and 748–757.

Lines 653–654: The claim that a 720 h prediction window may be sufficient for many flood and drought applications should be qualified according to specific application requirements. The limitations of the current framework for long-term continuous simulations and potential measures to extend its effective prediction horizon should be discussed more explicitly in the Discussion and Conclusions.

We have revised the statement in Lines 653–654 to better qualify the potential applications of the 720-h prediction window according to specific application requirements (Lines 717–719). The limitations of the current framework for long-term continuous simulations and potential measures to extend its effective prediction horizon have already been discussed in the final paragraph of the Conclusions (Lines 748–757). We have also shared some additional perspectives on potential future developments in our response to the preceding comment.

References

Condon, L. E., Atchley, A. L., & Maxwell, R. M. (2020). Evapotranspiration depletes groundwater under warming over the contiguous United States. *Nature Communications, 11*(1), 873. doi:10.1038/s41467-020-14688-0

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., & Otero-Casal, C. (2017). Hydrologic regulation of plant rooting depth. *Proceedings of the National Academy of Sciences, 114*(40), 10572-10577. doi:10.1073/pnas.1712381114

Ferguson, I. M., & Maxwell, R. M. (2011). Hydrologic and land-energy feedbacks of agricultural water management practices. *Environmental Research Letters, 6*(1). doi:Artn 01400610.1088/1748-9326/6/1/014006

Kollet, S. J., & Maxwell, R. M. (2008). Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resources Research, 44*(2). doi:Artn W0240210.1029/2007wr006004

Maxwell, R. M., & Condon, L. E. (2016). Connections between groundwater flow and transpiration partitioning. *Science, 353*(6297), 377-380. Retrieved from <Go to ISI>://WOS:000380583400039

Maxwell, R. M., & Kollet, S. J. (2008). Interdependence of groundwater dynamics and land-energy feedbacks under climate change. *Nature Geoscience, 1*(10), 665-669. doi:10.1038/ngeo315

Yang, C., Li, H.-Y., Fang, Y., Cui, C., Wang, T., Zheng, C., . . . Yang, X. (2020). Effects of Groundwater Pumping on Ground Surface Temperature: A Regional Modeling Study in the North China Plain. *Journal of Geophysical Research: Atmospheres, 125*(9), e2019JD031764. doi:10.1029/2019jd031764

Yang, C., Maxwell, R., McDonnell, J., Yang, X., & Tijerina-Kreuzer, D. (2023). The Role of Topography in Controlling Evapotranspiration Age. *Journal of Geophysical Research: Atmospheres, 128*(18), e2023JD039228. doi:https://doi.org/10.1029/2023JD039228