Reviewer #1

In the paper entitled "A hybrid physics–ML framework for integrating groundwater dynamics into land surface modeling", the authors used a deep learning model to replace a physics-based solver in the ParFlow model. Overall this paper is well written. The structure is clear and description of the model is complete. However, there are still some issues to be clarified before the paper can be considered for a future publication in GMD.

We sincerely thank Reviewer #1 for the positive assessment of our manuscript and for the constructive comments. Below, we provide a brief point-by-point response, with detailed revisions to be presented in the revised manuscript.

1. In my understanding, because the training of the DL model is performed using the inputs and outputs of the simulations for the Pearl River Basin, China, the updated hybrid model is probably moslty suitable for this specific region. The authors may need to clarify it in the title and the abstract of the paper. Moreover, the name of the model ParFlow may also need to appear in the title and the abstract as well.

We agree that the trained surrogate is specific to the Pearl River Basin, whereas the proposed coupling framework is more generally applicable. We will clarify this distinction in the title and abstract and explicitly identify ParFlow as the physics-based model replaced by the surrogate.

2. L251, in my mind, the rest of WY2021 should be used for model testing instead of the full year, because the beginning of the year has been used for validating the model.

We agree that the first 720 h of WY2021, which were used for validation and model selection, should not be included in the independent evaluation. We will therefore evaluate the trained model over the remaining period of WY2021, beginning from hour 721, and update the corresponding results and figures.

3. L360-361, the authors should use some numbers such as the computing time for one-month simulation, to indicate the speedup influence of the replacement more clearly.

We will report the computational time required for a representative one-month simulation using both PF-CoLM and the hybrid model, together with the corresponding speedup and computational configuration.

Reviewer #2

General comment:

This manuscript develops a groundwater surrogate that combines static spatial information with dynamic hydrological inputs and is designed to replace ParFlow within a coupled land-groundwater modeling framework. The approach is demonstrated over a ~34,000 km^2 basin in the Pearl River Basin, China, and achieves an approximately 20-fold speedup while maintaining generally good agreement with the physics-based reference model. The proposed design preserves the coupling interface with the land-surface model and therefore has potential value for efficient hybrid land-groundwater simulation.

However, the Results and Discussion section should be substantially reorganized. The authors should structure this section into clearly defined thematic subsections and provide more detailed discussion of the unresolved issues revealed by the results, particularly regarding spatial errors, long-term stability, and the advantages and limitations of the Hybrid model relative to the physics-based PF-CoLM system.

We sincerely thank Reviewer #2 for the careful evaluation of our manuscript and for the numerous constructive comments, which will help substantially improve its clarity and quality. Below, we provide a brief point-by-point response. More detailed responses and corresponding revisions will be provided with the revised manuscript. In particular, we will reorganize the Results and Discussion into thematic subsections and expand the discussion of the unresolved issues identified by the reviewer.

Detailed comments:

2 Methodology

2.1 Physics-based CoLM/ParFlow model

Line 118-120: The rooting-zone depth (3.43 m) and aquifer thickness (100 m) differ from those commonly used in previous ParFlow applications. Could the authors clarify the rationale for this vertical discretization and whether it was mainly designed to match the CoLM configuration?

The overall vertical configuration follows the design of PFCONUSv1 (Maxwell et al., 2015), which consists of four soil layers totaling 2 m overlying a 100 m aquifer layer. Our configuration retains the same 100 m aquifer representation, while the upper root zone is discretized into the 10 native CoLM soil layers, totaling 3.43 m. Thus, the difference mainly reflects the vertical discretization of the coupled land surface model rather than a fundamentally different subsurface configuration. We will clarify this rationale in the revised manuscript.

2.2 ParFlow deep-learning surrogate model

Lines 176-177 and 234-235: The manuscript emphasizes that, unlike the original FSTR using meteorological forcing directly, the modified surrogate uses CoLM-derived root-zone net fluxes as external inputs. However, these fluxes are themselves responses to meteorological forcing processed through land-surface processes. The authors should clarify the advantage of using root-zone fluxes instead of meteorological forcing directly and how this flux-conditioned design differs from a conventional forcing-driven surrogate model. Does this modification mainly enable coupling with CoLM by preserving the ParFlow-CoLM interface, or does it also improve surrogate accuracy, stability, or physical consistency?

The primary purpose of using CoLM-derived root-zone net fluxes is to preserve the original ParFlow–CoLM coupling interface, thereby allowing the surrogate to directly replace ParFlow within the online coupled system. This design was not intended to improve surrogate accuracy or stability. We will clarify this distinction in the revised manuscript.

Lines 250-252 and 255-256: The surrogate is trained only on WY2019–2020 using 120 h temporal windows. The authors should justify the choice of these two years and clarify whether the training dataset sufficiently represents wet, dry, and extreme hydrological conditions.

WY2019–2020 were selected because they provided a continuous simulation period generated during the development of the ParFlow–CoLM model, rather than being deliberately chosen to represent specific hydroclimatic conditions. In the revised manuscript, we will further analyze and clarify the extent to which the training period represents wet, dry, and extreme hydrological conditions.

Lines 250-256: The surrogate is trained exclusively using meteorological forcing and ParFlow simulations from the Pearl River Basin. It is therefore unclear whether the resulting surrogate represents a general approximation of ParFlow dynamics or is specific to the hydrological and spatial characteristics of the training basin. The authors should clarify this point.

The trained surrogate is basin-specific and should not be regarded as a general approximation of ParFlow dynamics. It is expected to be reliable primarily within the range of hydrological and spatial conditions represented in the training data, while its performance under unseen extreme conditions or in other basins cannot be guaranteed. The coupling framework itself is general, but applications to other domains may require additional ParFlow simulations and retraining or fine-tuning. We will clarify this distinction and add a corresponding discussion of this limitation in the revised manuscript.

Lines 267-268: Pressure heads are standardized using their mean and standard deviation for each vertical layer. Please discuss whether this standardization affects the representation of vertical hydraulic gradients, particularly under very dry or saturated conditions.

We will assess the potential effect of layer-wise standardization by comparing the vertical hydraulic gradients produced by the Hybrid and physics-based models under dry and saturated conditions. The corresponding results and discussion will be added to the revised manuscript.

2.3 Additional architectural experiments

Lines 277-278 and 302-303: The purpose and added value of the temporal self-attention module are not sufficiently clear. The authors should clarify whether it is primarily intended to improve long-term dependency learning, prediction accuracy, autoregressive stability, or computational efficiency. If the improvement is limited, this section should be streamlined and greater emphasis placed on the core surrogate and coupling framework.

The temporal self-attention module was introduced to better capture the long-term memory of groundwater dynamics and thereby potentially improve prediction accuracy and autoregressive stability, rather than computational efficiency. Although the improvement was limited, we believe this result provides a useful reference for the community. We will therefore streamline this section in the main text, move additional details to the Supplement, and place greater emphasis on the core surrogate and coupling framework.

2.4 ParFlow surrogate-CoLM physics hybrid coupling

Lines 342-343: Given that the hybrid model is evaluated only on WY2021, please discuss whether a single evaluation year is sufficient to demonstrate long-term robustness across different hydroclimatic conditions.

As noted above, evaluation over a single year cannot fully demonstrate long-term robustness across all hydroclimatic conditions. The surrogate is expected to perform better under conditions represented in the training data, whereas unseen conditions may require additional training or fine-tuning. We will clarify this limitation in the revised manuscript.

Lines 358-360: Does the reported speedup apply only to the ParFlow-CoLM coupled simulation, excluding the ParFlow spin-up? If a full ParFlow spin-up is still required for application to a new basin, the overall computational benefit of the surrogate may be considerably smaller. Clarify this limitation and discuss the end-to-end computational cost.

We agree that the reported speedup applies to the coupled simulation and does not include the ParFlow spin-up required for a new basin. We will clarify this limitation and report the end-to-end computational cost. However, with our continental-scale modeling infrastructure for China, a new basin can be initialized from an existing equilibrated continental state, substantially reducing or largely avoiding the additional ParFlow spin-up cost.

3 Results and discussion

I do not recommend presenting the Results and Discussion primarily in the order of individual figures. In its current form, this section reads largely as a figure-by-figure description, making it difficult to identify the main scientific findings and distinguish the presentation of results from their interpretation.

The authors should reorganize this section into clearly defined subsections, with each subsection addressing a specific aspect of the results. The authors may determine the specific subsection structure according to the content. For example, Lines 373-391 could be consolidated into a subsection focusing on differences in spatial patterns between the hybrid and physics-based simulations. Such a structure would substantially improve the readability and scientific focus of this section.

We agree with the reviewer's suggestion. In the revised manuscript, we will reorganize the Results and Discussion into clearly defined thematic subsections, with each subsection focusing on a specific scientific finding.

The manuscript evaluates WTD, soil moisture, and latent heat flux, but runoff is not discussed. Please clarify why runoff is excluded and include an evaluation of runoff performance, given its importance in assessing the coupled hydrological system.

The primary focus of this study is groundwater dynamics and groundwater–land surface interactions, and runoff was therefore not emphasized in the original manuscript. Nevertheless, we agree that runoff provides a useful complementary measure of coupled-system performance. We will add a concise runoff evaluation in the revised manuscript while retaining the primary focus on groundwater-related processes.

Figure S2 and Lines 347-348: The hybrid model is trained five times using different random initializations. Clarify what the labels in Fig. S2a represent and what specifically differs among these five realizations.

We will clarify in the figure caption and text that the labels represent five independently trained realizations that differ only in their random weight initialization.

Line 356: The checkpoint with the lowest validation RMSE was selected for subsequent evaluation. Identify the selected checkpoint and report the corresponding random initialization and training epoch/iteration. Please also clarify how sensitive the reported evaluation is to the checkpoint selection.

All five realizations were compared after the same initial training period of approximately (8×10^4) iterations, and the realization with the lowest validation RMSE at that stage was selected. Two runs were subsequently extended only to examine whether further training would improve performance, but no meaningful improvement was observed and these extensions were not involved in model selection. The selection was therefore made across random initializations at the same training stage, rather than across checkpoints from different epochs. We will identify the selected realization and clarify this procedure in the revised manuscript.

Lines 378-380: The authors attribute the finer-scale heterogeneity to soil texture and land-surface properties. However, if the hybrid model exhibits stronger or finer-scale spatial variability than the physics-based simulation, the authors should clarify why this occurs and whether these additional spatial features are physically meaningful or artifacts introduced by the surrogate.

The referenced statement describes the fine-scale spatial heterogeneity generally observed in the simulated fields and attributes it to soil texture and land-surface properties. It was not intended to suggest that the hybrid model produces stronger or additional spatial variability relative to the physics-based model. We will revise the wording to make this distinction clearer.

Line 385: Large errors appear to be concentrated in the southeastern part of the domain. The authors should provide a more detailed discussion of the possible causes of this spatial error pattern.

Our preliminary interpretation is that this pattern arises from accumulated errors in the ParFlow surrogate, which may be further amplified through iterative coupling with CoLM. We will examine the southeastern region in greater detail and discuss the underlying causes in the revised manuscript.

The Results and Discussion should focus more clearly on the performance gain of the hybrid model relative to the original PF-CoLM system. The central contribution of this manuscript is the development of a surrogate model to replace or accelerate ParFlow within the coupled framework. Therefore, the key comparisons should be hybrid vs. PF-CoLM and hybrid vs. observations/reference datasets.

Lines 412-419 and 443-449: Considerable emphasis is currently placed on comparisons between PF-CoLM and the PFFD/no-lateral-flow configuration. While PFFD can serve as a useful secondary benchmark for illustrating the effects of simplified groundwater treatment, these comparisons do not directly demonstrate the added value of the hybrid model and should not dominate the main Results and Discussion.

I therefore suggest substantially reducing the discussion of PFFD vs. PF-CoLM and retaining PFFD only as a secondary reference case.

The three comments above concern the same issue and are therefore addressed together. We agree that Hybrid–PF-CoLM should remain the primary comparison. However, the central contribution of this study is not the development of a new high-accuracy ParFlow surrogate, but the use of surrogate-enabled coupling to represent groundwater and its land-surface feedbacks. PFFD therefore provides a useful secondary benchmark representing the absence of groundwater feedback, allowing us to assess whether the hybrid model retains meaningful groundwater–land surface interactions as surrogate errors accumulate. We will rebalance the Results and Discussion accordingly while retaining this interpretive role of PFFD.

Figure S4d: I do not recommend including the Spearman-correlation comparison between the free-drainage and lateral-flow ParFlow configurations. Since both are physics-based ParFlow configurations representing different subsurface treatments, this comparison provides limited information for evaluating the hybrid model.

We respectfully retain a different view on this point. The purpose of Figure S4d is not to directly evaluate the hybrid model, but to establish the contrast between the full groundwater representation and the free-drainage configuration without lateral groundwater feedback. This comparison provides a secondary reference for assessing how much groundwater-related behavior the hybrid model retains as errors accumulate, which is central to interpreting its effective prediction horizon. As this analysis is supplementary rather than part of the main Results, we consider its inclusion appropriate and will clarify its purpose.

Lines 421-423 and 438-440, Figure 7: The authors should provide a quantitative discussion of the latent heat flux errors shown in Figs. 7b and 7c, particularly for the comparison between hybrid model and ERA5-Land. The key performance metrics and their values shown in Fig. 7 should also be reported explicitly in the text to allow readers to assess model performance without repeatedly referring to the figure.

The relevant performance metrics are already presented in Figure 7d but are not reported explicitly in the text. We will add their numerical values and provide a more quantitative discussion of the latent heat flux errors in Figures 7b and 7c, particularly for the comparison between the hybrid model and ERA5-Land.

Lines 450-452: Although the hybrid model provides substantial computational speedup, its improvement in simulation performance appears relatively modest. For example, the physics-based model already shows a high KGE against ERA5-Land for latent heat flux, with only a small improvement from hybrid. The authors should clearly demonstrate and discuss the specific advantages of hybrid over the physics-based model beyond computational efficiency.

The hybrid model is not intended or expected to outperform PF-CoLM in simulation accuracy. Its advantage lies in substantially reducing computational cost while retaining the key groundwater–land surface interactions represented by PF-CoLM. The slightly higher KGE against ERA5-Land should therefore not be interpreted as an improvement over the physics-based model. We will clarify this point in the revised manuscript.

Figure 9: The river-network-related spatial patterns remain clearly visible in the physics-based simulation but are much weaker in the hybrid results, including the bias and correlation maps. Could this indicate a deficiency of the hybrid model in reproducing the river-network features captured by the physics-based model? The authors should explicitly discuss this issue and explain its possible causes.

We agree that the weaker river-network patterns indicate a limitation of the hybrid model in reproducing sharp, localized hydraulic features. Our preliminary interpretation is that the abrupt transition between positive pressure heads in river cells and below-surface water tables in surrounding areas is more difficult for the surrogate to capture. We will further examine and quantitatively evaluate the performance in river-network areas and discuss its possible causes. We will also clarify that the primary focus of this study is domain-scale groundwater–land surface interactions rather than channel-scale dynamics.

Lines 481-485 and Figure 8: Figure 8 appears to provide limited additional information, as the upward propagation of subsurface-state errors into surface fluxes has already been demonstrated and discussed earlier. The authors should either clarify the distinct contribution of this figure or consider removing/streamlining it to avoid repetition.

Figure 8 provides information distinct from the preceding domain-averaged temporal analysis by examining the relationship at the grid-cell level. It shows that WTD errors generally increase with groundwater depth, while their effects on latent heat depend strongly on the groundwater regime: large WTD errors under deep water tables may have limited surface impacts, whereas smaller errors under shallow, hydraulically connected conditions can produce substantial flux deviations. We will revise and streamline the discussion to make this distinct contribution clearer.

Lines 510-511: Large hybrid-Physics discrepancies in the southeastern region are mentioned repeatedly, but the manuscript does not provide a detailed explanation for this persistent spatial error pattern. The authors should discuss the possible causes of the larger errors in this region.

This comment concerns the same persistent spatial error pattern raised in the comment on Line 385 and is addressed together with that comment. As noted above, our preliminary interpretation is that accumulated surrogate errors may be further amplified through iterative coupling with CoLM. We will investigate this region in greater detail and provide a unified explanation in the revised manuscript.

Line 525: The progressive deterioration is attributed to heavy precipitation, but the supporting evidence is unclear. Please specify whether this interpretation is based on Fig. 6 or Fig. 10 subplot a and cite the corresponding figure. A spatial precipitation map should also be provided to further examine whether regions of large hybrid errors coincide with areas of high precipitation.

We agree that the evidence supporting this interpretation should be clarified. In the revised manuscript, we will further examine the relationship between precipitation and hybrid-model errors using quantitative analysis, specify the relevant supporting figure, and provide a spatial precipitation map to assess whether and to what extent their spatial patterns correspond.

Figures 5, 9, and 11: In the full-year simulation shown in Fig. 9, the hybrid model does not reproduce the river-network features in soil moisture as clearly as the physics-based model. The same behavior is also evident in Fig. 5, whereas Fig. 11 shows much clearer river-related WTD patterns during a selected short-term heavy-rainfall event. The authors should discuss the possible reasons for this difference between long-term and short-term predictions.

As noted in our response above, river-network features involve sharp and localized hydraulic transitions and are therefore particularly sensitive to surrogate errors. In the continuous full-year simulation shown in Figures 5 and 9, autoregressive errors accumulate over time and progressively weaken these features. In contrast, the 720 h segmented prediction in Figure 11 limits error accumulation and therefore preserves the river-related patterns more clearly. We will clarify this distinction in the revised manuscript.

Figures 6 and 10: For the same period (hours 5761-6480), the median WTD bias is approximately 0.3-0.4 m in the continuous full-year simulation (Fig. 6) but remains below about 0.1 m in the segmented 720-h prediction (Fig. 10). Please explain why the same period produces such different error magnitudes under the two prediction methods.

Although the two simulations cover the same period, they have different prediction histories. The continuous simulation enters this period after approximately 5760 h of autoregressive prediction and therefore carries forward the errors accumulated previously. In contrast, the segmented simulation is reinitialized at hour 5761 and accumulates errors only within the subsequent 720 h window, resulting in a substantially smaller WTD bias. We will clarify this distinction in the revised manuscript.

If Fig. 9 represents a long-term simulation and Fig. 11 represents a short-term heavy-rainfall event, the results seem to suggest that the hybrid model loses clear river-network features during long-term simulation but can reproduce them during short-term high-rainfall conditions. The authors should discuss whether this indicates that the hybrid model is more suitable for short-term heavy-rainfall simulations than for long-term hydrological simulations.

The current results indicate that the hybrid model is more reliable for 720 h predictions than for continuous long-term simulations. Its strong performance during the selected heavy-rainfall period demonstrates that it can capture rapid hydrological responses under strong forcing, whereas the degradation in the full-year simulation is primarily associated with accumulated autoregressive errors. This suggests that periodic reinitialization or future data assimilation may help extend the effective prediction horizon. We will make this distinction clearer in the revised manuscript.

Lines 573-574: I agree that the hybrid model appears promising for short-term flood simulations, as it provides substantial computational speedup while reproducing river-network features and overall model performance comparable to the physics-based simulation. However, the current results do not convincingly demonstrate its suitability for long-term or decadal simulations. Long-term application requires substantial training data, and the hybrid model appears to lose clear river-network features, whereas the physics-based model maintains good performance. Under these conditions, the main demonstrated advantage of the hybrid model appears to be computational efficiency. The authors should discuss this limitation more explicitly.

We acknowledge that the current results do not demonstrate the suitability of the hybrid model for decadal simulations. We will discuss this limitation more explicitly in the revised manuscript, particularly regarding long-term error accumulation, training-data requirements, and the degradation of fine-scale spatial features.

Conclusion

The hybrid model is developed and evaluated only for the Pearl River Basin. The authors should clarify whether the trained surrogate model can be directly applied to other basins, or whether new ParFlow simulations, training data, and retraining would be required for each new domain. The manuscript should also distinguish between the general applicability of the proposed framework and the transferability of the trained surrogate model.

This issue is consistent with the related comment above and with Reviewer #1's comment regarding the regional scope of the study. The trained surrogate is specific to the Pearl River Basin and cannot be assumed to transfer directly to other basins; application to a new domain would generally require domain-specific ParFlow simulations and retraining or fine-tuning. In contrast, the proposed coupling framework is generally applicable. We will clarify this distinction throughout the revised manuscript, including the title and abstract.