This manuscript proposes a novel application of non-local neural networks for multi-depth soil moisture prediction and presents promising results on both synthetic and in-situ datasets. The idea of reformulating the problem as a single-step, profile-wise prediction and visualizing non-local weights to infer inter-layer interactions is valuable and interesting. The manuscript is well organized, but may need improvement in conceptual clarity, methodological transparency, and consistency of physical interpretation. Please see below for my comments.

**Major comments:**

1. Ambiguity of "Interpretability" and "Physics / Knowledge Guidance": A central claim of the paper is that the proposed framework is *interpretable* and *physics- or knowledge-guided*, but these concepts are not clearly or rigorously defined.

2. Methodological Clarity and Internal Consistency: Several aspects of the model formulation and workflow are insufficiently specified or internally inconsistent.

3. Fairness and Transparency of Model Comparisons: The manuscript claims superior performance of KG-NLNN over LSTM and SA-NLNN, but it is unclear whether the comparisons are fair and well controlled.

4. Interpretation of Weight Maps and Physical Meaning: The visual analysis of non-local weight matrices is central to the manuscript's interpretability claims, but the physical interpretation is currently speculative.

5. Claims of Generalizability and Flexibility: The manuscript claims that the framework is flexible and easily customizable, yet also indicates that models are trained site-specifically.

**Specific comments:**

Title: Does the "interpretable" mean that the deep learning model is interpretable, or the modeling results are interpretable? Additional, what are the new findings, which the interpretable approach can obtain while traditional models cannot

Line 20 and 25: As pointed in the prior comments, "guided by physics" and "physics guidance" can be misleading.

Line 20: While the authors showed that KG-NLNN outperformed the others, they should justify the fairness of the comparisons e.g., the number of input and tuning parameters are similar across the models.

Line 50: What is an "extreme learning machine"?

Line 70: The introduction of physics-informed neural networks appears to be distractive.

Line 75: The citation formats are not consistent.

Line 90: Are you referring to time stepping or iterative solver at each time step? The current statement does not provide a clear explanation of how the spatial interactions were done, nor the difference between the current method and physical methods. Please elaborate.

Line 90: "Temporal variations …" does not logically flow in the current context.

Line 95: What does it mean by "GNN rely on explicit, pre-defined graph structures"?

Line 100: The NLNNs sound like a GNN that considers hopping (message aggregation) on the entire graph. Can you clarify if that is the case? If not, please elaborate the difference then.

Line 100: How generalizable the model will be?

Line 115: What “interpretability” are you referring to? What are the new findings that were obtained due to this “interpretability”?

Line 120: “three-dimentional soil column” Did you model the three-dimmentional flow problem?

Line 125: Why is the current model not time series data processing?

Line 130: What are the physical meanings of the weights and how will they help interpret the results?

Line 135: Again, the citation format should be corrected.

Figure 1: I would suggest providing the full name of “SM” before using the acronym.

Line 150: Should it be ”manuscript” or “paper”?

Line 155: What does it mean by “under the assumptions that preferential flow”?

Line 170: The assumption appears the opposite of the statement in Line 90. Can you clarify which one should be the case?

Line 180: Are both methods new? Please clarify.

Line 185: The predictions did not include “sm_0^t+1’” in Figure 2. Additionally, the vector size of the predictions is not consistent with that of the ground truth.

Figure 2: Was SA score, KG score done separately or together? The current plot appears to show they are two comparts within one framework.

Figure 2: For the fully connected NN, what are the dimensions of nodes and layers?

Equation (2): Please define y_i.

Line 210: Does the weight matrix Wg depend on the number of locations being examined and their specific depths? In other words, will the Wg be directly applied to other problems?

Line 215: “It is evident that xxx”: NLNN reads like a transformer, please elaborate the difference.

Line 240-250: Is the r_score new? If not, I would either clarify that in the main text or move them to the SI.

Line 255: “The four masks in Figure 2 …” Is this conceptually similar to PCA?

Line 255: “Each of these components …” looks not consistent with the figure.

Line 270: Can the two layers be defined as “deep” learning?

Line 270: Commonly used in your research may not justify the decision. I would add references to support your choice.

Line 300: Can you justify the reason for such selection?

Line 330: Can you show the training loss and validation loss over epochs?

Line 425: “The color brightness on the …” Can you explain how this can be understood conceptually?

Line 435: Why LSTM_1 is introduced after LSTM_4? How about how about 2 and 3? I would suggest using a more descriptive naming instead of using numbers.

Figure 7: What do the weights represent?

Line 460: “Figure 7 depicts the weight matrix maps …” How about the homogeneous and heterogeneous cases?

Line 470: What case is being discussed? What is the difference between Figure 7 and Figure 8.

Line 470: “In the scenario is beath loam …” Which figure should be looked at?

Line 480: “the weight map of the SA-NLNN model appears slightly chaotic” Can you perform a test on more layers, with increasing or decreasing Ks to see if this trend holds?

Line 490: What is the reason that the homogeneous case was shown first in figure 4, but here was discussed after the two-layer case?

Line 510: Why is the comparison challenging? Can you elaborate the reasoning?

Figure 9: Are the RMSE average across both depths and sites?

Table 3: Why were some numbers bold?

Figure 11: Can you label the sites on the figures?

Line 590: Can you quantify the accuracy?

Figure 13: Why were some red curves cut off?

Line 650: “What’s more, the proposed network framework is flexible and easily customizable to suit specific requirements,” is not consistent with the fact that the model is “site-specific”. Can you elaborate on your reasoning?