Sam Houston State University
DEPARTMENT OF ENVIRONMENTAL AND GEOSCIENCES

Manuscript Number: egusphere-2025-3540

Renjie Zhou and Shiqi Liu. "A hybrid Kolmogorov-Arnold networks-based model with attention for predicting Arctic River streamflow"

Reviewer 1:

Zhou and Liu present a novel approach for a data-driven model for discharge modelling. It is based on a Kolmogorov-Arnold network combined with a Long-Short Term Memory (LSTM) model, an attention mechanism that includes a trigonometric depiction of seasonal patterns, as well as a physics-based constrain. The newly developed model aimed at improving the prediction of discharge within arctic areas with their special characteristics like perma frost and accumulation and melting of snow over longer periods. Therefore, the model was applied to the discharge data of the Kolyma River in Siberia and the prediction evaluated against the predictions of several other simpler models.

I have found the presented modelling approach to be a novel and valuable contribution to the hydrological modelling community. I believe it to be fitting for the scope of the Journal. However, the presented manuscript needs work regarding the methodology section as well as the discussion.

Reply: We are grateful for the reviewer's positive feedback and constructive suggestions. We have thoroughly revised the manuscript, corrected errors, added references, addressed each comment, and provided the necessary clarifications as outlined below.

Major comments:

1. Line 30: I can't really support the statement that the presented framework is (better) suited for predicting Arctic River discharge under changing climate conditions. It is well likely that climate change impacts the respective catchments in a way that the general behaviour changes - which also alters how discharge forms. I then get to a model space where the model has to extrapolate - which data-driven models are unsuited for.

Reply: Implemented. We thank the reviewer for raising this point. We have revised the statement in the manuscript and added a short paragraph about its limitations. While the RCPIKLA model demonstrates robust performance for the Kolyma River prediction under historical and current hydroclimatic conditions, several limitations should be acknowledged. As a data-driven model trained on historical observations, the model's performance may degrade if climate change induces fundamental shifts in watershed behavior that extend beyond the range of training conditions. Such regime changes may include but are not limited to scenarios like transitions from continuous to discontinuous permafrost, and significantly altered seasonal patterns. Under such scenarios, the model would need to extrapolate beyond its training data range, which remains a challenge for data-driven approaches. Future applications under changing climate conditions should include regular model retraining and validation as new observations become available.

2. Line 137, Figure 1: I personally don't think the figure to be well chosen, as the important aspects are missing. I would rather use a fogure that shows the catchment itself with its topography.
   Reply: Implemented. We thank the reviewer for this advice. The figure has been updated to show the catchment itself with its topography. Also, the input variables over the entire time span will be plotted and provided along with the catchment map.

3. Line 138-143: These lines are unnecessary here and probably can be deleted. All those things have already been said within the introduction and are explained over the methodology section anyways.
   Reply: Implemented. These lines have been deleted and revised to increase the flow. The sections of introduction, study area, and data acquisition, and methodology are reorganized to improve the flow and reduce the overlap.

4. Line 144-145: All steps, that are necessary for actual model runs should come after the model description. Otherwise, the order is confusing.
   Reply: Implemented. It is reorganized to improve readability and clarity. The preprocessing step has been improved with the following introduction after the model description.
   Prior to model training, the input variables, including monthly precipitation, temperature and evapotranspiration data, are preprocessed and standardized using the Z-score normalization technique: $X_{std} = \frac{X-\mu}{\sigma}$, where μ and σ are the mean and standard deviation computed from the training dataset; X and X_{std} denote the input values before and after standardization, respectively. This standardization process ensures that features with different scales contribute appropriately to the training process and improves model convergence.

5. Line 146-164: The description of the whole model structure should be done after the individual parts are explained. Figure 2 also should be moved there.
   Reply: Implemented. We have moved the whole model struction description after introducing all individual components.
   In summary, this newly proposed hybrid model leverages the KAN component as a feature transformation layer to extract and learn complex nonlinear patterns from hydrological and meteorological datasets. The LSTM component captures short- and long-term dependencies and effectively simulates sequential patterns and discharge variability. To further refine temporal learning, the attention mechanism is introduced and integrated, which allows the proposed model to selectively emphasize historically significant time steps, particularly those driving major and seasonal hydrological transitions. An important innovation is the residual compensation structure, which explicitly addresses the challenges of predicting extreme discharge events. By learning systematic error patterns, the

residual structure can adjust simulations based on residual predictions and improve performance during high-variability scenarios. Unlike conventional data-driven models that completely ignore fundamental physical constraints, the newly developed model incorporates physics-informed loss functions. Additionally, the model employs seasonality-aware encoding using trigonometric transformations to recognize the cyclic nature of hydrological processes. This architecture is designed to provide an accurate and robust framework for forecasting river discharge in Arctic and permafrost-dominated environments.

6. I do recommend the inclusion of an additional efficiency measure like KGE, that is complementary to the other ones and also incorporates different aspects of the discharge like bias for example. Please also cite and mention, which version of the KGE you use then.

   <u>Reply:</u> Implemented. We have added KGE' (2012) as an additional evaluation metrics.

   The following introduction is added to the subsection of 3.6 Evaluation Metrics. Also, pictures, references and discussion are revised and updated accordingly.

   In addition to NSE and RMSE, the Kling-Gupta Efficiency (KGE) is employed to provide a balanced assessment of model performance. The KGE metric was developed to address certain limitations of NSE, particularly its sensitivity to extreme values and the potential compensation of errors in mean, variance, and correlation (Gupta et al., 2009). Unlike other metrics, KGE explicitly decomposes model performance into three components: linear correlation, bias ratio, and variability ratio. In this study, the modified KGE is employed, which addresses issues with the original formulation's sensitivity to the magnitude of standard deviations (Kling et al., 2012). The modified KGE (KGE') is calculated as:

$$KGE' = 1 - \sqrt{\left(r_{kge} - 1\right)^2 + \left(\beta_{kge} - 1\right)^2 + \left(\gamma_{kge} - 1\right)^2},$$

   where r_{kge} refers to the linear correlation coefficient between observed and simulated discharge; β_{kge} refers to the ratio of simulated mean to observed mean; γ_{kge} denotes the variability ratio. The KGE' ranges theoretically from -∞ to 1, with KGE' = 1 indicating perfect agreement between observations and predictions in terms of correlation, bias, and variability. A KGE' value of -0.41 represents the performance of using the mean flow as a predictor, serving as a natural benchmark below which model predictions are no better than simply using the long-term average (Knoben et al., 2019). In hydrological modeling applications, KGE' values above 0.75 are generally considered very good, values between 0.5 and 0.75 indicate satisfactory performance, and values below 0.5 suggest unsatisfactory model performance (Towner et al., 2019). The use of multiple complementary metrics (NSE, RMSE, and KGE') provides a comprehensive evaluation framework. While NSE emphasizes matching variance and is sensitive to peak flows, KGE' provides balanced assessment across correlation, bias, and variability. RMSE quantifies absolute error magnitude in original units, which is particularly important for operational applications. Together, these metrics enable thorough assessment of model performance across different aspects of discharge prediction, from overall pattern matching to peak flow accuracy.

The KGE' metric provides additional insights into model performance by decomposing errors into correlation, bias, and variability components. The RCPIKLA model achieves KGE' values ranging from 0.74 to 0.82 across all time steps. Similar to NSE, the RCPIKLA model reaches its peak KGE' performance of approximately 0.82 at the 9-month time step. The baseline models demonstrate modest KGE' performance, with values ranging from 0.64 to 0.73. A notable degradation in KGE' performance is observed at the 12-month time step, where the RCPIKLA value drops to approximately 0.74, falling below the 0.75 threshold. This decline likely reflects the challenges of maintaining balanced performance across all three KGE' components (correlation, bias, and variability) at very long forecasting horizons. At 12 months, accumulated prediction errors and the increased difficulty in capturing seasonal phase transitions may cause the model's predictions to exhibit greater bias or variability mismatch compared to observations, despite maintaining reasonable correlation.

Reference:

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: inherent benchmark or not? Comparing nash–sutcliffe and kling–gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan De Perez, E., and Stephens, E. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin, Hydrol. Earth Syst. Sci., 23, 3057–3080, https://doi.org/10.5194/hess-23-3057-2019, 2019.

7. Why does the methodology end here? Important parts that come up later within the results part are missing. The methodology should explain that the final model is compared to certain baseline models and how they distinguish from the new model presented here. Furthermore, the whole part is missing about how the model is trained on the data, with how many runs, ending criterion, hyper parameters and so on.

   Reply: Implemented. We have restructured amd revised the manuscript and have add a subsection (Section 3.7) of model implementation and training to introduce the models and model differences, such as with how many runs, ending criterion, hyperparameters and so on.

   3.7 Model implementation and training

   As shown in Fig. 4, prior to model training, the input variables, including monthly precipitation, temperature and evapotranspiration data, are preprocessed and standardized using the Z-score normalization technique: $X_{std} = \frac{X-\mu}{\sigma}$, where μ and σ are the mean and standard deviation computed from the training dataset; X and X_{std} denote the input values before and after standardization, respectively. This standardization process ensures that features with different scales contribute appropriately to the training process and improves model convergence (LeCun et al., 1998).

In regions dominated by permafrost, snow accumulation and melt typically exhibit strong seasonal periodicity (Andersson et al., 2021; Ernakovich et al., 2014). Discharge patterns are strongly influenced by annual cycles of temperature, snow accumulation, and melt in Arctic hydrological systems (Häkkinen and Mellor, 1992). Accurately capturing such periodic behaviors can help develop robust long-term forecasting models. To include these cyclical patterns and facilitate smooth temporal transition, a trigonometric encoding (TE) of seasonal features is incorporated as input variables using sine and cosine transformations of the calendar month. Specifically, the timestamp is encoded to two features using the following trigonometric transformations:

$$\text{Month}_{\sin} = \sin\left(2\pi \frac{m}{12}\right); \ \text{Month}_{\cos} = \cos\left(2\pi \frac{m}{12}\right),$$

where m refers to the calendar month $m \in \{1,2,\cdots,12\}$. These encodings aim at capturing cyclical temporal patterns without introducing artificial discontinuities between December and January. The trigonometric features are concatenated with other input variables, including temperature, precipitation and evapotranspiration, and fed into the residual-compensated physics-informed KAN-LSTM model with attention.

Table 1 summarizes the hyperparameters and configuration settings used in this study. The choice of hyperparameters balances model capacity with overfitting risk, given the limited training data available. The LSTM hidden dimension of 64 units and a dropout rate of 0.3 prevent overfitting while capturing essential temporal patterns. The batch size and epoch size are set to 32 and 150, respectively. The optimal physics constraint weight (β = 0.3) and the MSE weight (α = 0.7) are adopted by conducting grid search over $\alpha \in \{0.1,0.3,0.5,0.7,0.9\}$ (Figure S1 in Supplementary Material). With these hyperparameters, the newly proposed model trained in the training dataset of the Kolyma River, and then the fine-tuned models are applied to the unseen testing dataset for the assessment of the predictive performance. The prediction performance is compared with several popular temporal baseline models, including the simple RNN, LSTM, and GRU models. To assess model stability and minimize the effects of stochastic processes in the training procedure, each model configuration is trained 10 times independently on Google Colab. This repeated training protocol allows assessment of performance variability arising from the inherent stochasticity in the optimization process, including random batch shuffling and numerical precision variations.

Table 1 Model hyperparameters and configuration settings

Parameters	**Values**
Training Epochs	150
Batch size	32
Learning rate	0.0005
Optimizer	Adam
Early stopping patience	10
MSE weight (α)	0.7
Physics constraint weight (β)	0.3
KAN grid size	5
KAN number of layers	2
LSTM hidden dim	64
Baseline models hidden dim	64
Dropout	0.3
Attention activation	Tanh
Output activation	ReLU
Number of runs	10

References:

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R.: Efficient BackProp, in: Neural networks: tricks of the trade, vol. 1524, edited by: Orr, G. B. and Müller, K.-R., Springer Berlin Heidelberg, Berlin, Heidelberg, 9–50, https://doi.org/10.1007/3-540-49430-8_2, 1998.

Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D. C., Wilkinson, J., Phillips, T., Byrne, J., Tietsche, S., Sarojini, B. B., Blanchard-Wrigglesworth, E., Aksenov, Y., Downie, R., and Shuckburgh, E.: Seasonal arctic sea ice forecasting with probabilistic deep learning, Nat. Commun., 12, 5124, https://doi.org/10.1038/s41467-021-25257-4, 2021.

Ernakovich, J. G., Hopping, K. A., Berdanier, A. B., Simpson, R. T., Kachergis, E. J., Steltzer, H., and Wallenstein, M. D.: Predicted responses of arctic and alpine ecosystems to altered seasonality under climate change, Global Change Biol., 20, 3256–3269, https://doi.org/10.1111/gcb.12568, 2014.

Häkkinen, S. and Mellor, G. L.: Modeling the seasonal variability of a coupled arctic ice-ocean system, J. Geophys. Res.: Oceans, 97, 20285–20304, https://doi.org/10.1029/92JC02037, 1992.

8. Line 323-327: This is methodology and should not be within the results part - as it is missing within the methods section.

   Reply: Implemented. This part has been removed from the results section to the methodology section.

9. Line 328-329: As mentioned earlier, the baseline models cannot be newly introduced within the results.

   Reply: Impelemented. We have changed the order of introduction. The baseline models are introduced in the methodology section before the resuls.

10. Line 343-344: You can't conduct boxplots. Do you mean you conducted the model application 10 times?

    Reply: Implemented. We have rephrased the manuscript to improve its clarity and readability here We trained and evaluated each model in 10 independent runs. This repeated training quantifies performance variability due to the inherent stochasticity of the optimization process. Results from the 10 runs are summarized using boxplots.

11. Line 357: Figure 6 y axis seems to be cut, the numbers are partly missing
    Reply: Implemented. We thank the reviewer for pointing this out. The figures have been fixed.

12. Line 361: I dont see how this represents the "spectrum of hydrological variability". From my understanding, it is more of a possibility to see, how the model performs if the data is only available in lesser resolution. How does this assess the depiction of the hydrological variability?
    Reply: Implemented. We thank the reviewer for this important clarification. The reviewer is correct that our analysis examines model performance under varying flow conditions, from low to high discharge events. The corresponding description is rephrased for clarification.

13. Line 405: Figure 8, are these for a aggregation period of 1 month?
    Reply: We thank the reviewer for requesting this clarification. The boxplots in Figure 8 show results aggregated across all forecasting time steps. Each model variant is trained 10 times independently at each time step (1-12 months), yielding 120 total evaluations per model. The results of all 120 evaluations for each model are summerized in the boxplots. The manuscript has been revised for clarification.

14. Line 407-415: This is all methodology and not results.
    Reply: Implemented. We thank the reviewer for identifying this issue. The contents have been reorganized and moved to methodology.

15. Line 437-448: I dont think this part is really necessary here. The conclusion is not

a whole summary of the paper, but points out the key findings again.

Reply: Impelemened. This long paragraph has been removed.

16. Line 455-456: The river discharge has a long memory? The sentence does not make sense. I feel like there is a more thorough discussion necessary of why the model shows this behaviour regarding the model efficiency for different aggregation periods - where the reason must be within model structure and how it fits the discharge pattern over time.

    Reply: Implemented. The sencence has been rephrased to avoid confusion. A more thorough discussion will be added here.

    The optimal performance at the 9-month input sequence length reflects important temporal characteristics of this permafrost-dominated watershed and the model's capacity to capture structured temporal dependencies. In the Kolyma River basin, current discharge is influenced by hydrometeorological conditions that could span multiple seasons, such as snow accumulation, snowmelt dynamics, and subsequent baseflow recession controlled by active layer storage and permafrost-restricted groundwater flows. The 9-month optimal input window captures the information of seasonal dynamics which provides the model with sufficient temporal context. The attention mechanism further refines this by assigning higher importance to specific antecedent months that strongly influence current discharge. Shorter sequences may fail to capture full seasonal cycles and snow accumulation processes, while longer sequences (10-12 months) likely introduce temporal uncertainties.

17. I generally feel like the discussion part is lacking depth. While I personally recommend to separate results and discussion, you can keep both together if it makes sense overall. But in the current state, the results lack depth regarding the explanation of observed model behaviour. For example, line 462-463: has this been the same for the application of other models? Is this a common problem? Like this, a few more citations and comparisons to other studies would help

putting the paper within a broader context.

Reply: Implememted. We will improve the results and discussion.

This systematic underestimation of peak flows represents a common challenge in data-driven hydrological modeling, particularly for Arctic river systems, where extreme discharge events are relatively rare but carry significant implications for water resource management and hazard mitigation. Kratzert et al. (2019) observed similar patterns in LSTM-based rainfall-runoff modeling across diverse catchments. For Arctic rivers specifically, Gelfan et al. (2017) and Chang et al. (2025) reported that process-based models and machine learning approaches struggle with extreme conditions due to the complex processes and events that are poorly represented in limited observational records. In our study, extreme high discharge events (>80 mm) constitute less than 5% of the training dataset, creating a class imbalance problem common in hydrological time series (Nearing et al., 2021). The squared error loss function (MSE) used in model training inherently weights all samples equally, which can lead to optimization that favors the more numerous moderate flow events at the expense of rare extremes. Future work could address this limitation through specialized sampling techniques or physics-informed constraints specifically designed to better captures high-magnitude discharge events.

Reference:

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Gelfan, A., Gustafsson, D., Motovilov, Y., Arheimer, B., Kalugin, A., Krylenko, I., and Lavrenov, A.: Climate change impact on the water regime of two great arctic rivers: modeling and uncertainty issues, Clim. Change, 141, 499–515, https://doi.org/10.1007/s10584-016-1710-5, 2017.

Chang, S. Y., Schwenk, J., and Solander, K. C.: Deep learning advances arctic

river water temperature predictions, Water Resour. Res., 61, e2024WR039053, https://doi.org/10.1029/2024WR039053, 2025.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

18. Also, I am currently missing a graphical depiction of the gauging curve and the simulated discharge. I believe a figure for that would help to give the reader an idea of how the model behaves, where it might deviate from gauging data and where it is strongly in congruence with it.

    Reply: Implememted. A new graphic depiction of observed and simulated discharge will be added to the manuscript to provide the readers with a better idea of how different models behave.

Minor comments:

19. Line 22: structure

    Reply: Implemented. We have corrected the spelling/grammar error.
20. Line 24: dominated by permafrost

    Reply: Implemented. We have corrected the spelling/grammar error.
21. Line 27: ...that these components improve the predictive performance.

    Reply: Implemented. We have corrected the spelling/grammar error.
22. Line 46: These temperature dependent transitions...?

    Reply: Implemented. We have corrected the spelling/grammar error.
23. Line 128-129: Why is there no citation for the Dataset?

    Reply: Implemented. The data source and citation have been added to the manuscript.
24. Line 178: 1) Input expansion

    Reply: Implemented. We have corrected the spelling/grammar error.

25. Line 183-185: Kolmogorov-Arnold theorem while avoiding the computational overhead

    Reply: Implemented. We have corrected the spelling/grammar error.

26. Line 195: GELU

    Reply: Implemented. We have corrected the spelling/grammar error.

27. Line 196: Figure 3 not referenced within the text.

    Reply: Implemented. We have added Figure 3 in the text.

28. Line 200: ...mechanism and a hidden state, an LSTM can efficiently regulate...

    Reply: Implemented. We have corrected the spelling/grammar error.

29. Line 209: The memory cell of an LSTM is primarily composed...

    Reply: Implemented. We have corrected the spelling/grammar error.

30. Line 240: "Q refers the discharge prediction using the context vector calculated from the context vector." It has to be "refers to" and what is "using the context vector calculated from the context vector" supposed to mean?

    Reply: Implemented. It has been rephrased to improve clarity. We have corrected the spelling/grammar error.

31. Line 273: I recommend a semicolon after water.

    Reply: Implemented. We have corrected the spelling/grammar error.

32. Line 279: caused by sources, such as model simplifications...

    Reply: Implemented. We have corrected the spelling/grammar error.

33. Line 285-286: Maybe its better to reformulate the sentence and describe alpha and beta as parameters that have to be fitted through model application?

    Reply: Implemented. α and β are weighting coefficients that control the relative importance of the data-driven loss (MSE) and physics-informed constraint terms in the combined loss function. The optimal physics constraint weight ($\beta = 0.3$) and the MSE weight ($\alpha = 0.7$) are adopted by conducting grid search over $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

34. Line 299: beneficial

    Reply: Implemented. We have corrected the spelling/grammar error.

35. Line 303-304: What is cited here? The Nash-Sutcliffe efficiency measure should

be properly cited.

Reply: Implemented. A reference has been added regarding the Nash-Sutcliffe efficiency measure.

36. Line 330: I would recommend to implement the name RCPIKLA of the new model earlier, instead of within the results.

    Reply: Implemented. We have move it earlier.

37. Line 396: change "better captures"

    Reply: Implemented. We have corrected the spelling/grammar error.

Sincerely yours,

Renjie Zhou

Renjie Zhou
Associate Professor
Department of Environmental and Geosciences
Sam Houston State University
Huntsville, TX 77340