Sensitivity of hydrological machine learning prediction accuracy to information quantity and quality

RC1.1: The manuscript entitled "Sensitivity of hydrological machine learning prediction accuracy to information quantity and quality" present a valuable discussion about the influence of information quantity and quality on the performance of machine-learning-based (ML) models for hydrological prediction.

Response to RC1.1: Thank you for recognizing the value of this study.

RC1.2: Below are some points regarding its methodology, results, and potential areas for improvement:

It is quite trivial that calibrated models can offer training samples with high quality and thus help machine learning models achieve significant performance improvement. Could you please further clarify which key scientific findings/insights can be offered by this study?

Response to RC1.2: We appreciate your feedback and understand your point. Indeed, it is reasonable to expect that calibrated models provide higher-quality training data compared to uncalibrated models. However, our question is: what specifically makes the outputs of a calibrated model high-quality training data? In other words, if calibrated model outputs do improve machine learning model accuracy, why is that the case? The aim of this study is not simply to recommend using calibrated mechanistic model outputs to enhance accuracy. Instead, we seek to understand how to improve the accuracy of hydrological machine learning models efficiently by exploring the underlying qualities of training data that contribute to this improvement. In this study, we examined the relationship between the quantity and quality of information in training datasets and the prediction accuracy of hydrological machine learning models. Our assumption is that training sample quantity and quality can be quantified using information theory measures, specifically marginal entropy for quantity and transfer entropy for quality.

To clarify our approach, we have added our research questions and hypotheses to the article (after the sentence ending in Line 69 on Page 3), providing readers with a clear understanding of the core objectives of this study: “The research question that this study tried to answer was how the quantity and quality of information in training datasets, as measured by marginal entropy and transfer entropy, can affect the prediction accuracy of hydrological machine learning models? Our hypothesis was that a higher quantity and quality of information in training datasets, as indicated by increased marginal entropy and transfer entropy, would positively correlate with improved prediction accuracy in these models.”

RC1.3: Figure 1. classifies Random Forest (RF), Support Vector Machine (SVM) as clustering methods, Artificial Neural Network (ANN) as neural network method. What are the essential differences between the two categories of ML models and whether such differences will influence the following discussion?

Response to RC1.3: The essential differences between clustering methods and neural network methods in machine learning are rooted in their objectives, underlying mechanisms, and applications. The clustering methods are unsupervised learning techniques that aim to group data points into clusters based on similarity or distance measures without prior knowledge of labels. They are primarily used for data exploration, identifying inherent patterns, and segmenting data into meaningful clusters. Neural networks are typically used in supervised or reinforcement learning contexts, aiming to learn complex patterns in labeled data for prediction, classification, or decision-making tasks. They are known to be versatile and well-suited for high-dimensional and non-linear data. These differences have important implications for the choice of machine learning models. We have revised the last paragraph of the discussion section and added another paragraph with a focus on how the distinctions between clustering methods and neural networks influence their respective effectiveness and applications:

"Negative IUE-TE values were observed when watershed responses were predicted using RF and SVM models (red star in Fig. 8[b]), particularly in the WD+UC case, suggesting challenges in leveraging additional information from training data. The RF and SVM models, which rely on "piecewise" linear decision boundaries or hyperplanes to partition input space, struggled to manage the "curse of dimensionality" (Bellman, 1961) and complex non-linear relationships between variables. While SVM models use kernel functions to transform non-linear decision spaces into linear ones, and RF models employ non-linear decision boundaries, prior studies indicate that such methods are not always effective in resolving high-dimensional issues, often sampling less informative features (Wang and Xia, 2016; Ye et al., 2013). Despite the radial basis kernel function and Bayesian optimization employed in this study to enhance SVM performance (Shawe-Taylor and Sun, 2011), the model's predictive accuracy remained inconsistent. Conversely, the ANN model avoided negative IUE scores, demonstrating its resilience and ability to more efficiently utilize quality information, even with lower-quality training data in cases such as WD+UC (Table S3).

Neural networks, particularly the ANN model, excel in handling high-dimensional, non-linear data, making them more effective than RF and SVM for this study's hydrological predictions. With diverse features such as precipitation, temperature, and watershed characteristics contributing to accurate predictions, the ANN model utilized the rich, high-dimensional data from calibrated and uncalibrated SWAT outputs to achieve strong performance. Unlike clustering methods, which primarily group data without a predictive function, neural networks improve prediction accuracy through learning from labeled data and adapting to input quality. The

absence of negative IUE scores for ANN underscores its flexibility and robustness. These findings affirm the ANN model's suitability for high-dimensional, quality-driven hydrological modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity."

RC1.4: For Sect. 2.2, the input variables of machine learning models are not clear. It might need further explanation about the setting-up process of machine learning models.

Response to RC1.4: We appreciate your observation regarding the insufficient explanation of the machine learning model setup process. To address this, we have provided a more detailed description of the input variables and clarified the dataset division process in the revised manuscript. We have also elaborated on the optimization procedures for each model, which were conducted using Bayesian optimization to ensure efficient and accurate parameter tuning.

In the sub-section of "2. Method and Materials: 2.2 Data-driven (or machine learning) models" of the revised manuscript, we state: "The optimization of three machine learning models—RF, SVM, and ANN—was carried out using Bayesian optimization, a method that improves decision-making efficiency by iteratively identifying the most promising hyperparameter configurations (Jones, 2001). Compared to traditional grid or random search methods, Bayesian optimization is notably more efficient in finding optimal hyperparameters (Yu and Zhu, 2020). For the RF model, key parameters such as the maximum number of splits, the number of predictors per split, and the number of trees were optimized. In the case of the SVM model, the kernel scale, epsilon, and cost parameters were fine-tuned. For the ANN model, optimization focused on activation functions and layer sizes. These optimizations were designed to enhance each model's performance by leveraging input variables—including precipitation, temperature, and watershed characteristics—that were carefully selected to align with the study's objectives."

References

Jones, D.R., 2001. A Taxonomy of Global Optimization Methods Based on Response Surfaces. Journal of global optimization, 21, pp.345-383.

Yu, T., Zhu, H., 2020. Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689.

RC1.5: Line 151: why is the threshold correlation arbitrarily selected as 0.30?

Response to RC1.5: The interpretation of a correlation coefficient of 0.30 varies depending on the context and field. While some define 0.30 as the threshold for a "medium" correlation (Woolf, 2009), others describe it as a low correlation (Asuero et al., 2006) or weak correlation (Schober and Schwarte, 2018). Low or weak correlations suggest some degree of structure or association, as opposed to no meaningful correlation when values fall below this threshold. Since the correlation coefficient measures linear agreement between two variables (e.g., observed vs.

predicted), achieving high linear correlations in hydrological modeling is uncommon due to the inherently complex and nonlinear nature of hydrological processes, especially on relatively fine temporal scales such as daily. Based on the literature, we determined that a correlation value of at least 0.30 represents the minimum acceptable strength of correlation between observed and predicted hydrological variables for this study.

References

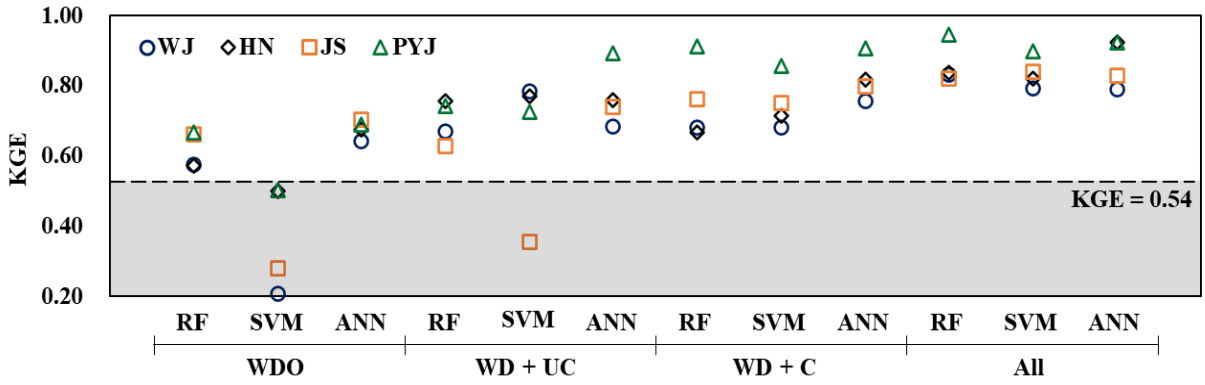
Woolf, P. J., 2009. Chemical Process Dynamics and Controls. United States, University of Michigan Engineering Controls Group. pp.13.13.1. (https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Chemical_Process_Dynamics_and_Controls_(Woolf)/13%3A_Statistics_and_Probability_Background/13.13%3A_Correlation_and_Mutual_Information)

Asuero, A.G., Sayago, A. and González, A.G., 2006. The correlation coefficient: An overview. Critical reviews in analytical chemistry, 36(1), pp.41-59.

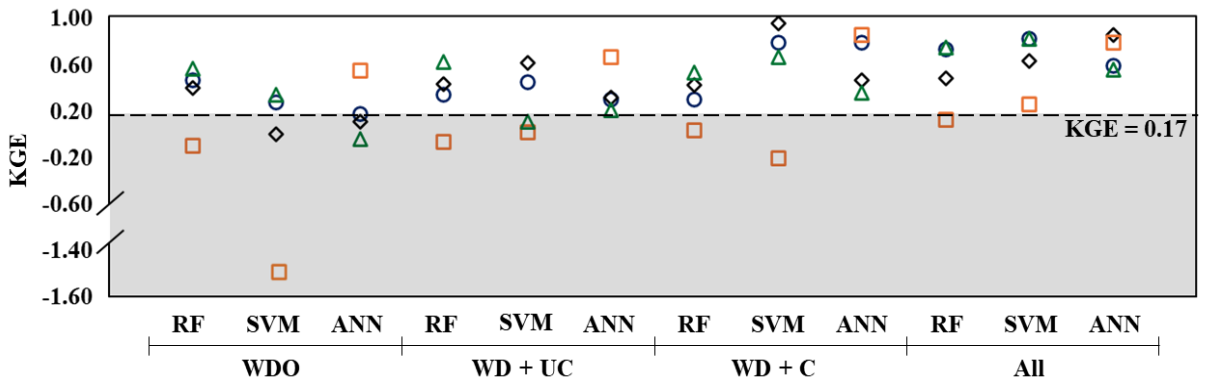
Schober, P., Boer, C. and Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. Anesthesia & analgesia, 126(5), pp.1763-1768.

RC1.6: Figure 4. uses 3D plotting which might make comparison between different cases and models difficult. Could you please use a 2D figure with legends instead?

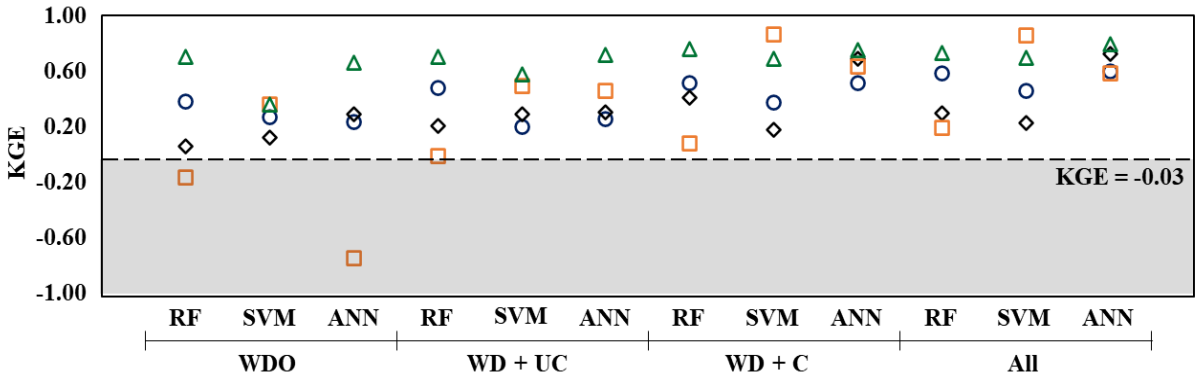
Response to RC1.6: Thank you for your valuable feedback; we understand your concerns and have made the necessary adjustments. The figures have been revised to a 2D format for improved clarity, with each watershed represented by a distinct symbol. Additionally, the training dataset is now plotted along the x-axis for better visualization.



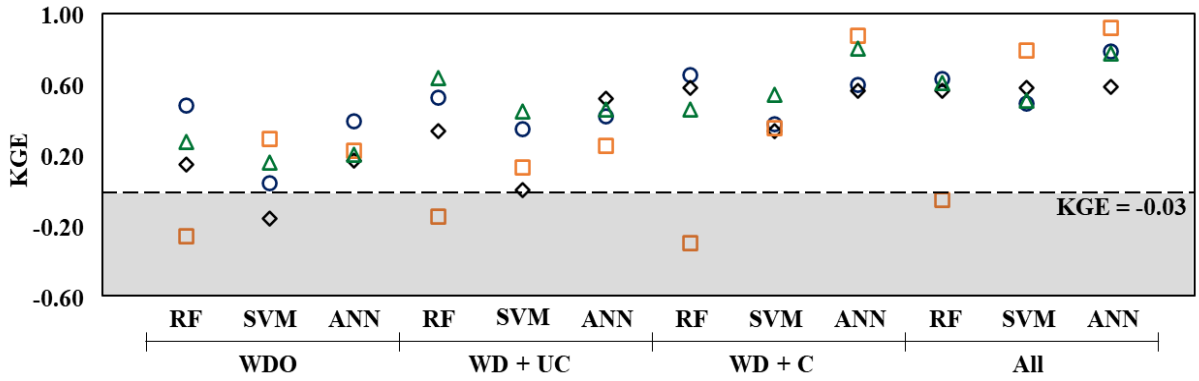
(a) Flow



(b) SS



(c) TN



(d) TP

Figure 4. Prediction accuracy (KGE) of hydrological ML models trained with the different training data set combinations. The KGE values that do not satisfy the acceptable accuracy levels (e.g., i.e., 0.54 for flow, 0.17 for SS, and -0.03 for TN/TP) are included in gray areas.