Revision Note

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

Editor

EC1: Thank you for your submission and thorough responses. The study addresses an important and timely topic, and both reviewers recognize its potential contribution to hydrological ML modeling.

However, key concerns remain regarding clarity, terminology, methodological transparency, and presentation of uncertainty. Your initial revisions have improved the manuscript, particularly the restructured Discussion and clarified definitions, but further refinement is needed to meet publication standards.

I invite you to submit a major revision that fully addresses the reviewers' comments.

Response to EC1: Thank you for your positive and constructive feedback on our study and manuscript. In response to the reviewer's comments on the structure, terminology, methodological clarity, and uncertainty presentation, we revised the manuscript to improve clarity and consistency. The Discussion section was reorganized into three focused subsections addressing (i) the influence of target variables and watershed characteristics on ML performance, (ii) the effects of ML model structures under varying information quantity and quality, and (iii) implications, limitations, and future directions for integrating ML with process-based models, with detailed performance metrics moved to the Results section. Terminology was refined by replacing "resilience" with "robustness" and by explicitly defining "information-use efficiency" as achieving equal or better predictive performance with fewer, more informative inputs, rather than reduced computational cost. We also clarified the computation of marginal and transfer entropy, with transfer entropy calculated based on Shannon entropy and quantile-based discretization, a zero-lag setting to reflect synchronous information-use in regression models, and all entropy values reported in bits. To prevent data leakage, ML training and testing periods were explicitly aligned with SWAT calibration and validation periods. Finally, model performance uncertainty was clarified by revising the main figure and adding a supplementary figure with compact uncertainty indicators.

Reviewer 1

RC1.1: The manuscript investigates how the input information quantity and quality, quantified by marginal and transfer entropy, influence the machine-learning-based (ML) hydrological prediction performance. The results demonstrate that increased information quantity does not necessarily enhance model performance whereas improved information quality can more efficiently boost predictive accuracy. However, some points might need to be improved or clarified before publication.

Response to RC1.1: We thank the reviewer for recognizing this study's central contributions and for highlighting areas needing clarification. In response to the comments regarding structure, terminology, and interpretation in the discussion, we have made several key revisions to improve clarity and emphasize the study's contributions.

First, as suggested, we split the original Discussion section into three clearer parts: (1) how target variables and watershed characteristics affect ML performance, (2) how ML model structures affect predictive performance across information quantity and quality, and (3) implications, limitations, and future directions for combining ML with process-based models. We also moved detailed model results (e.g., RMSE and NSE) to the Results section to clearly separate evidence from interpretation.

Second, we replaced the word "resilience" with "robustness" when describing the ANN model's ability to maintain high performance even when low-quality inputs (i.e., WD+UC) were utilized.

Third, we clarified our use of the term "efficiency" to avoid confusion. In this study, "information-use efficiency" refers to a model's ability to achieve equal or better predictive performance using fewer, but more informative, input variables, not to reductions in training time or computational cost. We revised the text to reflect this meaning consistently.

RC1.2: For the discussion part, the authors dump the discussion of ML modeling accuracy, the influence of data quantity and quality into one long section. It might be better to transfer the results of ML performance into the Results part, and then divide remaining discussions into several subsections, e.g., the influence of data quantity on ML performance, the influence of data quality on ML performance, implications on future integration of ML with mechanistic models. This breakdown might better clarify the contribution of this work.

Response to RC1.2: Thank you for this constructive suggestion. We agree that the original Discussion section combined several key themes, including ML model performance, the effects of data quantity and quality, and broader implications, into a single continuous narrative, which may have reduced clarity and obscured the study's main contributions. In response, we have made the following revisions:

1. Relocated quantitative results (e.g., RMSE and NSE comparisons across data scenarios) from the Discussion to the Results section to more clearly separate evidence from interpretation.
2. Reorganized the Discussion section into three focused sub-sections:
    - 4.1 Influence of target variables and watershed characteristics on ML performance
    - 4.2 Influence of model structure on ML performance across information quantity and quality
    - 4.3 Implications, limitations, and future directions for integrating ML with mechanistic models

These changes better align our discussion structure with the study's main objectives and improve the clarity of our contributions. Specifically:

Section 4.1 discusses how differences in information content across target variables and watersheds influence prediction performance.

Section 4.2 focuses on how differences in model structure among ML methods affect performance, with particular emphasis on why RF and SVM experience performance degradation when handling high-dimensional input data.

Section 4.3 explores the broader implications and limitations of our findings, including how entropy metrics can guide input selection and data assimilation in mechanistic models, inform efficient training strategies, and support the development of hybrid modeling approaches.

“4.1 Influence of target variables and watershed characteristics on ML performance

The prediction accuracy of the ML models varied according to the variables of interest (flow, SS, TN, and TP). The RF, SVM, and ANN ML models were best at predicting the flow of the study watersheds compared to the other variables (Figs. 4 and S5). For example, the ML models provided KGEs of 0.557 (WDO) to 0.854 (All) when predicting flow versus KGEs of 0.093 (WDO) to 0.607 (All) for SS loads. This variance is presumably because of the previously described differences in the amount of information contained in the watershed responses or target variables. For example, the flow hydrographs commonly have relatively higher marginal entropies (8.582 on average) than the other variables’ hydrographs (5.144 for SS, 6.180 for TN, and 5.981 for TP on average) for all study watersheds (Table 3). In the frequency domain, normalized SS load data have the most biased distributions toward low values (small SS loads) and the highest frequencies among the watershed responses or target variables, leading to relatively low entropy in the SS data (Fig. S11). The SS, TN, and TP concentrations observed at the watershed outlets have relatively small variations compared to the flow (Table S4), which might be attributed to the fact that water quality variables were much less frequently measured (or sampled) than flow (Table S4, the number of observations); thus, potentially large concentration variations might not be apparent in the observations. These comparison results

imply that the frequency of water quality sampling can affect the amount of information in training data and the accuracy of hydrological ML model prediction.

The accuracy of the ML modeling also varied by the watersheds. Regardless of the training data sets, the ML models provided the best prediction accuracy for the PYJ watershed, which has the largest drainage area, while they did the worst for the JS watershed, which has the smallest drainage area; this implies the potential impact of watershed features and responses (flow, SS, TN, and TP) on ML prediction accuracy. For example, entropy in the watershed responses of the PYJ watershed was consistently higher than that of the JS watershed (Table 3). In the case of WDO, the amount of information contained in the independent variables (i.e., only weather records observed at a single station) of the training data set should be the same for the PYJ and JS watersheds. However, their responses (dependent or target variables) differ and thus have different entropy (or information) scores. The responses of the PYJ watershed are spread out over wide value ranges, which means relatively high entropies compared to those of the other watersheds, especially the JS watershed (Fig. S11). The flow observed at the outlet of the JS watershed are relatively highly biases toward low flow ranges. ML model prediction accuracy was found to be associated with the entropy in the watershed responses (Fig. 9). The results indicated that the KGE (i.e., prediction accuracy) scores of the ML models generally increased with increases in the amount of information contained in the target variables (Fig. 9).

4.2 Influence of model structure on ML performance across information quantity and quality

Despite being trained on identical datasets, ML models exhibited different predictive abilities in response to increases in information quantity and quality. The ANN and SVM models could improve their predictions more efficiently in terms of the amount of information (i.e., marginal entropy) added to the training data compared to the RF model (Fig. 8(c)). The RF model uses a random sampling method to select the feature subspace for each node in growing the trees (Breiman, 2001), which is a model parameter called the “number of variables.” Previous studies (Wang and Xia., 2016; Ye et al., 2013) have argued that when applying the random sampling method to a high-dimension data set, model may select many subspaces that do not include informative features and will increase error bounds for the RF model. This study agrees with previous studies: RF performed relatively poorly when dimension of a training data set was higher (i.e., the large number of independent variables) than SVM and ANN. A traditional ANN model with one or two hidden layers is known to suffer performance degradation due to its rapid growth in the number of connection weights (Krenker et al., 2011). However, one study demonstrated that a deep neural network that employs numerous hidden layers, such as the one used in this study, could yield promising performance with high-dimensional training data (Liu et al., 2017).

In addition, 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. While SVM models use kernel function 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-dimensionality 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 robustness by effectively exploiting additional information even when lower-quality data (WD+UC) were added to the training dataset (Figs. 8(b) and (e)). This indicates that the ANN model excels at handling high-dimensional, non-linear data, making it 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 labelled 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 hydrological modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity.

4.3 Implications, limitations, and future directions for integrating ML with mechanistic models

The quantitative evaluation of data quality and quantity in relation to model performance provides actionable insights for future modeling efforts. Our findings highlight the importance of prioritizing high-quality, high-relevance inputs to improve prediction accuracy. This suggests feature selection techniques, informed by our results, could help identify the most impactful variables, reducing the risk of overfitting and computational inefficiencies. The demonstrated sensitivity of ML models to data quality suggests the need for rigorous preprocessing steps, such as outlier detection, imputation of missing data, and validation of sensor measurements. For example, ensuring consistent and accurate data collection at critical watershed locations can enhance model reliability. Our results could guide the development of benchmarks or thresholds for data quality metrics (e.g., measurement error limits) to ensure datasets meet the minimum requirements for effective modeling.

Our findings also have implications for hybrid modeling approaches, particularly guiding the integration of ML with mechanistic models. These results can help select variables where mechanistic models provide better physical realism (e.g., water and nutrient transport) while leveraging ML for complementary predictions (e.g., extreme event responses or data gap interpolation). Improving the quality of key inputs can also help reconcile discrepancies between data-driven and theory-driven predictions, enhancing overall model accuracy. Expanding beyond

standalone ML applications, this study provides a foundation for adaptive modeling frameworks that dynamically assess and adjust data inputs based on their predictive contributions.

Future studies could build on our quantitative evaluation by applying these insights to develop guidelines or workflows for selecting and preparing datasets tailored to specific hydrological and water quality objectives. The study underscores the need to refine hydrological ML prediction models by emphasizing the connection between training data quantity, quality, and prediction accuracy. These results suggest that higher-quality training data improved the IUE of ML models, enabling them to maintain or improve prediction accuracy while using a reduced number of inputs. The integration of theory-driven and data-driven approaches will not only improve prediction accuracy but also streamline model training by ensuring that the data used contains both sufficient quantity and quality. Moreover, a deeper understanding of the interaction between different data types will inform more effective training strategies for ML models, leading to more accurate and reliable hydrological predictions.

While this study provides valuable insights into the relationship between ML prediction accuracy and the quality and quantity of input data in hydrological modeling, several limitations and uncertainties must be acknowledged. The study's findings are based on a specific set of watersheds with unique hydrological, climatic, and geographical characteristics. The variability in watershed conditions across different regions may limit the generalizability of the results. The study assumes that the available datasets accurately represent real-world hydrological processes. However, biases, errors, and inconsistencies in the input data, such as measurement inaccuracies or missing values, could influence the results. The study evaluates model performance at specific temporal and spatial scales. Fine temporal resolutions (e.g., daily predictions) may introduce additional complexities not captured in coarser scales (e.g., monthly or annual). Despite optimization efforts, ML models remain susceptible to overfitting, particularly when trained on small datasets or when irrelevant features are included. By acknowledging these limitations and uncertainties, this study provides a helpful starting point for future work to build upon, with the goal of enhancing the reliability of ML models in hydrological applications."

RC1.3: Line 526. The authors mentions that the ANN model exhibit its resilience to more efficiently utilize quality information. What does the term of "resilience" mean here?

Response to RC1.3: We appreciate the reviewer's request for clarification. In the original sentence, we wrote: "The ANN model exhibits its resilience to more efficiently utilize quality information …" Our intention was to convey that the ANN model retained a high level of predictive skill even after low-entropy (low-quality) inputs were added, whereas the other models showed a sharper decline in performance. In other words, the ANN model was robust to reductions in the quantity of information as long as the remaining inputs were of high informational quality. To avoid ambiguity, we have replaced the word "resilience" with "robustness," a term more commonly used in the modelling literature to describe stability of

performance under adverse or reduced-information conditions. In addition, we have added a definition in the text.

The revised paragraph now reads: “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 robustness by effectively exploiting additional information even when lower-quality data (WD+UC) were added to the training dataset (Figs. 8(b) and (e)). This indicates that the ANN model excels at handling high-dimensional, non-linear data, making it 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 labelled 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 hydrological modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity.”

RC1.4: Line 555-557. The authors mentions that high-quality training data can improve the efficiency of ML models. The term of “efficiency” might be ambiguous here since it can either refer to the information use efficiency or reduced training/computation time of ML models, especially given later comments on potential advantages of streamlined model training. Similar unclarified issues also exist in other parts, e.g., Line 539-540, and might cause reader’s misunderstanding. Please check related unclarified terms and keep consistency through discussions.

Response to RC1.4: Thank you for your comment regarding the ambiguous use of the term “efficiency.” We agree that in the original text, the use of “efficiency” could be misinterpreted as either referring to information-use efficiency or to computational/training efficiency. Our intention was to emphasize that high-quality training data enhanced the information-use efficiency of machine learning models, meaning the models achieved equal or better predictive accuracy using fewer, yet more informative, input variables. To clarify this, we revised the sentence in Lines 523–527 to state: “These results suggest that higher-quality training data improved the IUE of ML models, enabling them to maintain or improve prediction accuracy while using a reduced number of inputs.” We also reviewed and revised other instances throughout the manuscript, such as on Lines 506–508, to ensure that the term “efficiency” consistently refers to information-use efficiency unless otherwise specified: “Our findings highlight the importance of prioritizing high-quality, high-relevance inputs to improve prediction

accuracy. This suggests feature selection techniques, informed by our results, could help identify the most impactful variables, reducing the risk of overfitting and computational inefficiencies."

Reviewer 2

RC2.1: This is a useful, well-motivated study on how information quantity (marginal entropy) and quality (transfer entropy) affect hydrological ML performance. The core message—that more data does not guarantee better predictions while higher-quality information is more impactful—is clear and relevant. I recommend moderate revision focused on clarity of structure, terminology, and methods transparency. Detailed comments are listed as follows.

Response to RC2.1: We appreciate the reviewer's detailed and constructive suggestions for clarifying the manuscript structure and improving methodological transparency. In response to them, we have made several key revisions. First, as suggested, we split the original Discussion section into three parts: (1) how target variables and watershed characteristics affect ML performance, (2) how ML model structures affect predictive performance across information quantity and quality, and (3) implications, limitations, and future directions for combining ML with process-based models. We also moved detailed model accuracy statistics (e.g., RMSE and NSE) included in the Discussion section to the Results section to better separate evidence from interpretation. Second, we replaced the word "resilience" with "robustness" when describing the ANN model's ability. Third, we clarified our use of the term "efficiency" to avoid confusion. In this study, "information-use efficiency" refers to a model's ability to achieve equal or better predictive performance using fewer, but more informative, input variables, not to reductions in training time or computational cost. Fourth, we added whisker-box plots to present the inter-dataset and inter-model variability of the key metrics (KGE and IUE). These plots allow us to examine how performance varies across different ML models and data combinations and to discuss which configurations yield more consistent results.

RC2.2: Restructure: move quantitative ML results to Results; keep Discussion for interpretation, organized into quantity effects, quality effects, and implications for ML–process model integration.

Response to RC2.2: Thank you for this constructive suggestion. We agree that the original Discussion section combined several key themes, including ML model performance, the effects of data quantity and quality, and broader implications, into a single continuous narrative, which may have reduced clarity and obscured the study's main contributions. In response, we have made the following revisions:

1. Relocated quantitative results (e.g., RMSE and NSE comparisons across data scenarios) from the Discussion to the Results section to more clearly separate evidence from interpretation.
2. Reorganized the Discussion section into three focused sub-sections:
    - 4.1 Influence of target variables and watershed characteristics on ML performance
    - 4.2 Influence of model structure on ML performance across information quantity and quality

- 4.3 Implications, limitations, and future directions for integrating ML with mechanistic models

These changes better align our discussion structure with the study's main objectives and improve the clarity of our contributions. Specifically:

Section 4.1 discusses how differences in information content across target variables and watersheds influence prediction performance.

Section 4.2 focuses on how differences in model structure among ML methods affect performance, with particular emphasis on why RF and SVM experience performance degradation when handling high-dimensional input data.

Section 4.3 explores the broader implications and limitations of our findings, including how entropy metrics can guide input selection and data assimilation in mechanistic models, inform efficient training strategies, and support the development of hybrid modeling approaches.

“4.1 Influence of target variables and watershed characteristics on ML performance

The prediction accuracy of the ML models varied according to the variables of interest (flow, SS, TN, and TP). The RF, SVM, and ANN ML models were best at predicting the flow of the study watersheds compared to the other variables (Figs. 4 and S5). For example, the ML models provided KGEs of 0.557 (WDO) to 0.854 (All) when predicting flow versus KGEs of 0.093 (WDO) to 0.607 (All) for SS loads. This variance is presumably because of the previously described differences in the amount of information contained in the watershed responses or target variables. For example, the flow hydrographs commonly have relatively higher marginal entropies (8.582 on average) than the other variables’ hydrographs (5.144 for SS, 6.180 for TN, and 5.981 for TP on average) for all study watersheds (Table 3). In the frequency domain, normalized SS load data have the most biased distributions toward low values (small SS loads) and the highest frequencies among the watershed responses or target variables, leading to relatively low entropy in the SS data (Fig. S11). The SS, TN, and TP concentrations observed at the watershed outlets have relatively small variations compared to the flow (Table S4), which might be attributed to the fact that water quality variables were much less frequently measured (or sampled) than flow (Table S4, the number of observations); thus, potentially large concentration variations might not be apparent in the observations. These comparison results imply that the frequency of water quality sampling can affect the amount of information in training data and the accuracy of hydrological ML model prediction.

The accuracy of the ML modeling also varied by the watersheds. Regardless of the training data sets, the ML models provided the best prediction accuracy for the PYJ watershed, which has the largest drainage area, while they did the worst for the JS watershed, which has the smallest drainage area; this implies the potential impact of watershed features and responses (flow, SS,

TN, and TP) on ML prediction accuracy. For example, entropy in the watershed responses of the PYJ watershed was consistently higher than that of the JS watershed (Table 3). In the case of WDO, the amount of information contained in the independent variables (i.e., only weather records observed at a single station) of the training data set should be the same for the PYJ and JS watersheds. However, their responses (dependent or target variables) differ and thus have different entropy (or information) scores. The responses of the PYJ watershed are spread out over wide value ranges, which means relatively high entropies compared to those of the other watersheds, especially the JS watershed (Fig. S11). The flow observed at the outlet of the JS watershed are relatively highly biases toward low flow ranges. ML model prediction accuracy was found to be associated with the entropy in the watershed responses (Fig. 9). The results indicated that the KGE (i.e., prediction accuracy) scores of the ML models generally increased with increases in the amount of information contained in the target variables (Fig. 9).

4.2 Influence of model structure on ML performance across information quantity and quality

Despite being trained on identical datasets, ML models exhibited different predictive abilities in response to increases in information quantity and quality. The ANN and SVM models could improve their predictions more efficiently in terms of the amount of information (i.e., marginal entropy) added to the training data compared to the RF model (Fig. 8(c)). The RF model uses a random sampling method to select the feature subspace for each node in growing the trees (Breiman, 2001), which is a model parameter called the "number of variables." Previous studies (Wang and Xia., 2016; Ye et al., 2013) have argued that when applying the random sampling method to a high-dimension data set, model may select many subspaces that do not include informative features and will increase error bounds for the RF model. This study agrees with previous studies: RF performed relatively poorly when dimension of a training data set was higher (i.e., the large number of independent variables) than SVM and ANN. A traditional ANN model with one or two hidden layers is known to suffer performance degradation due to its rapid growth in the number of connection weights (Krenker et al., 2011). However, one study demonstrated that a deep neural network that employs numerous hidden layers, such as the one used in this study, could yield promising performance with high-dimensional training data (Liu et al., 2017).

In addition, 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. While SVM models use kernel function 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-dimensionality 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 robustness by effectively exploiting additional information even when lower-quality data (WD+UC) were added to the training dataset (Figs. 8(b) and (e)). This indicates that the ANN model excels at handling high-dimensional, non-linear data, making it 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 labelled 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 hydrological modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity.

4.3 Implications, limitations, and future directions for integrating ML with mechanistic models

The quantitative evaluation of data quality and quantity in relation to model performance provides actionable insights for future modeling efforts. Our findings highlight the importance of prioritizing high-quality, high-relevance inputs to improve prediction accuracy. This suggests feature selection techniques, informed by our results, could help identify the most impactful variables, reducing the risk of overfitting and computational inefficiencies. The demonstrated sensitivity of ML models to data quality suggests the need for rigorous preprocessing steps, such as outlier detection, imputation of missing data, and validation of sensor measurements. For example, ensuring consistent and accurate data collection at critical watershed locations can enhance model reliability. Our results could guide the development of benchmarks or thresholds for data quality metrics (e.g., measurement error limits) to ensure datasets meet the minimum requirements for effective modeling.

Our findings also have implications for hybrid modeling approaches, particularly guiding the integration of ML with mechanistic models. These results can help select variables where mechanistic models provide better physical realism (e.g., water and nutrient transport) while leveraging ML for complementary predictions (e.g., extreme event responses or data gap interpolation). Improving the quality of key inputs can also help reconcile discrepancies between data-driven and theory-driven predictions, enhancing overall model accuracy. Expanding beyond standalone ML applications, this study provides a foundation for adaptive modeling frameworks that dynamically assess and adjust data inputs based on their predictive contributions.

Future studies could build on our quantitative evaluation by applying these insights to develop guidelines or workflows for selecting and preparing datasets tailored to specific hydrological and water quality objectives. The study underscores the need to refine hydrological ML prediction models by emphasizing the connection between training data quantity, quality, and prediction

accuracy. These results suggest that higher-quality training data improved the IUE of ML models, enabling them to maintain or improve prediction accuracy while using a reduced number of inputs. The integration of theory-driven and data-driven approaches will not only improve prediction accuracy but also streamline model training by ensuring that the data used contains both sufficient quantity and quality. Moreover, a deeper understanding of the interaction between different data types will inform more effective training strategies for ML models, leading to more accurate and reliable hydrological predictions.

While this study provides valuable insights into the relationship between ML prediction accuracy and the quality and quantity of input data in hydrological modeling, several limitations and uncertainties must be acknowledged. The study's findings are based on a specific set of watersheds with unique hydrological, climatic, and geographical characteristics. The variability in watershed conditions across different regions may limit the generalizability of the results. The study assumes that the available datasets accurately represent real-world hydrological processes. However, biases, errors, and inconsistencies in the input data, such as measurement inaccuracies or missing values, could influence the results. The study evaluates model performance at specific temporal and spatial scales. Fine temporal resolutions (e.g., daily predictions) may introduce additional complexities not captured in coarser scales (e.g., monthly or annual). Despite optimization efforts, ML models remain susceptible to overfitting, particularly when trained on small datasets or when irrelevant features are included. By acknowledging these limitations and uncertainties, this study provides a helpful starting point for future work to build upon, with the goal of enhancing the reliability of ML models in hydrological applications."

RC2.3: Clarify terms: replace/define "resilience" precisely; reserve "efficiency" for information-use efficiency (IUE) and use "computational efficiency" for runtime/training remarks.

Response to RC2.3: We appreciate the reviewer's request for clarification. In the original sentence, we wrote: "The ANN model exhibits its resilience to more efficiently utilize quality information …" Our intention was to convey that the ANN model retained a high level of predictive skill even after low-entropy (low-quality) inputs were added, whereas the other models showed a sharper decline in performance. In other words, the ANN model was robust to reductions in the quantity of information as long as the remaining inputs were of high informational quality. To avoid ambiguity, we have replaced the word "resilience" with "robustness," a term more commonly used in the modelling literature to describe stability of performance under adverse or reduced-information conditions. In addition, we have added a definition in the text.

The revised paragraph now reads: "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 robustness by effectively exploiting additional

information even when lower-quality data (WD+UC) were added to the training dataset (Figs. 8(b) and (e)). This indicates that the ANN model excels at handling high-dimensional, non-linear data, making it 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 labelled 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 hydrological modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity."

In addition, the use of "efficiency" could be misinterpreted as either referring to information-use efficiency or to computational/training efficiency. Our intention was to emphasize that high-quality training data enhanced the information-use efficiency of machine learning models, meaning the models achieved equal or better predictive accuracy using fewer, yet more informative, input variables. To clarify this, we revised the sentence in Lines 523–527 to state: "These results suggest that higher-quality training data improved the IUE of ML models, enabling them to maintain or improve prediction accuracy while using a reduced number of inputs." We also reviewed and revised other instances throughout the manuscript, such as on Lines 506–508, to ensure that the term "efficiency" consistently refers to information-use efficiency unless otherwise specified: "Our findings highlight the importance of prioritizing high-quality, high-relevance inputs to improve prediction accuracy. This suggests feature selection techniques, informed by our results, could help identify the most impactful variables, reducing the risk of overfitting and computational inefficiencies."

RC2.4: Methods transparency: briefly specify how marginal/transfer entropy are estimated (estimator, lags/embedding/discretization) and note comparability of "bits" across variables.

Response to RC2.4: We agree that a more detailed description of how marginal and transfer entropy were computed would improve reader understanding. Transfer entropy (TE) was calculated using the methodology offered by the RTransferEntropy package (Behrendt et al., 2019), applying Shannon TE with a quantile-based discretization scheme. This choice enhances robustness to outliers and better captures information transfer associated with relatively high and low values (Nie, 2021; Zhang and Zhao, 2022). The lag parameter was set to zero, because the ML models used in this study are standard regression models without explicit temporal memory (e.g., no LSTM); accordingly, we quantified synchronous information-use between inputs and outputs (relationship between same t; lag = 0), which aligns with our primary objective. Marginal entropy was computed with log base 2, and all marginal/transfer entropy magnitudes are reported in bits; where missing, units have been added to figures and tables.

“This study measured the quantity and quality of information contained in the training data using marginal and transfer entropies. In general, a data set that is spread out has relatively high entropy, while another data set that is concentrated on a small range of values has relatively small entropy. The marginal entropy is defined as the information content of a variable and used to calculate randomness in time series using Eq. 3 (Shannon, 1948; Cover and Thomas, 2006; Silva et al., 2017):

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 P(x_i) \quad (3)$$

where $H(X)$ is a measure of information of a discrete random variable X, and $P(x)$ is the probability mass function of variable x in the i^{th} step. The base-2 logarithm is used in the entropy calculation to express information content in bits (Shannon, 1948).

While the amount of information contained in a variable can be calculated using the marginal entropy, we can also calculate the amount of information shared between two variables based on mutual information theory using Eq. 4 (Cover and Thomas, 2006):

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (4)$$

where $I(X,Y)$ is the quantified value between X and Y. The mutual information $I(X,Y)$ represents the expected information gained in Y from measuring X, or vice versa. From these definitions, we can calculate the conditional entropy by subtracting the amount of information shared between X and Y from $H(X)$, which indicates how much information remains about the entire time series X in case we already know the information content of Y.

$$H(X|Y) = H(X) - I(X;Y) \quad (5)$$

These quantities are all symmetrical and do not explain the amount of information exchanged between variables (Bennett et al., 2019). The transfer entropy was devised to consider the asymmetric transfer of information between any two-time series X and Y (the information flow from one to another variable), and can be defined as conditional mutual information (Schreiber, 2000):

$$T_{X\to Y} = I(Y_t; X_t|Y_t) \quad (6)$$

where $T_{X\to Y}$ is the transfer entropy from X to Y, and X_t or Y_t denotes the variables X and Y in time *t*. The lag parameter, which defines the time delay between variables X and Y, was set to zero because the ML models used in this study are standard regression models without explicit temporal memory (e.g., no long-short term memory model); accordingly, we quantified synchronous information-use between inputs and outputs (i.e., the lag time or time delay between X and Y is zero), which aligns with our primary objective.

Transfer entropy was calculated using a quantile-based discretization scheme provided by the RTransferEntropy package (Behrendt et al., 2019) for R software. This method enhances

robustness to outliers and better captures information transfer associated with relatively high and low values (Nie, 2021; Zhang and Zhao, 2022)."

Table 3. Marginal entropy quantified for target variables observed at the watershed outlets in the training period.

Watershed	Flow (bits)	SS (bits)	TN (bits)	TP (bits)
WJ	8.680	5.507	6.080	5.757
HN	8.868	5.252	5.946	5.828
JS	7.896	4.014	5.924	5.760
PYJ	8.884	5.801	6.770	6.578
Average	8.582	5.144	6.180	5.981

References:

Behrendt, S., Dimpfl, T., Peter, F.J., Zimmermann, D.J., 2019. RTransferEntropy — Quantifying information flow between different time series using effective transfer entropy. SoftwareX 10, 100265.

Nie, C.-X., 2021. Dynamics of the price–volume information flow based on surrogate time series. Chaos: An Interdisciplinary Journal of Nonlinear Science 31.

Zhang, N., Zhao, X., 2022. Quantile transfer entropy: Measuring the heterogeneous information transfer of nonlinear time series. Communications in Nonlinear Science and Numerical Simulation 111, 106505.

RC2.5: Data splits & leakage: clearly diagram time windows (SWAT calibration vs. ML train/test) and state how leakage is avoided.

Response to RC2.5: Thank you for the constructive suggestion. The primary objective of this study is to evaluate how the quantity and quality of input data influence the predictive accuracy of ML models by using both intentionally uncalibrated and calibrated mechanistic modeling (SWAT) outputs as inputs.

To further ensure that data leakage was avoided, we carefully aligned the training and testing periods of the ML models with the calibration and validation periods of the SWAT model, respectively. For example, the ML models were evaluated using the same period (January 1, 2016 to December 31, 2017) employed for SWAT model validation. In addition, the ML models were trained exclusively on SWAT-simulated nutrient loads from the calibration period, while

observed discharge and concentration data were used only for SWAT calibration and validation and never as ML inputs, thereby preserving the independence of the datasets. These ensured that no observed data used for SWAT calibration was involved in ML model training or testing, thereby maintaining strict independence between datasets.

To enhance clarity, these methodological safeguards and the rationale behind our data-separation strategy have been explicitly described in the revised manuscript (**Section 2.1 – Overall procedure and Section 2.6 – Study Watersheds and Training Data Acquisition**). In addition, we revised the diagram (Figure 1) to clearly illustrate the data-splitting scheme and the training workflow.

Section 2.1 – Overall procedure (line 95–99): "The SWAT models were calibrated to flow (or streamflow discharges), SS, TN, and TP loads measured at the outlets. Then, the outputs (i.e., flow discharge, SS, TN, and TP loads) of the uncalibrated (i.e., SWAT models with the default parameter values) and calibrated SWAT models were used as additional data sets for the training of the three ML models. Before model training, we carefully separated the ML training and testing datasets to match the SWAT model's calibration and validation periods and to prevent data leakage."

Section 2.6 – Study Watersheds and Training Data Acquisition (line 277–286): "These monitoring data were divided into two non-overlapping subsets: one for model training and the other for testing. To prevent potential data leakage, we applied a consistent temporal split such that the training and testing periods for the ML models matched the calibration and validation periods of the mechanistic model (i.e., SWAT), respectively. In this setup, the three ML models were trained on data corresponding to the SWAT model calibration period (July 12, 2013 to December 31, 2015), while their prediction accuracy of the ML models was evaluated using the testing dataset from the SWAT validation period (January 1, 2016, to December 31, 2017), ensuring a clear separation between the data used for model training and for model evaluation (Fig. 1). To ensure the integrity of the experiment, observed discharge and concentration data were used solely for the calibration and validation of the SWAT model and were not reused in training the ML models. This separation of data sources was a deliberate aspect of the study design to maintain the independence of the ML training process and to avoid any unintended data leakage."

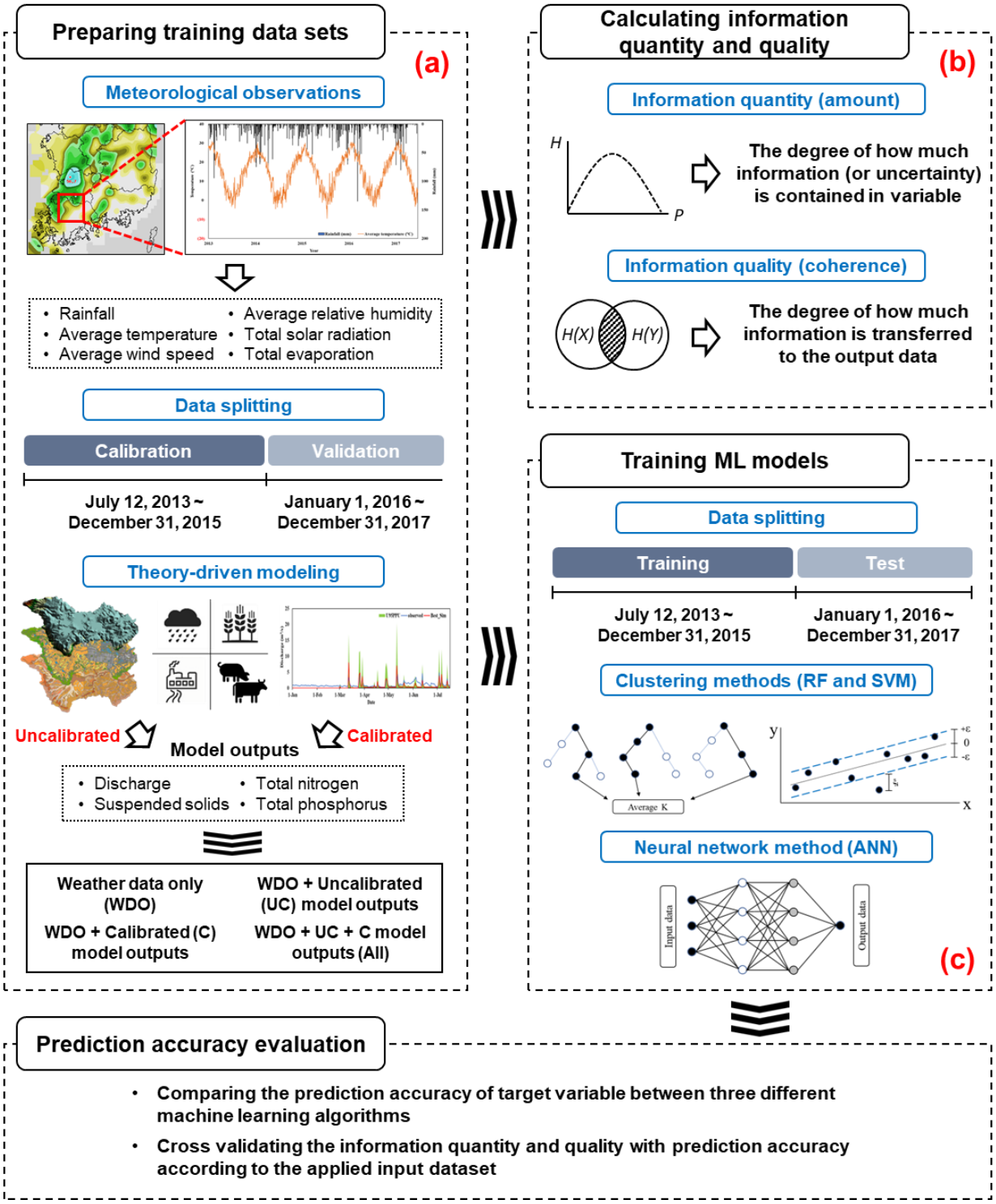

Figure 1. Overall procedure to investigate the contribution of information quantity and quality to the prediction accuracy of hydrological machine learning (ML) modeling.

RC2.6: Uncertainty & presentation: add compact uncertainty cues (e.g., CIs/whiskers or paired tests) to key figures; simplify dense plots and fix minor typos/formatting.

Response to RC2.6: We appreciate the reviewer's suggestion to clarify the presentation of model performance uncertainty. In response to it, we have added new figures (**Figure S5**) and revised the key figure (**Figure 8**) to include compact uncertainty indicators as suggested. Specifically, we added box-whisker plots to present the inter-dataset, inter-model, and inter-watershed variability of the key metrics (IUE-ME and IUE-TE). In addition, we have carefully revised typographical errors and formatting throughout the manuscript.

Section 3.3 – Prediction accuracy of machine learning modeling: "The four ML models were trained with different sets of training data: weather data only (WDO), the uncalibrated SWAT modeling outputs added to WDO (WD+UC), the calibrated SWAT modeling outputs added to WDO (WD+C), and all training data (All or WD+UC+C). The trained ML models yielded unique performances in the predictions depending on the training data set types (Fig. 4). Overall, the ML models' flow prediction accuracy consistently improved as additional data sets were added to the training data, including WDO to WD+UC, WDO+C, and All (Fig. S5(a)). For example, the WDO case provided acceptable accuracy (KGE of 0.67 greater than the threshold of 0.54) in the prediction of flow using the RF algorithm at the outlet of the PYJ watershed. When the outputs of the uncalibrated and/or calibrated SWAT modeling were added to the training data, the accuracy of the ML modeling was increased to KGEs of 0.74 (11.6% increase with WD+UC) and 0.91 (37.2% increase with WD+C) in the case of using the RF model. The additional training data sets also improved the accuracy of the water quality ML modeling. However, ML models trained only using the weather data and uncalibrated mechanistic modeling outputs failed to meet the acceptable accuracy levels (i.e., 0.17 for SS and -0.03 for TN/TP; Fig. 4). In Fig. 4, the KGE scores overall increase from left to right. Negative KGE scores are frequently found in the JS watershed, indicating the models relatively poorly performed for the watershed.

None of the ML models consistently provided more accurate predictions than the others (Figs. 4 and S5(b)). This finding aligns with other studies that identified no ML model that is universally applicable to all data sets or problems (Alzubi et al., 2018; Domingos, 2012). SomeA study has demonstrated that the RF model is more accurate compared to the SVM and ANN models (Al-Mukhtar, 2019). Conversely, another study has determined that the SVM or ANN model outperform the RF model (Ahmad et al., 2018). In this study, the RF model provided relatively better accuracy than other ML models when predicting the streamflow of the PYJ watershed using all training data sets (the All case). The ANN model was the only one that could provide acceptable accuracy (KGE of 0.84, which is greater than the threshold of 0.17 for SS) in the prediction of SS loads in the JS watershed with the WD+C training data. The SVM model provided a relatively greater KGE score than the other models when predicting the SS loads of the HN watersheds using the WD+UC training data."

Section 4.2 – Influence of model structure on ML performance across information quantity and quality: "Despite being trained on identical datasets, ML models exhibited different predictive abilities in response to increases in information quantity and quality. The ANN and SVM models could improve their predictions more efficiently in terms of the amount of information (i.e., marginal entropy) added to the training data compared to the RF model (Fig. 8(c)). The RF model uses a random sampling method to select the feature subspace for each node in growing the trees (Breiman, 2001), which is a model parameter called the "number of variables." Previous studies (Wang and Xia., 2016; Ye et al., 2013) have argued that when applying the random sampling method to a high-dimension data set, model may select many subspaces that do not include informative features and will increase error bounds for the RF model. This study agrees with previous studies: RF performed relatively poorly when dimension of a training data set was higher (i.e., the large number of independent variables) than SVM and ANN. A traditional ANN model with one or two hidden layers is known to suffer performance degradation due to its rapid growth in the number of connection weights (Krenker et al., 2011). However, one study demonstrated that a deep neural network that employs numerous hidden layers, such as the one used in this study, could yield promising performance with high-dimensional training data (Liu et al., 2017).

In addition, 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. While SVM models use kernel function 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-dimensionality 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 robustness by effectively exploiting additional information even when lower-quality data (WD+UC) were added to the training dataset (Figs. 8(b) and (e)). This indicates that the ANN model excels at handling high-dimensional, non-linear data, making it 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 labelled 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 hydrological modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity."

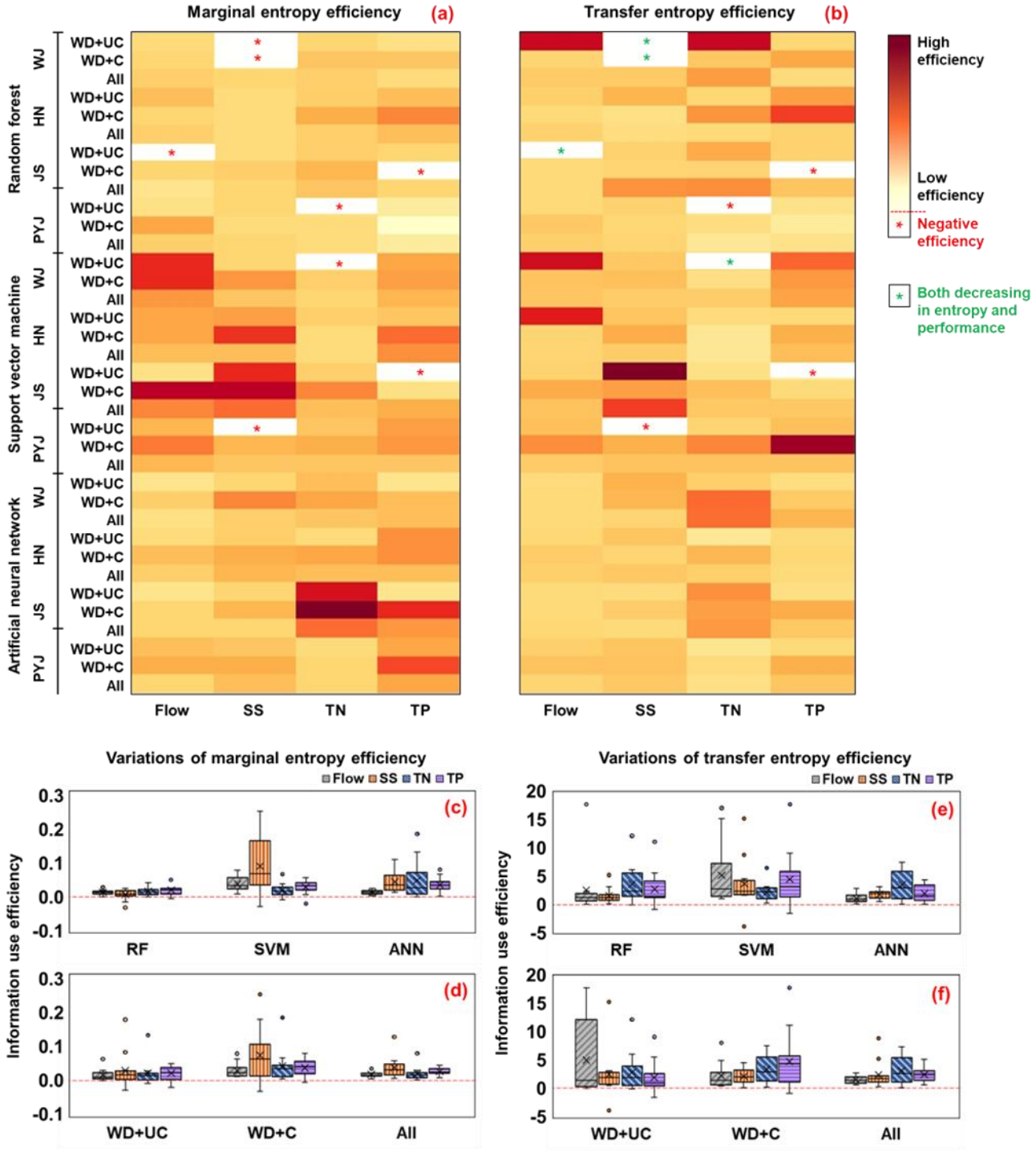

Figure 8. Comparison of changes in information use efficiency (IUE) calculated from the entropy (ME and TE) and accuracy (KGE) statistics provided by using the different training sets. (a) and (b) exhibit relative changes in information use efficiency (IUE-ME and IUE-TE) across all factors, including the ML models, watersheds, training datasets, and predicted variables. Red asterisks marks represent "negative efficiency," describing the cases where prediction accuracy decreased with increases in entropy, and green asterisks marks indicate the cases where both IUE

and KGE decrease. (c) and (d) summarize, with box-whisker plots, the variations of changes in IUE-ME for the watersheds and training datasets (c) and for the watersheds and ML models (d), respectively; (e) and (f) show the variations of changes in IUE-TE for the watersheds and training datasets (e) and for the watersheds and ML models (f), respectively.

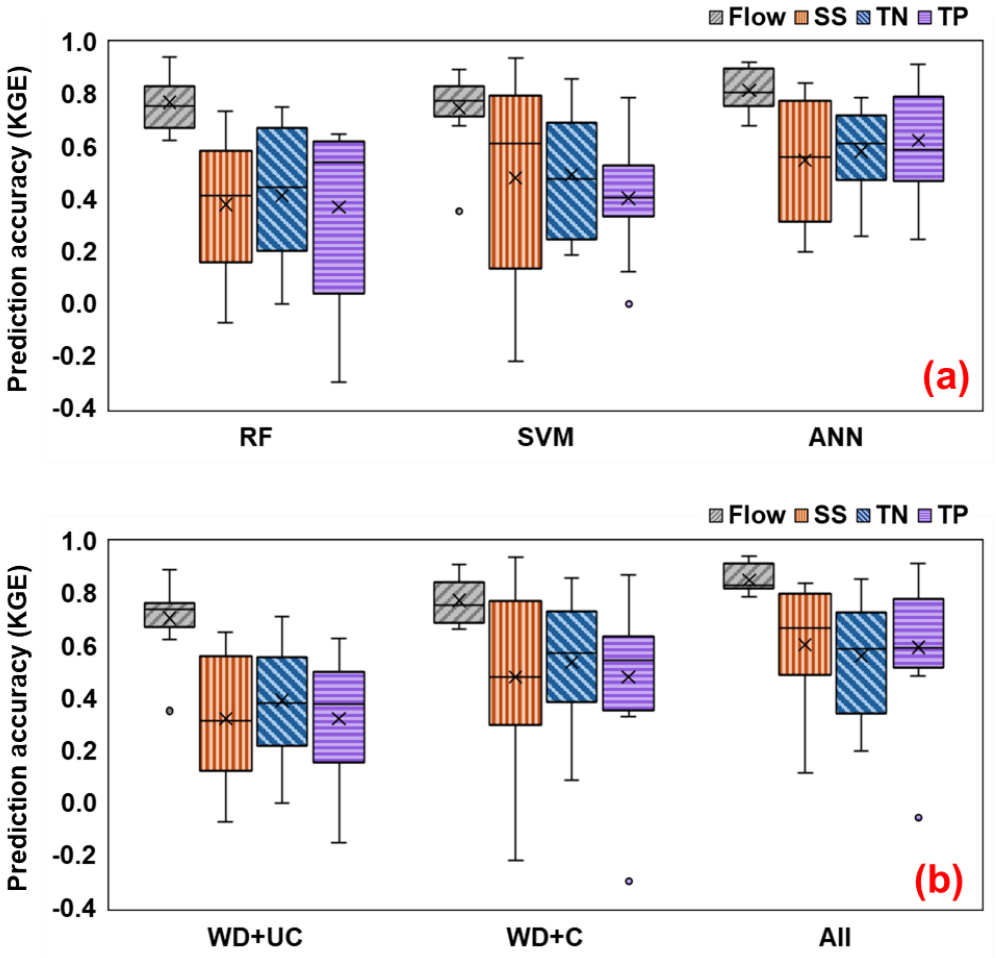

Figure S5. Variations in prediction accuracy (KGE) of the ML models for the watersheds and training datasets (a) and for the watersheds and ML models (b).