https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

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

Minhyuk Jeung[1], Younggu Her[2], Sang-Soo Baek[3], Kwangsik Yoon[1]

[1] Department of Rural & Biosystems Engineering (Brain Korea 21), Chonnam National University, Gwangju 61186, Republic of Korea
[2] Department of Agricultural and Biological Engineering / Tropical Research and Education Center, University of Florida, Homestead, Florida 33186, USA
[3] Department of Environmental Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Correspondence to: Kwangsik Yoon (ksyoon@chonnam.ac.kr)

Abstract. Machine learning (ML) is now commonly employed as a tool for hydrological prediction due to recent advances in computing resources and increases in data volume. The prediction accuracy of ML (or data-driven) modeling is known to be improved through training with additional data; however, the improvement mechanism needs to be better understood and documented. This study explores the connection between the amount of information contained in the data used to train an ML model and the model's prediction accuracy. The amount of information was quantified using Shannon's information theory, including marginal and transfer entropy. Three ML models were trained to predict the flow discharge, sediment, total nitrogen, and total phosphorus loads of four watersheds. The amount of information contained in the training data was increased by sequentially adding weather data and the simulation outputs of uncalibrated and/or calibrated mechanistic (or theory-driven) models. The reliability of training data was considered a surrogate of information quality, and accuracy statistics were used to measure the quality (or reliability) of the uncalibrated and calibrated theory-driven modeling outputs to be provided as training data for ML modeling. The results demonstrated that the prediction accuracy of hydrological ML modeling depends on the quality and quantity of information contained in the training data. The use of all types of training data provided the best hydrological ML prediction accuracy. ML models trained only with weather data and calibrated theory-driven modeling outputs could most efficiently improve accuracy in terms of information use. This study thus illustrates how a theory-driven approach can help improve the accuracy of data-driven modeling by providing quality information about a system of interest.

1 Introduction

Machine learning (ML) techniques have become commonly employed for hydrological prediction due to the availability of large hydrological data repositories and advances in computing resources and techniques (Sun et al., 2020; Xu and Liang, 2021). Studies have demonstrated that ML techniques can predict hydrological variables as accurately or even better than other statistical methods and mechanistic (or theory-driven) modeling (Panidhapu et al., 2020). The prediction accuracy of ML modeling is known to increase with the volume of data used to train the models (Jha et al., 2018); as such, the accuracy is

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

expected to improve further as hydrological observations and records accumulate over time. However, it remains unclear how prediction accuracy is associated with the characteristics of training data: can any data added to a training set improve the accuracy?

Information theory has served as a mathematical tool to measure the amount of information contained in data and its transfer to another set of data (Shannon, 1948a; Shannon, 1948b). This tool can help us understand the correlations or dependencies among multiple interconnected data sets (Pechlivanidis et al., 2018), which helps determine whether the training data contains information that could improve the accuracy of the model (Nearing et al., 2020). Shannon's entropy, often called marginal entropy, is one of the most commonly used information theories that can quantify information content in a set of data (Silva et al., 2017). The concept of transfer entropy was proposed to measure the amount of information transferred from one variable to another (Schreiber, 2000). Previous studies have employed marginal entropy to quantify the amount of information in hydrological datasets (Silva et al., 2017) and transfer entropy to qualify the interactions between input and output data in hydrological analyses (Bennett et al., 2019; Konapala et al., 2020). Both marginal and transfer entropies have great potential as concepts and methods to evaluate the informatic characteristics of training data and their impacts on hydrological ML model performance.

Data-driven methods, including ML modeling, rely on historical records and estimates from other analyses, while theory-driven approaches employ existing hydrological concepts and knowledge for prediction. Mechanistic modeling can be classified as a theory-driven method even when its parameter calibration has the nature of a data-driven approach. Mechanistic models employ different assumptions, knowledge, and methods to conceptualize a hydrological system of interest, which is why they provide unique predictions. For example, streamflow hydrographs predicted using Hortonian and Dunne's concepts might be substantially different from each other even after parameter calibration (Loague et al., 2010). Information embedded in hydrological theories and models can help improve the performance of data-driven modeling, and the information is considered in the predictions of mechanistic modeling. Weather records are one of the data sets commonly used to train hydrological ML models (Chen et al., 2020). Previous studies have demonstrated that ML models trained only with meteorological data provide limited accuracy; this is unsurprising given that hydrological processes are usually complicated by many other factors, including topography, soil, land use and cover, geological features, and management practices (Srinivasan et al., 2010; Srivastava et al., 2020). Hence, mechanistic model predictions can be an alternative source of data for the training of hydrological ML models.

Mechanistic modeling often or always requires parameter calibration to consider the hydrological characteristics of an area of interest. In a technical sense, parameter calibration is an effort to improve the statistical similarity between observed and predicted variables of interest. The prediction accuracy of mechanistic modeling is usually improved through the calibration process. As a result, the amount and/or quality of information in a relatively accurate prediction may be greater and/or higher than that of information in a relatively inaccurate one. When the prediction accuracy of mechanistic modeling is improved by calibrating its parameters, the calibrated model may have more and/or better-quality information than the uncalibrated model.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

Thus, a pair of uncalibrated and calibrated mechanistic models for a watershed can be a useful tool to create training data sets with different amounts and/or qualities of information for hydrological ML modeling.

Recent advancements in hydrological modeling highlight the importance of combining ML approaches with diverse data sources and process-based models to improve prediction accuracy. Kratzert et al. (2021) demonstrated how deep learning models could leverage the synergy among multiple meteorological datasets to enhance rainfall-runoff predictions, emphasizing the role of data integration in improving model performance. Similarly, Razavi et al. (2022) advocated for the coevolution of machine learning and process-based models, suggesting that their combined use can address limitations inherent to each approach and revolutionize Earth and environmental sciences. Reichstein et al. (2019) explored the intersection of data-driven methods and process understanding, illustrating how deep learning can advance Earth system science by extracting insights from complex datasets while maintaining a connection to fundamental physical principles. These studies underscore the critical interplay between information quantity, quality, and model design, which is central to this study.

This study attempted to relate the quantity and quality of information contained in sets of training data to the prediction accuracy of hydrological ML models, with the goal of understanding how to improve the accuracy efficiently. Information quantity and quality were quantified using information theory, including marginal and transfer entropies statistics. The research question that this study tried to answer was how the quantity and quality of information in training datasets, as measured by marginal and transfer entropies, can affect the prediction accuracy of hydrological ML models. We hypothesized that both higher information quantity and quality in training datasets, as reflected by increased marginal and transfer entropies values, would together positively correlate with improved model prediction accuracy. Three different ML algorithms (or models) were tested in the evaluation. The quantity of information was systematically increased by adding weather records and uncalibrated and calibrated theory-driven (or mechanistic) model outputs to training data sets. This study employed a mechanistic model commonly used to predict flow and water quality to represent a theory-driven approach. The data-driven (i.e., ML) and theory-driven (i.e., mechanistic) modeling approaches were applied to predict the flow discharge, suspended solid (SS), total nitrogen (TN), and total phosphorus (TP) loads of four watersheds. Then, the implications of the evaluation results and the limitations of this study were discussed, and the future direction of hydrological ML modeling was suggested based on the findings.

2 Methods and Materials

2.1 Overall procedure

Three ML models, including Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), were employed in this study (Fig. 1). The ML models were first trained using data sets collected from four study watersheds, including weather data observed at weather stations within or close to the watersheds and flow discharge, SS, TN, and TP measured at the outlets of the four study watersheds. The Soil and Water Assessment Tool (SWAT) model was selected to represent a mechanistic (or theory-driven) model. The SWAT model was used to produce additional data sets to which the ML models would be trained. The SWAT models were calibrated to flow (or streamflow discharges), SS, TN, and TP loads

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

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. Here, we assumed there would be a difference in the quality of information contained in the uncalibrated and calibrated outputs of the SWAT model, and we employed the marginal and transfer entropy based on Shannon's information theory, to quantify the amount and quality of information contained in the training data sets. Finally, the weather data and uncalibrated and calibrated SWAT model outputs were sequentially fed to the three ML models.

Specifically, four datasets were prepared using different types of hydrological data, including weather records and outputs from theory-driven (or mechanistic) models. The first dataset consisted solely of weather data, serving as the baseline. The second and third datasets incorporated uncalibrated and calibrated hydrological model outputs into the baseline, respectively, to evaluate how integrating hydrological knowledge from mechanistic models improves predictions. Additionally, we compared how variations in input data quality influenced model performance. The final dataset combined all input variables, including weather data and both uncalibrated and calibrated model outputs. The baseline model in our study was not intended to replicate a fully developed nutrient model; rather, it was purposefully designed as a foundational benchmark for systematically assessing the effects of incorporating additional information—particularly theory-driven outputs from uncalibrated and calibrated mechanistic models. This framework allowed us to isolate and quantify the incremental gains in ML prediction accuracy as the training data increased in both informational quantity and quality. The use of calibrated mechanistic model outputs as ML training data was a deliberate methodological choice to evaluate how improvements in data reliability influence prediction performance. By excluding nutrient inputs and management practices from the baseline model, we established a consistent reference point for evaluating the added value of such variables when subsequently introduced through mechanistic model outputs. To rigorously assess these effects, we applied Shannon's marginal and transfer entropy to quantify the amount and quality of information embedded in each training dataset. This structured approach enabled a theory-grounded evaluation of how ML model accuracy responds to changes in both data quantity and quality (Fig. 1 and Table 1). Furthermore, we calculated information use efficiency to measure how effectively each ML model leveraged added information, based on the improvement in prediction accuracy per unit increase in entropy across the four training data sets.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

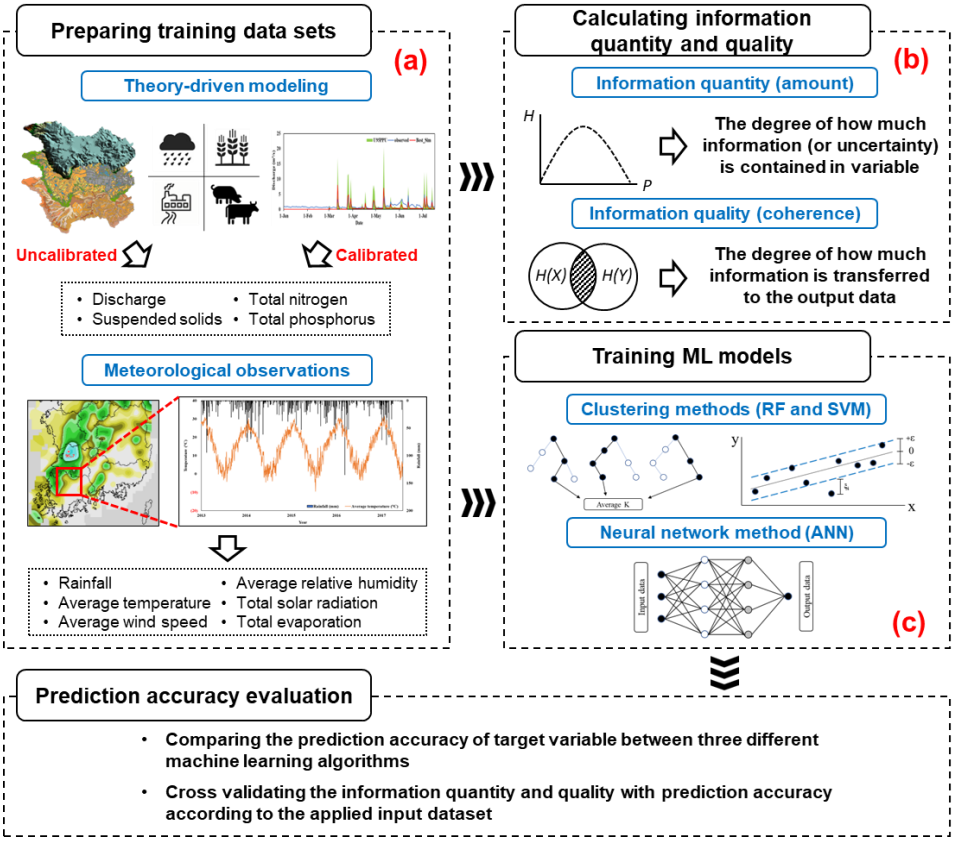


Figure 1. Overall procedure to investigate the contribution of information quantity and quality to the prediction accuracy of hydrological machine learning (ML) modeling.

Table 1. Combinations of data sets used to train hydrological ML models.

Training Data Sets	Variables
WDO	P, AT, WS, RH, SR, E
WD+UC	P, AT, WS, RH, SR, E, Q_Uncal*, SS_load_Uncal*, TN_load_Uncal*, TP_load_Uncal*
WD+C	P, AT, WS, RH, SR, E, Q_Cal*, SS_load_Cal*, TN_load_Cal*, TP_load_Cal*
All	P, AT, WS, RH, SR, E, Q_Uncal*, SS_load_Uncal*, TN_load_Uncal*, TP_load_Uncal*, Q_Cal*, SS_load_Cal*, TN_load_Cal*, TP_load_Cal*

* P, AT, WS, RH, SR, and E represent precipitation, average temperature, wind speed, relative humidity, solar radiation, and evaporation, respectively. The SWAT outputs, including Q, SS load, TN load, and TP load, were used to train ML models

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

separately depending on the target variables. For example, SWAT's SS load simulation results were used only when predicting SS load using ML models.

2.2 Data-driven (or machine learning) models

The RF model is based on a regression tree, but it differs as it does not grow with a single tree but rather an entire forest of numerous trees using the bootstrap aggregating technique or bagging technique to help decrease model variance (Breiman et al., 1984; Breiman, 2001). The RF model is known for its ability to be used when there are more variables than observation data, and it does not result in overfitting due to the pruning process (Diaz-Uriate and de Andrés, 2006). The RF model was also reported to offer excellent performance even when predictive variables are irregular (Diaz-Uriate and de Andrés, 2006). In RF modeling, a decision tree grows by splitting a tree node, and it is pruned by removing tree nodes or sections with relatively low explanatory power compared to others (Hasanipanah et al., 2017).

The SVM model divides a high- or infinite-dimensional space using hyperplanes until all data points are separated (Vapnik, 1995; 1998). The SVM model is known to be able to avoid overfitting and produces highly accurate predictions (Aktan, 2011). The goal of the SVM procedures is to identify the optimal hyperplane separating two classes in the high-dimensional space that maximizes the distance between the two data point groups (Ahmed et al., 2017). SVM modeling transforms training data using the kernel function so that a linear hyperplane can separate the data points in high dimensions. Three kernel functions are commonly used: radial basis function (RBF), linear function, and polynomial function. This study employed the RBF, which is the most widely used kernel function (Tao et al., 2008).

The ANN model has been widely used to solve various modeling problems (Khashei and Bijari, 2010). The structure of the ANN model was inspired by the biological structure of the human brain, which is composed of many interconnected processing elements called neurons (Tosun et al., 2016). The structure is characterized by a network of three layers: input, hidden, and output. The number of input and hidden layers is determined by the number of input variables and the complexity of the problem (Yilmazkaya et al., 2018). Neurons are a critical parameter used in interconnected processing, which is characterized by weights (Tosun et al., 2016). The weights of individual neurons determine how input values are transferred to other values on the output nodes. The weights of connections between layers are calculated by the backpropagation process, which calculates the gradient of prediction error with respect to weights (Siddique and Tokhi, 2001).

The optimization of three machine learning models (i.e., 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

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

variables, including precipitation, temperature, and watershed characteristics that were carefully selected to align with the study's objectives.

2.3 Data normalization and accuracy evaluation

ML modeling is known to have low learning rates when some types of training data have value ranges substantially different from those of others (Ioffe and Szegedy, 2015). Data normalization techniques are commonly used to rescale the training data from their original ranges into a common value range so that the ML models can be efficiently and quickly trained. Several data normalization methods are available; linear scaling is one of the most widely used, presumably due to its simplicity and efficacy (Raju et al., 2020; Eq. 1).

$$X' = (x - \min(x)) / (\max(x) - \min(x)) \quad (1)$$

where χ' is the normalized value of the data set (ranges from 0 to 1), and x is an original value.

The prediction accuracy of the three ML models was evaluated using the Kling-Gupta efficiency coefficient (KGE; Gupta et al., 2009). The KGE considers the strength of the correlation between observed and predicted variables while also comparing the variables' biases and variances. Thus, compared to the Nash-Sutcliffe efficiency and the coefficient of determination, the KGE is less sensitive to relatively large values that lead to biases toward such values (Nash and Sutcliffe, 1970; Gupta et al., 2009; Eq. 2).

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \quad (2)$$

where σ_{obs} and σ_{sim} are the standard deviations of observations and simulation results, respectively, and μ_{obs} and μ_{obs} are the averages of observed and simulated variables, respectively.

A KGE of 1 indicates perfect agreement between observations and predictions (Andersson et al., 2017). Knoben et al. (2019) mathematically demonstrated that the KGE value approaches -0.41 when the predicted (or simulated) values of a variable are equal to the average value of its observations. Thus, a KGE value of -0.41 can be interpreted similarly to an NSE value of 0.00, meaning that the predictions may not be a better than the observed mean (Schaefli and Gupta, 2007). In this study, we assumed that predictions would be acceptable or satisfactory when the differences between observed and simulated averages of a variable (or percentage biases) and the variances of the differences are less than 25% for flow, 55% for SS, and 70% for TN/TP (Moriasi et al., 2007), which correspond to KGEs of 0.54, 0.17, and -0.03 for flow, SS, and TN/TP, respectively, with an arbitrarily selected threshold correlation of 0.30.

2.4 Theory-driven (or mechanistic) model

The SWAT model was designed to predict watershed processes based on theories and known mechanisms that control the generation and transport processes of water, sediment, and nutrients (Nietsch et al., 2002). The SWAT model is popularly used to predict water and nutrient loadings at the watershed and basin scales due to its proven applicability to a variety of landscapes and climate zones as well as its simple but defendable modeling strategies. Moreover, the SWAT model can consider various

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

management practices, including application rates and timing of fertilizers and herbicides/pesticides; tillage and low-impact development practices; and agricultural conservation practices such as filter strips, nutrient management plans, terraces, and tile drainage (Her et al., 2017; Her and Jeong, 2018; Li et al., 2021a). Several studies have attempted to improve the prediction accuracy of SWAT modeling by coupling it with ML techniques, for example, to predict peak flow (Senent-Aparicio et al., 2019), water quality (Noori et al., 2020), and aquifer vulnerability (Jang et al., 2020).

Two versions of the SWAT model, namely uncalibrated and calibrated mechanistic modeling outputs, were prepared to generate two sets of training data for the ML models. The agricultural management practices, including fertilizer application, planting and harvest dates, compiled from the study watersheds were incorporated into both models (RDA, 2014; Table S2). The values of all parameters of the uncalibrated SWAT model remained unchanged; thus, the uncalibrated SWAT models do not necessarily represent the hydrological processes of the study watersheds, and they are not likely to reproduce the observed flow, SS, TN, and TP at an acceptable accuracy level. Accordingly, the quality of information contained in the outputs of the uncalibrated SWAT models may be relatively low compared to that of the calibrated SWAT models. The parameter values of the SWAT models were calibrated to flow, SS, TN, and TP observations made at the study watersheds' outlets (Table S3). While the SWAT model includes many parameters, previous studies (Arnold et al., 2012; Douglas-Mankin et al., 2010; El-Sadek and Irvem, 2014) have identified key parameters that relatively substantially influence model outputs. In this study, we focused on these parameters for each target variable and calibrated each watershed independently. Calibration was performed sequentially; upstream watersheds were calibrated first, and their calibrated parameter values remained unchanged while calibrating the corresponding parameters for their downstream watersheds. For example, the WJ watershed (which is nested by the HN watershed) was calibrated prior to the HN watershed, and then the parameters for areas that are not included in the WJ watershed but only in the HN watershed were calibrated to observations made at the outlet of the HN watershed. For the target variable, previous studies (Arnold et al., 2012; Engel et al., 2007; Santhi et al., 2001) have recommended that streamflow should be calibrated first, followed by SS, and then TN and TP, due to the interdependencies among these constituents resulting from shared transport processes. The flow, SS, TN, and TP loads predicted using the calibrated SWAT models were assumed to have relatively high-quality information compared to those of the uncalibrated SWAT model. The quantity and quality of information were quantified using the marginal and transfer entropies described in the following section.

The SUFI-2 algorithm, widely used for SWAT model calibration, was used to explore the multi-dimensional parameter spaces of the SWAT models and locate a solution (or a parameter set) close to the global optimum in this study (Sao et al., 2020). The simulation period was split into three: a warm-up period from January 1, 2008, to July 11, 2013; a calibration period from July 12, 2013, to December 31, 2015; and a validation period from January 1, 2016, to December 31, 2017. The types and value ranges of the calibration parameters were determined based on the previous SWAT modeling experience, the understanding of the calibration parameters, and the literature (Tobin and Bennett, 2017; Tang et al., 2021).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

2.5 Marginal and transfer entropy

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.

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. Once the marginal and transfer entropies were calculated for the modeling experiments with the unique combinations of the ML models and the training data sets (Fig. 1 and Table 1), the prediction accuracy gain was divided by the increases in the quantity (marginal entropy) and quality (transfer entropy) of information contained in the training data to calculate information use efficiency (IUE):

$$IUE_{ME} = \frac{P_{WD\to ID}}{\sum H(x_{WD}\to x_{ID})} \quad (7)$$

$$IUE_{TE} = \frac{P_{WD\to ID}}{\sum T_{WD_i\to Y} \to \sum T_{ID_i\to Y}} \quad (8)$$

where $P_{WD\to ID}$ is the prediction accuracy gain or increase from using additional straining data sets, as compared to the case of only using weather data for the training. The $H(x_{WD\to}x_{ID})$ and $\sum T_{WD_i\to Y} \to \sum T_{ID_i\to Y}$ denotes the marginal entropy and transfer

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

entropy gain or increase from using additional straining data sets, compared to the case of only using weather data for the training.

2.6 Study watersheds and training data acquisition

The Pung-Yeong-Jung (PYJ) river watershed was selected for the modeling experiment of this study. The PYJ watershed can be divided into three sub-watersheds from upstream to downstream: the Wall-Jeong (WJ), Ha-Nam (HN), and Jang-Su (JS) watersheds (Fig. 2 and Table S1). The WJ watershed is nested by the HN watershed, and the HN and JS watersheds are nested by the PYJ watershed. Thus, all direct runoff drained from the three nested watersheds passes the outlet (35°09'58.87" N, 126°49'08.93" E) of the PYJ watershed. The streamflow, SS, TN, and TP concentrations were monitored at the outlets of the four study watersheds for four years and six months, from July 12, 2013, to December 31, 2017. Most of the drainage areas were covered by agricultural land uses, including upland and rice paddy fields (covering 41% of the JS watershed and 62% of the WJ watershed) and forest. Urbanized areas cover 5% (WJ watershed) to 31% (JS watershed) of the watersheds.

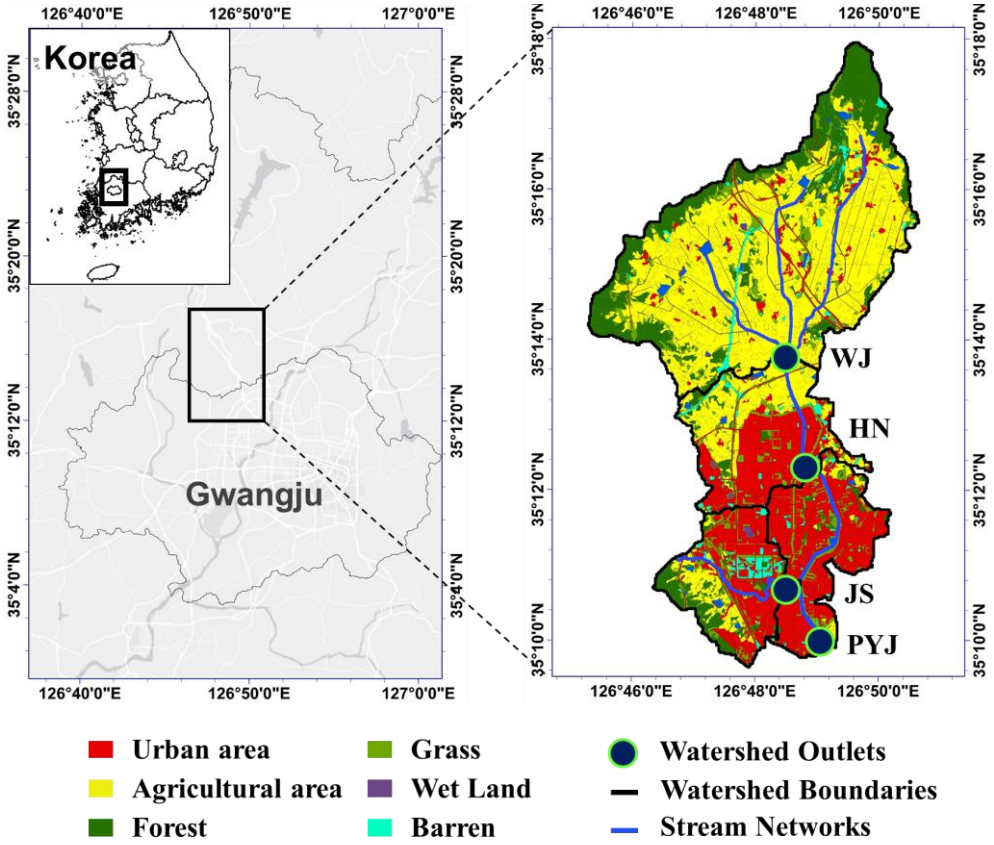


Figure 2. Location of the study watersheds and their land uses and covers.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

Streamflow, SS, TN, and TP concentrations were monitored at the outlets of the four study watersheds over a period of four years and six months, from July 12, 2013, to December 31, 2017. 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 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. 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.

The Korean Meteorological Administration monitors weather variables, including daily AT, P, E, WS, RH, and SR, at a weather station located approximately 7 km away from the study watersheds. Water pressure sensors and data loggers (OTT Orpheus Mini, Germany) were deployed at the monitoring sites close to the watershed outlets. The cross-sections of the streams were surveyed at the monitoring sites, and the velocity of streamflow was then measured using a flow meter (VALEPORT model 002, UK) across the sections to estimate flow discharge rates. Water quality samples were manually collected every one or two weeks. During a rainfall event, stream water was collected using an automatic sampler (ISCO portable sampler 6712, USA), and the sampling interval was reduced to 1 hour to catch the expected large variations of flow rates and the corresponding water quality concentrations for improved observation accuracy. During the monitoring period, a total of 17 large rainfall events were sampled.

3 Results

3.1 Training data: Weather records and monitoring data

Sets of training data were prepared using the daily weather records, the watershed monitoring data, and SWAT modeling results (uncalibrated and calibrated outputs; Figs. 3 and S1–S4). The watersheds have four seasons, with relatively short springs and falls. The watersheds are fairly wet in the summer and dry in the spring. For example, the watersheds receive precipitation of 831–1333 mm annually, with more than half (59% on average) of the precipitation occurring in the summer (from June to September). In spring, the stream might dry up due to the small amount of precipitation and warm air. In the case of the PYJ watershed, streamflow discharges can be large, with as much as 2.64 m^3/s on average in summer, but they are limited (e.g., 1.21 m^3/s) enough to reveal the bottom of the stream in spring (from March to May).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

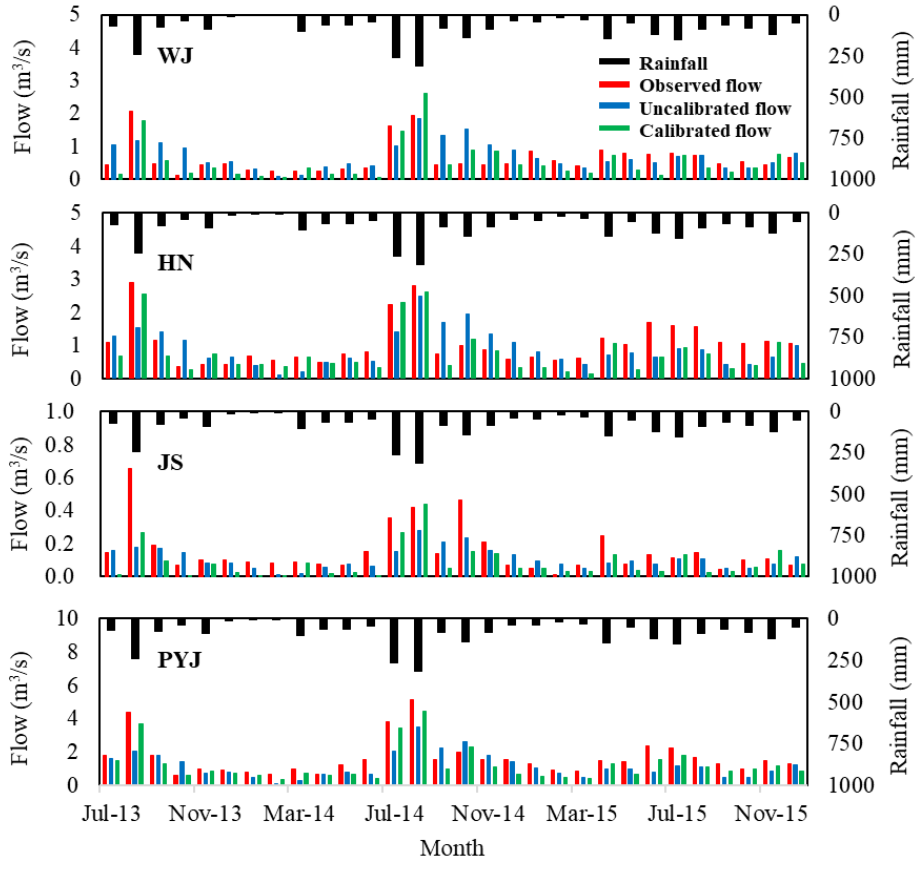


Figure 3. Comparison of monthly streamflow predicted using the mechanistic models (i.e., uncalibrated and calibrated SWAT models) and observed during the training period (July 12, 2013, to December 31, 2015). The daily-scale comparisons can be found in the supplementary document (Figs. S1–S4).

300

The PYJ and JS watersheds had the largest and smallest average daily discharge of 1.69 and 0.16 m^3/s, respectively (Table S4). The JS watershed had relatively higher SS concentrations compared to the other watersheds as it includes large construction sites (Mendie, 2005; Pullanikkatil et al., 2015; Adeola-Fashae et al., 2019). In addition, the first flush effects of the urbanized watersheds (e.g., the JS watershed; Table S1) led to higher peak SS, TN, and TP concentrations (Chaudhary et
305 al., 2022). The WJ and HN watersheds had relatively higher TN and TP concentrations, presumably due to agricultural management activities such as fertilizer application and livestock farming in their large agricultural areas (Liu et al., 2012; Table S4).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

3.2 Training data: Outputs of the mechanistic modeling

The calibrated SWAT model provided acceptable performance in all watersheds (e.g., KGEs equal to or greater than 0.54 for flow, 0.17 for SS, and -0.03 for TN/TP). The average KGE values for all watersheds were 0.68 for flow, 0.45 for SS, 0.40 for TN, and 0.44 for TP (Table 2). However, as expected, the uncalibrated model could not accurately predict the variables; average KGEs for all watersheds were less than 0.41 for flow, 0.02 for SS, -0.20 for TN, and -0.35 for TP. As such, the information quality of the outputs of the calibrated SWAT modeling may be greater than that of the uncalibrated modeling. The quantity and quality of information were evaluated with marginal and transfer entropies.

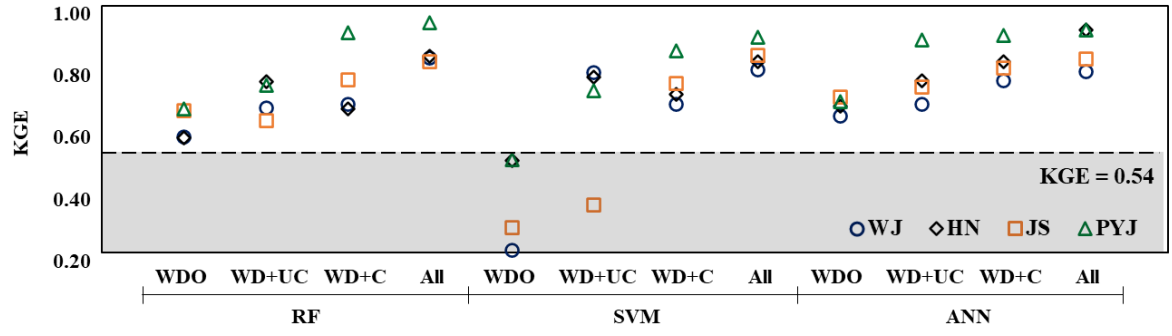
Table 2. Accuracy statistics (KGEs) of a theory-driven (or SWAT) model in the training period. The KGE scores that satisfy the acceptable accuracy criteria (i.e., 0.54 for flow, 0.17 for SS, -0.03 for TN/TP) are in bold.

Watershed	Flow		SS		TN		TP	
	Uncal	Cal	Uncal	Cal	Uncal	Cal	Uncal	Cal
WJ	0.49	**0.71**	**0.28**	**0.52**	-0.28	**0.41**	-0.39	**0.43**
HN	0.50	**0.70**	-0.06	**0.36**	-0.09	**0.43**	-0.33	**0.47**
JS	0.18	**0.57**	-0.35	**0.45**	-0.44	**0.37**	-0.41	**0.27**
PYJ	0.46	**0.72**	**0.22**	**0.48**	**0.01**	**0.40**	-0.27	**0.57**

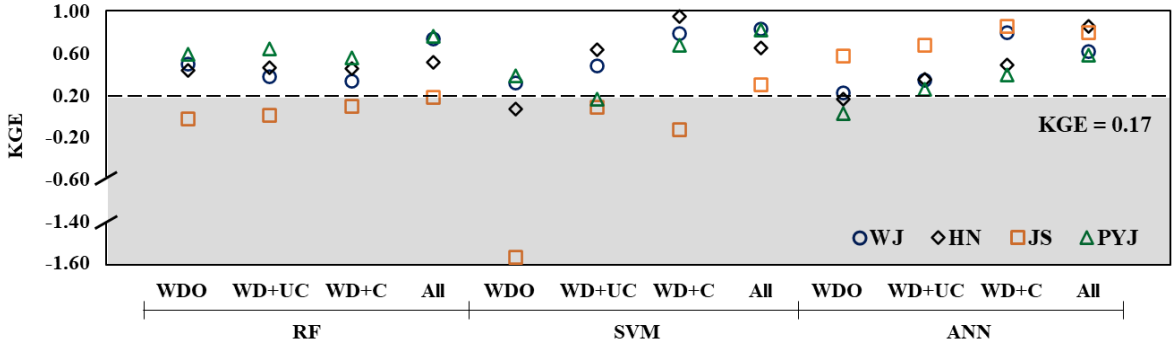
3.3 Prediction accuracy of machine learning modeling

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. 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.

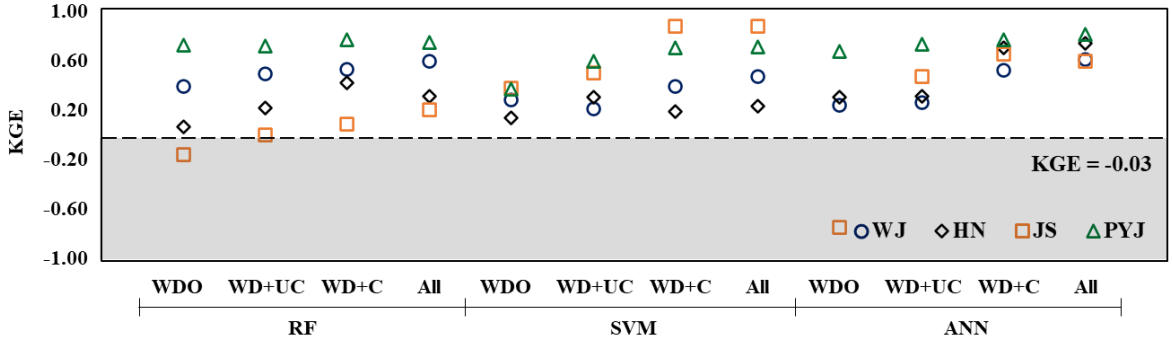
https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025



(a) Flow

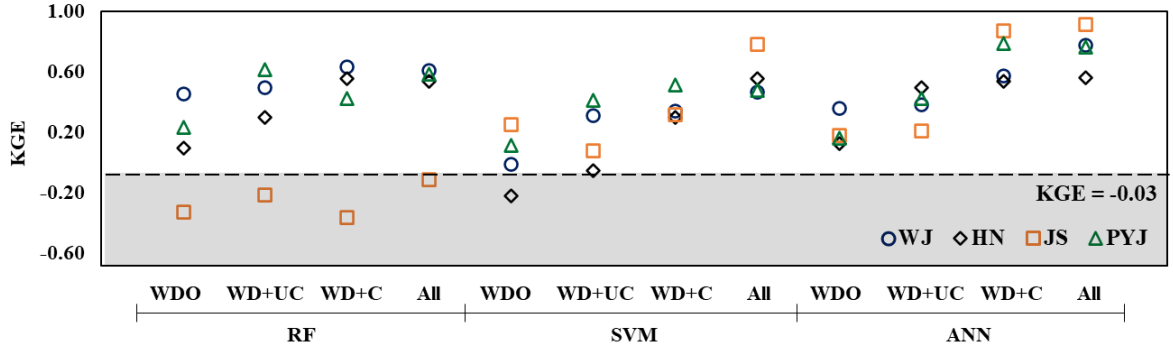


(b) SS



(c) TN

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

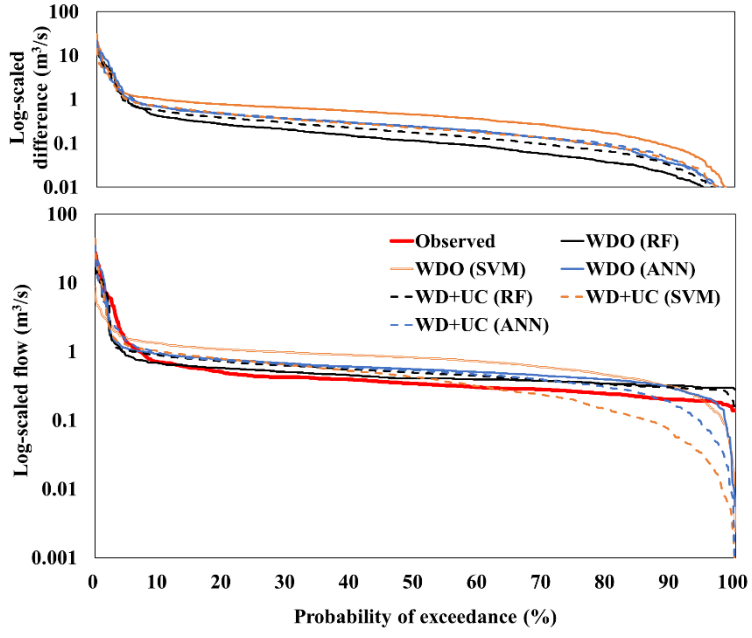


(d) TP

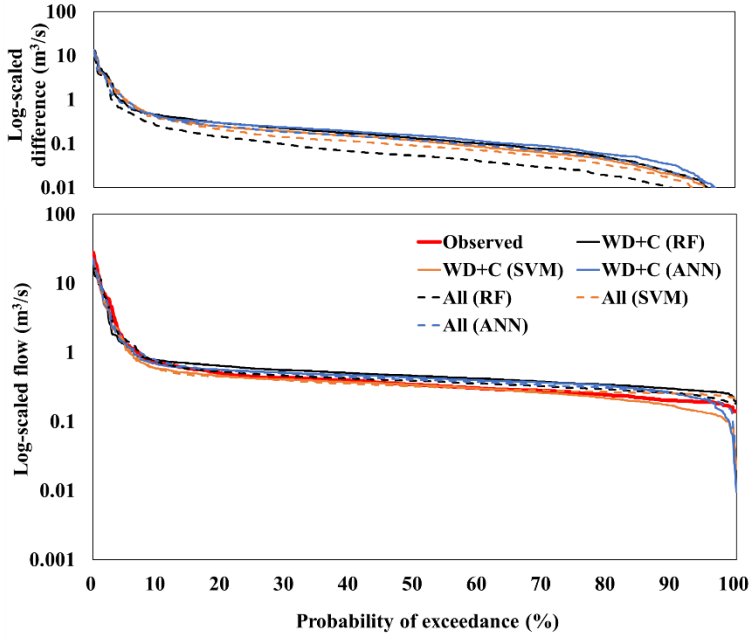
Figure 4. Prediction accuracy (KGE) of hydrological ML models trained with the different training data set combinations. The KGE values that do not satisfy the acceptable accuracy levels (e.g., i.e., 0.54 for flow, 0.17 for SS, and -0.03 for TN/TP) are included in gray areas.

A flow duration curve (FDC) provides a graphical way of investigating the frequency of extreme events, such as floods and droughts. The FDCs were derived from the observed and predicted flow hydrographs and compared to evaluate prediction accuracy in the frequency domain (Figs. 5 and S5). The FDCs created from flow predictions made using the ML models trained with all training data (the All case) were the closest to the observed FDC in both high (e.g., flooding) and low (e.g., drought) exceedance probability regions. The WDO and WD+UC cases created relatively large differences (under- and over-estimations) between the predicted and observed FDCs, especially for extreme events (i.e., flooding and drought). For example, the differences between the ANN predictions and observations for the 5% (flooding) and 95% (drought) exceedance probabilities of the WJ watershed were 18.1% and 13.8% in the All case respectively, and they increased to 28.8% and 77.9% in the WD+UC case. The findings indicate that the ML models trained with all available training data sets (the All case) can more accurately predict the extremes than the relatively less trained ML models (the WDO and WD+UC cases).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025



(a) the WDO and WD+UC cases.



(b) the WD+C and All cases.

Figure 5. Comparison between observed and ML-predicted FDCs at the outlet of the WJ watershed.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

3.4 Information quantity and quality

The amount and quality of information contained in the training data sets for the ML training were quantified using the marginal and transfer entropy concepts, respectively. Then, INFORMATION USE EFFICIENCY scores were calculated to understand how efficiently information quantity and quality can improve the prediction accuracy of hydrological ML modeling. marginal entropy of the training data sets generally increased as additional data (i.e., the outputs of the uncalibrated and calibrated mechanistic or SWAT modeling) were added to the weather data (i.e., WDO case; Figs. 6 and S6–S7). In the case of predicting the flow of the WJ watershed, for example, marginal entropy increased by 7.8 to 17.1 bits when the uncalibrated and/or calibrated SWAT modeling outputs were added to the training data respectively, compared to the WDO case (Fig. 6). The All cases increased marginal entropy more substantially than the WD+UC and WD+C cases, regardless of the watersheds and variables. The WD+C cases did not always increase marginal entropy more (or efficiently) than the WD+UC cases, and the marginal entropy increases were negligible even when they did occur (e.g., in the case of predicting TP loads); this confirms that marginal entropy does not consider the association between two variables (i.e., watershed responses observed and simulated using the calibrated mechanistic models) in the training data sets, which is one of the features that marginal entropy has. Thus, marginal entropy does not change depending on the types of ML models as marginal entropy only counts information in the training data.

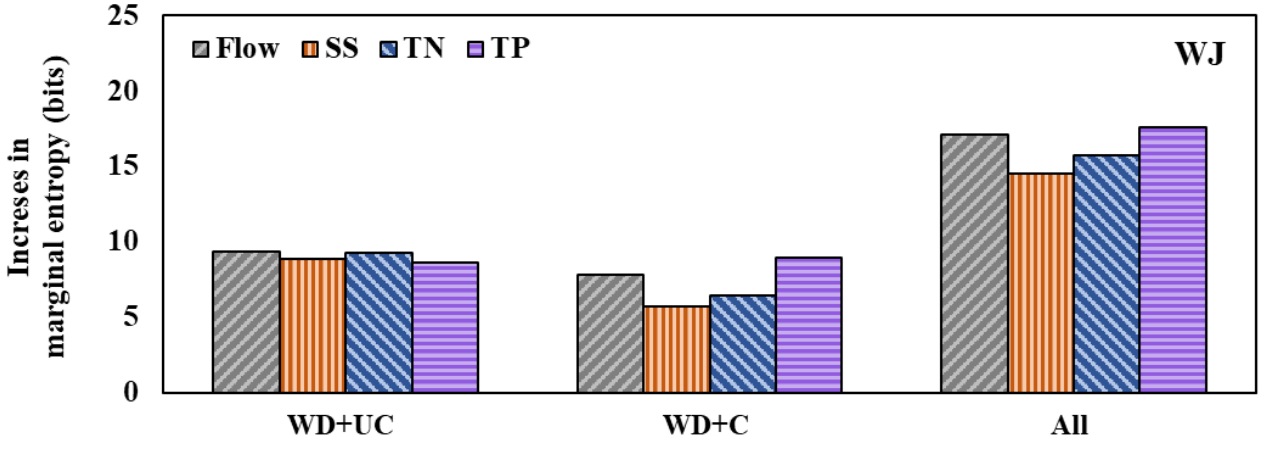


Figure 6. Increases in marginal entropy due to the addition of additional training data sets in the case of the WJ watershed. The WDO training data set serves as the baseline for this comparison.

Transfer entropy did not always increase with additional training data (Figs. 7 and S8, Table S6). For example, in the case of the RF modeling trained with the WDO data set for the WJ watershed, the transfer entropy of SS loads decreased from 0.385 to 0.190 and 0.294 when adding uncalibrated and calibrated mechanistic modeling outputs, respectively (Fig. 7); this

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

indicates that a loss of information was commonly found in the target and training data sets when adding additional data, such as uncalibrated and calibrated modeling outputs, to the training data set. The amount of precipitation information (0.174 bits) was transferred to the SS prediction for the WJ watershed in the case of WDO. However, when adding the uncalibrated mechanistic (i.e., SWAT) modeling output to the training data set, the amount of transferred precipitation information decreased to 0.066 bits, whereas only 0.044 bits were transferred from the uncalibrated SWAT modeling output. Here, the information loss of 0.064 bits can be calculated by subtracting 0.110 bits (amount of information on precipitation and uncalibrated mechanistic modeling output when applying WD+UC training data set) from 0.174 bits. Transfer entropy quantifies the amount of information contained in the training data sets that is effectively transferred into the predictions made using the trained ML models. It captures not only the shared information between input and output variables but also the directional flow of information from one variable to another (Schreiber, 2000). Thus, transfer entropy depends on the types of training data sets, prediction variables, and ML models (Figs. S7 vs. S9).

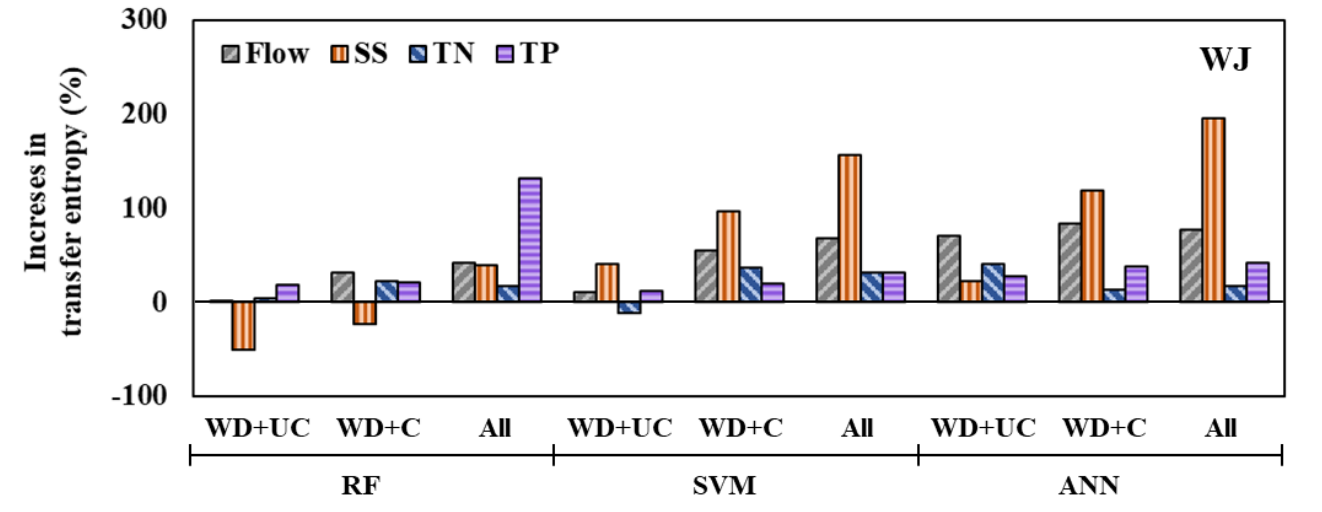


Figure 7. Increases in transfer entropy due to the addition of training data sets in the case of the WJ watershed. The WDO training data set serves as the baseline for this comparison.

3.5 Information use efficiency

Information use efficiency represents the relative improvement of prediction accuracy compared to the baseline per unit change of information quantity (Eqs. 7 and 8). Information use efficiency was calculated by dividing the increases in KGEs (the WDO training data set serves as the baseline) by the differences between the amount of information quantified using marginal entropy (IUE-ME) or transfer entropy (IUE-TE) contained in the training data sets (Fig. 8).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

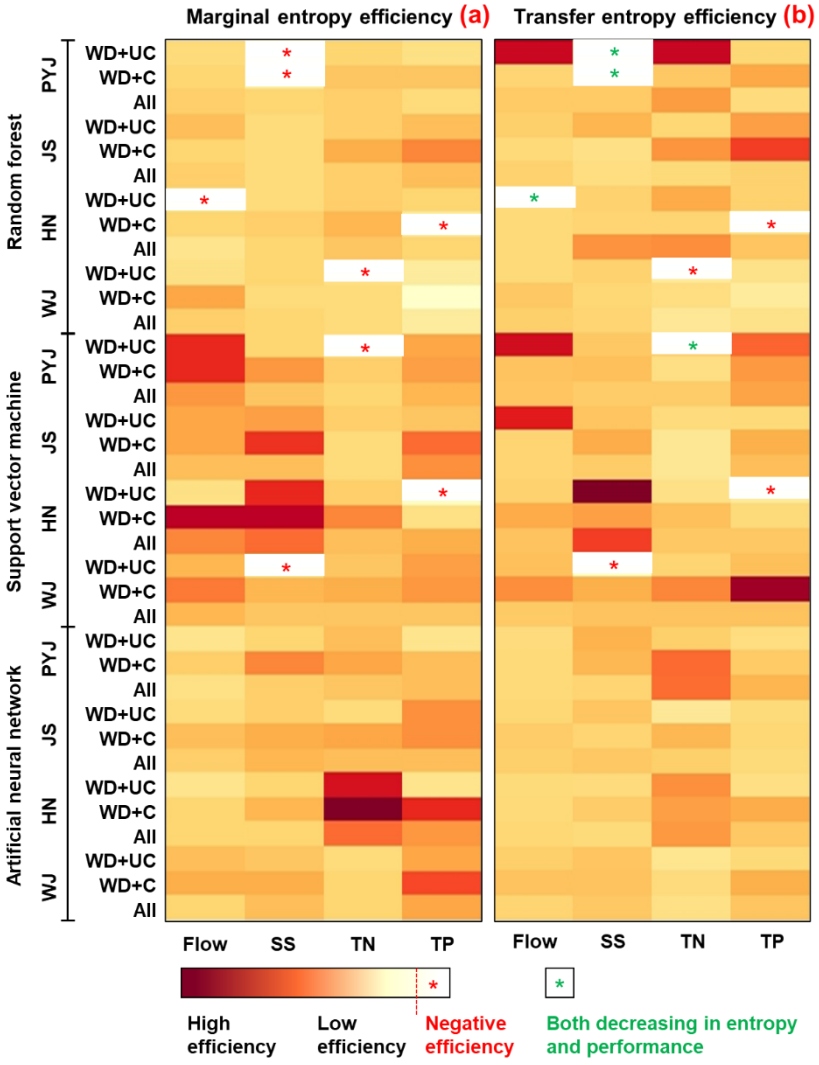


Figure 8. Comparison of information use efficiency calculated from the entropy (marginal and transfer entropies) and accuracy (KGE) statistics provided by using the different training sets. "Negative efficiency" describes the case where prediction accuracy decreased with increases in entropy, which is presented with a red symbol. In addition, decreases in both entropy and prediction accuracy are presented with a green symbol.

The case of WD+C provided a relatively higher IUE-ME compared to the other training data cases (Table S5). This means that the prediction accuracy of ML modeling can be most efficiently improved when the outputs of the calibrated mechanistic modeling are added to the training data sets (i.e., WD+C). Interestingly, WD+C may be more efficient than the All case, which added the uncalibrated theory-driven modeling outputs to WD+C. This finding implies that information quality can more efficiently improve the prediction accuracy of hydrological ML modeling than information quantity. However, it is worth

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

noting that the All case still provided the best prediction accuracy (or the highest KGE), but its efficiency in increasing KGE scores was lower than that of WD+C when considering the relative accuracy improvement to the amount of added information.

IUE-ME were often negative, especially in the WD+UC case, indicating that prediction accuracy decreased even when entropy increased (red star in Fig. 8[a]); this is because marginal entropy always increases with additional input variables, regardless of their quality or association with the target variables. IUE-TE also showed negative efficiency, which means the KGE decreased with increases in the transfer entropy. Model performance might not necessarily relate to transfer entropy because of complicated associations among weather forcings, management practices, watershed feathers, and responses (Konapala et al., 2020). KGEs also decreased when transfer entropy decreased (green star in Fig. 8[b]), which implies that transfer entropy can capture the decrease in information flow between independent (i.e., weather data, uncalibrated and calibrated modeling outputs) and dependent (or target) variables that may lead to decreased prediction accuracy (or decreased KGEs). This inverse relationship was primarily detected when adding uncalibrated mechanistic modeling outputs to the training data set, demonstrating the role of information quality in ML modeling training.

4 Discussions

This study investigated how the prediction accuracy of hydrological ML modeling is associated with the quantity and quality of information contained in the training data. The results exhibited that prediction accuracy (KGE scores) generally increased with the amount and quality of information contained in the training data sets (all cases except the cases with stars in Fig. 8). Hence, access to both a large quantity and high-quality information helps increase hydrological ML modeling accuracy. However, the prediction accuracy of hydrological ML modeling and its association with entropy scores were found to be dependent on the study watersheds, target variables, and the ML models.

The accuracy of the ML modeling 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 (Figs. 9 and S10). 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. 10). 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. 10).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

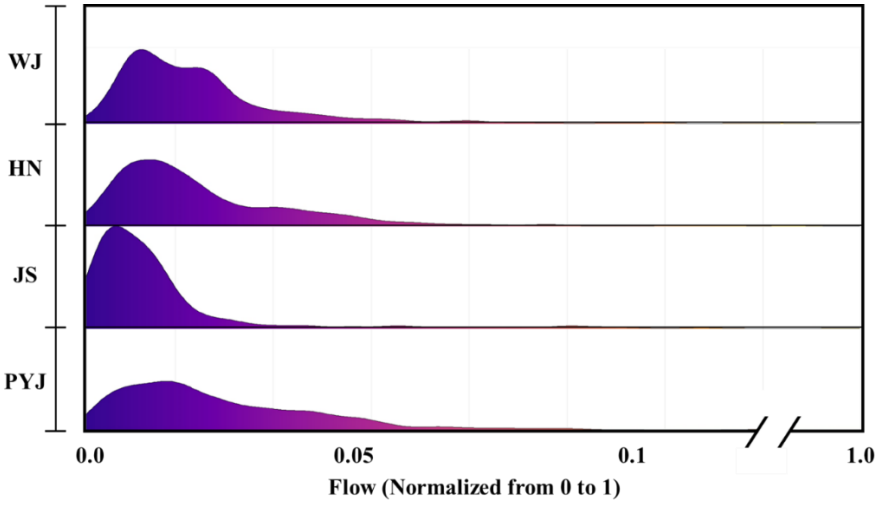


Figure 9. Density (or frequency) distributions of observed flow data (i.e., target variable) during the training period. The flow was normalized from 0 to 1 for each watershed.

440

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

Watershed	Flow	SS	TN	TP
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

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

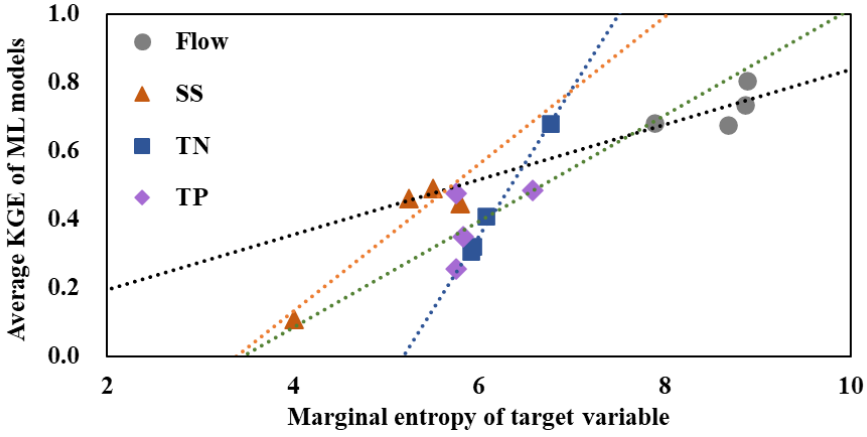


Figure 10. Linear relationship between the average KGE scores of three ML models trained using the four training data sets and the marginal entropy of the target variables.

The prediction accuracy of the ML models also 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 (Fig. 4). 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, which is also why prediction accuracy varied by watershed. For example, the flow hydrographs commonly have relatively higher 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 (Figs. 9 and S10). 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.

This study used continuous daily flow rate and nutrient load data that were derived from mechanistic models. However, raw flow rate and nutrient concentration observations can be valuable as part of the training data for machine learning models. Nutrient concentration records are not typically available at daily intervals; instead, they are usually collected at longer temporal intervals, likely due to cost concerns. In this study, the nutrient concentrations were measured every one or two weeks. In contrast, flow rate data—generated from flow level observations—are generally recorded at finer temporal resolutions (such as hourly or daily) and converted using a rating curve. As a result, nutrient concentration records provide less detailed

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

information compared to nutrient load estimates, which can be obtained from mechanistic model outputs. Consequently, nutrient concentration records may have a more limited impact on improving the predictive accuracy of hydrological machine learning models, compared to the daily nutrient load estimates.

To evaluate this assumption, we conducted seven additional experiments directly incorporating observed flow rates and nutrient concentration records into training data to assess their impact on the predictive accuracy of hydrological machine learning models: 1) WD+O_Flow (with observed flow rates), 2) WD+O_NCon (with observed nitrogen concentrations), 3) WD+O_PCon (with observed phosphorus concentrations), 4) WD+O_Flow+O_NCon (with both observed flow rates and nitrogen concentrations), 5) WD+O_Flow+O_PCon (with both observed flow rates and phosphorus concentrations), 6) WD+S_NLoad (with simulated nitrogen loads using the calibrated SWAT model), and 7) WD+S_PLoad (with simulated phosphorus loads using the calibrated SWAT model). Here, "Flow," "NCon," and "PCon" refer to observed flow rates, nitrogen concentrations, and phosphorus concentrations, respectively. In addition, “S_NLoad” and “S_PLoad” represent simulated nitrogen and phosphorus loads derived from using the calibrated mechanistic model, serving as reference cases that use continuous daily data.

The test results indicated that the WD+O_Flow case performed comparably or even better in predicting TN and TP loads compared to the WD+O_NCon and WD+O_PCon cases (Fig. S11). Previous studies have shown that pollutant loads are primarily influenced by streamflow rather than nutrient concentrations, as demonstrated by regression models relating streamflow to pollutant loads (Lee et al., 2016; Song et al., 2022; Song et al., 2024; Wu et al., 2024). This is likely because the energy, or transport capacity, of flowing runoff primarily drives the detachment of sediment and nutrients from the soil surface and their downstream transport to waterbodies, particularly under transport-limited (rather than supply-limited) conditions (Basu et al., 2010; Wainwright et al., 2015; Song et al., 2024). Additionally, while using all available variables (WD+O_Flow+O_NCon and WD+O_Flow+O_PCon) improved performance relative to cases that only incorporated nutrient concentrations alongside weather records (WD+O_NCon and WD+O_PCon), it still underperformed when compared to using datasets containing continuous daily nutrient load data derived from mechanistic models (WD+S_NLoad and WD+S_PLoad). This result was expected, given the significantly smaller number of observed nutrient concentration records compared to the daily nutrient load estimates produced by mechanistic models. For example, TN concentrations were measured every one to two weeks from July 12, 2013, to December 31, 2017, resulting in 109 to 229 data points per watershed. In contrast, the mechanistic model generated 1,634 daily TN load estimates for the same period. The average information content of the observed concentrations was 7.001 bits for TN and 6.808 bits for TP, substantially lower than the simulated loads, which contained 8.600 bits for TN and 9.437 bits for TP, on average. These findings highlight the crucial role of data quantity in improving the predictive accuracy of hydrological machine learning models.

None of the ML models consistently provided more accurate predictions than the others (Fig. 4). 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). Some study has demonstrated that the RF model is more accurate compared to the SVM and ANN models (Al-Mukhtar, 2019). Conversely, other study has determined that the SVM or ANN model outperform the RF model (Ahmad et al., 2018). In this

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

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.

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[a]). 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).

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 the input space, struggled to manage the "curse of dimensionality" (Bellman, 1961) and complex non-linear relationships between variables. 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 information use efficiency scores, demonstrating its resilience and ability to more efficiently utilize quality information, even with lower-quality training data in cases such as WD+UC (Fig. 8). 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 labelled data and adapting to input quality. The absence of negative information use efficiency scores for ANN underscores its flexibility and robustness. These findings affirm the ANN model's suitability for high-dimensional hydrological

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

modeling, highlighting its advantage over other methods in tasks requiring predictive precision and adaptability to data complexity.

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. Future studies can focus on including input data that have strong statistical relationships to target variables. 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. Incorporating high-quality training data into ML training can significantly enhance the reliability and efficiency of ML models. 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

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

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.

5 Conclusions

This study provides clear evidence that the predictive accuracy of hydrological ML models is intricately linked to both the quantity and the quality of information embedded within the training datasets. The highest overall accuracy was achieved when the full suite of available data sources was used to train the models (i.e., the All case). However, the most efficient improvements in predictive accuracy—relative to the amount of information added—were realized when high-quality outputs from calibrated mechanistic models were incorporated (i.e., the WD+C case).

To better characterize the nature of these improvements, we employed metrics from information theory, specifically marginal entropy and transfer entropy, to quantify the amount and relevance of information in each training dataset. Furthermore, we introduced information use efficiency as a novel indicator to evaluate how effectively additional information contributes to gains in prediction accuracy.

The findings underscore the critical role of information quality in hydrological ML modeling. In particular, augmenting training datasets with low-relevance or low-accuracy data (as in the WD+UC case) did not necessarily improve—and in some cases degraded—model performance. This highlights the potential risks associated with increasing data volume without due consideration of informational relevance.

Importantly, the combined application of marginal entropy, transfer entropy, and information use efficiency offers a robust and interpretable framework for evaluating and optimizing training data. These metrics facilitate a more rigorous assessment of how data characteristics influence model performance, providing actionable insights for both model development and data acquisition strategies. By leveraging this framework, hydrologists and modelers can more effectively select and structure training datasets that balance both informational richness and efficiency, ultimately improving the robustness and generalizability of hydrological ML predictions.

Author contributions. **MJ**: Conceptualization, Software, Validation, Formal analysis, Writing - Original Draft; **YH**: Conceptualization, Methodology, Supervision, Writing - Review & Editing; **SB**: Validation, Formal analysis, Data Curation; **KY**: Conceptualization, Supervision, Writing - Review & Editing.

Competing interests. The contact author has declared that none of the authors has any competing interests

Acknowledgements. This research was supported by a project titled “A Long-term Monitoring for the Nonpoint Sources Discharge” (Yeongsan and Seomjin River Water Management Committee).

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

References

Adeola-Fashae, O., Abiola-Ayorinde, H., Oludapo-Olusola, A., Oluseyi-Obateru, R., 2019. Landuse and surface water quality in an emerging urban city. Appl. Water Sci. 9(25), Doi: 10.1007/s13201-019-0903-2.

Ahmad, I., Basheri, M., Iqbal, M.J., Rahim, A., 2018. Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. IEEE Access 6, 33789-33795.

Ahmed, S., Khalid, M., Akram, U., 2017. A Method of Short-Term Wind Speed Time Series Forecasting Using Support Vector Machine Regression Model. "2017 6th International Conference on Clean Electrical Power (ICCEP), 190-195. Doi: 10.1109/ICCEP.2017.8004814.

Aktan, S., 2011. Application of machine learning algorithm for business failure prediction. Invest. Manage. And Financial Inno. 8(2), 52-65.

Al-Mukhtar, M., 2019. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. Environ. Monit. Assess. 191, 673.

Alzubi, J., Nayyar, A., Kumar, A., 2018. Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series 1142, 012012.

Andersson, J.C.M., Arheimer, B., Traoré, F., Gustafsson, D., Ali, A., 2017. Process refinements improve a hydrological model concept applied to the Niger River basin. Hydrol. Process. 31, 4540-54.

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Santhi, C., Harmel, R., Van Griensven, A., Van Liew, M.W., 2012. SWAT: Model use, calibration, and validation. Transactions of the ASABE 55, 1491-1508.

Basu, N.B., Destouni, G., Jawitz, J.W., Thompson, S.E., Loukinova, N.V., Darracq, A., Zanardo, S., Yaeger, M., Sivapalan, M., Rinaldo, A. and Rao, P.S.C., 2010. Nutrient loads exported from managed catchments reveal emergent biogeochemical stationarity. Geophys. Res. Lett., 37(23).

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press.

Bennett, A., Nijssen, B., Ou, G., Clark, M., Nearing, G., 2019. Quantifying Process Connectivity with Transfer Entropy in Hydrologic Models. Water Resour. Res. 55, 4613-4629.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5-32.

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. CRC press, Wadsworth.

Cabaneros, S.M.S., Calautit, J.K.S., Hughes, B.R., 2017. Hybrid Artificial Neural Network Models for Effective Prediction and Mitigation of Urban Roadside NO2 Pollution. Energy Procedia 142, 3524-3530.

Chaudhary, S., Chua, L.H.C., Kansal, A., 2022. Event mean concentration and first flush from residential catchments in different climate zones. Water Res. 219, 118594.

Chen, Z., Zhu, Z., Jiang, H., Sun, S., 2020. Estimating daily reference evapotranspiration based on limited meteorological data using deep learning and classical machine learning methods. J. Hydrol. 591, 125286.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

Choi, J.Y., Engel, B.A., Chung, H.W., 2002. Daily streamflow modelling and assessment based on the curve-number technique. Hydrol. Process. 16, 3131-3150.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. John Wiley & Sons, Inc., Hoboken.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7, 1-13.

Douglas-Mankin, K., Srinivasan, R., Arnold, J., 2010. Soil and Water Assessment Tool (SWAT) model: Current developments and applications. Transactions of the ASABE 53, 1423-1431.

El-Sadek, A., Irvem, A., 2014. Evaluating the impact of land use uncertainty on the simulated streamflow and sediment yield of the Seyhan River basin using the SWAT model. Turkish Journal of Agriculture and Forestry 38, 515-530.

Engel, B., Storm, D., White, M., Arnold, J., Arabi, M., 2007. A Hydrologic/Water Quality Model Applicati1. JAWRA Journal of the American Water Resources Association 43, 1223-1236.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80-91.

Hasanipanah, M., Faradonbeh, R.S., Amnieh, H.B., Armaghani, D.J., Monjezi, M., 2017. Forecasting blast-induced ground vibration developing a CART model. Eng. Comput. 33, 307-316.

Her, Y., Jeong, J., 2018. SWAT+ versus SWAT2012: Comparison of sub-daily urban runoff simulations. Trans. ASABE 61(4), 1287-1295.

Her, Y., Jeong, J., Arnold, J., Gosselink, L., Glick, R., Jaber, F., 2017. A new framework for modeling decentralized low impact developments using Soil and Water Assessment Tool. Environ. Model. Softw. 96, 305-322.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Acceleration Deep Network Training by Reducing Internal Covariate Shift. arXiv preprint.

Jang, W.S., Engel, B., Yeum, C.M., 2020. Integrated environmental modeling for efficient aquifer vulnerability assessment using machine learning. Environ. Model. Softw. 124, 104602.

Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C., Agrawal, A., 2018. ElemNet: Deep Learning the Chemistry of Materials from Only Elemental Composition. Sci. Rep. 8, 17593.

Jones, D.R., 2001. A Taxonomy of Global Optimization Methods Based on Response Surfaces. J. Glob. Optim. 21, 345-383.

Khashei, M., Bijari, M., 2010. An artificial neural network (p,d,q) model for timeseries forecasting. Expert Syst Appl. 37, 479-489.

Kim, N.W., Lee, J., 2008. Temporally weighted average curve number method for daily runoff simulation. Hydrol. Process. 22, 4936-4948.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci. 23, 4323-31.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

Konapala, G., Kao, S.C., Addor, N., 2020. Exploring Hydrologic Model Process Connectivity at the Continental Scale Through an Information Theory Approach. Water Resour. Res. 56, e2020WR027340.

Kratzert, F., Klotz, D., Hochreiter, S., Nearing, G.S., 2021. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 25, 2685-2703.

Krenker, A., Bester, J., Kos, A., 2011. Introduction to the artificial neural networks. In K. Suzuki (Ed.), Artificial neural networks-methodological advances and biomedical applications. Rijeka, Croatia: IntechOpen.

Lee, C.J., Hirsch, R.M., Schwarz, G.E., Holtschlag, D.J., Preston, S.D., Crawford, C.G. and Vecchia, A.V., 2016. An evaluation of methods for estimating decadal stream loads. J. Hydrol. 542, 185-203.

Li, S., Liu, Y., Her, Y., Chen, J., Guo, T., Shao, G., 2021a. Improvement of simulating sub-daily hydrological impacts of rainwater harvesting for landscape irrigation with rain barrels/cisterns in the SWAT model. Sci. Total Environ. 798, 149336.

Li, T.Y., Xiang, H., Yang, Y., Wang, J., Yildiz, G., 2021b. Prediction of char production from slow pyrolysis of lignocellulosic biomass using multiple nonlinear regression and artificial neural network. J. anal. appl. pyrolysis. 159, 105286.

Liu, B., Wei, Y., Zhang, Y., Yang, Q., 2017. Deep Neural Networks for High Dimension, Low Sample Size Data. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), 2287-2293.

Liu, Z., Yang, J., Yang, Z., Zou, J., 2012. Effects of rainfall and fertilizer types on nitrogen and phosphorus concentrations in surface runoff from subtropical tea fields in Zhejiang, China. Nutr. Cycl. Agroecosyst. 93, 297-307. Doi: 10.1007/s10705-012-9517-x.

Loague, K., Heppner, C.S., Ebel, B.A., VanderKwaak, J.E., 2010. The quixotic search for a comprehensive understanding of hydrologic response at the surface: Horton, Dunne, Dunton, and the role of concept-development simulation. Hydrol. Process. 24, 2499-2505.

Mendie, U.E., 2005. The theory and practice of clean water production for domestic and industrial use: Purified and package water. Lacto-Medal Ltd, Lagos.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASAE. 50 (3), 885–900. Doi: 10.13031/2013.23153.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10, 282-290.

Nearing, G.S., Ruddell, B.L., Bennett, A.R., Prieto, C., Gupta, H.V., 2020. Does Information Theory Provide a New Paradigm for Earth Science? Hypothesis Testing. Water Resources Research 56, e2019WR024918.

Nietsch, S.L., Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., 2002. SWAT: Soil and water assessment tool user's manual. Texas Water Resources Institute, USDA Agricultural Research Service, College Station, TX.

Noori, N., Kalin, L., Isik, S., 2020. Water quality prediction using SWAT-ANN coupled approach. J. Hydrol. 590, 125220.

Panidhapu, A., Li, Z., Aliashrafi, A., Peleato, N.M., 2020. Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. Water Res. 170, 115349.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

Pechlivanidis, I.G., Gupta, H., Bosshard, T., 2018. An Information Theory Approach to Identifying a Representative Subset of Hydro-Climatic Simulations for Impact Modeling Studies. Water Resources Research 54, 5422-5435.

Pullanikkatil, D., Palamuleni, L.G., Ruhiiga, T.M., 2015. Impact of land use on water quality in the Likangala catchment, southern Malawi. Afr. J. Aquat. Sci. 40(3), 277-286. Doi: 10.2989/16085914.2015.1077777.

Qiu, L., Zheng, F., Yin, R.-s., 2012. SWAT-based runoff and sediment simulation in a small watershed, the loessial hilly-gullied region of China: Capabilities and challenges. International J. Sediment. Res. 27, 226–234.

Raju, V.N.G., Lakshmi, K.P., Jain, V.M., Kalidindi, A., Padma, V., 2020. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 729-735.

Razavi, S., Hannah, D.M., Elshorbagy, A., Kumar, S., Marshall, L., Solomatine, D.P., Dezfuli, A., Sadegh, M., Famiglietti, J., 2022. Coevolution of machine learning and process-based modelling to revolutionize Earth and environmental sciences: A perspective. Hydrol. Process. 36, e14596.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, F., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195-204.

Rural Development Administration (RDA), 2014. Agricultural work schedule – Machine transplanting cultivation. http://www.nongsaro.go.kr.

Santhi, C., Arnold, J.G., Williams, J.R., Dugas, W.A., Srinivasan, R., Hauck, L.M., 2001. VALIDATION OF THE SWAT MODEL ON A LARGE RWER BASIN WITH POINT AND NONPOINT SOURCES. J. Am. Water Resour. Assoc. 37, 1169-1188.

Sao, D., Kato, T., Tu, L.H., Thouk, P., Fitriyah, A., Oeurng, C., 2020. Evaluation of Different Objective Functions Used in the SUFI-2 Calibration Process of SWAT-CUP on Water Balance Analysis: A Case Study of the Pursat River Basin, Cambodia. Water 12, 2901.

Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075-2080.

Schreiber, T., 2000. Measuring information transfer. Phys. Rev. Lett. 85, 461.

Senent-Aparicio, J., Jimeno-Sáez, P., Bueno-Crespo, A., Pérez-Sánchez, J., Pulido-Velázquez, D., 2019. Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction. Biosyst. Eng. 177, 67-77.

Shannon, C.E., 1948a. A mathematical theory of communication. The Bell system technical journal 27, 379-423.

Shannon, C.E., 1948b. A mathematical theory of communication. The Bell System Technical Journal 27, 623-656.

Shawe-Taylor, J., Sun, S., 2011. A review of optimization methodologies in support vector machines. Neurocomputing 74(17), 3609-3618. doi: 10.1016/j.neucom.2011.06.026.

Siddique, M., Tokhi, M.O., 2001. Training neural networks: backpropagation vs. genetic algorithms. IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222). IEEE, pp. 2673-2678.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

Silva, V.d.P.R.d., Belo Filho, A.F., Singh, V.P., Almeida, R.S.R., Silva, B.B.d., de Sousa, I. F., Holanda, R.M.d., 2017. Entropy theory for analysing water resources in northeastern region of Brazil. Hydrol. Sci. J. 62(7), 1029-1038. Doi: 10.1080/02626667.2015.1099789.

Song, J.-H., Her, Y., Guo, T., 2022. Quantifying the contribution of direct runoff and baseflow to nitrogen loading in the Western Lake Erie Basins. Sci. Rep. 12, 9216.

Song, J.-H., Her, Y., Park, Y.S., Yoon, K., Kim, H., 2024. Investigating the applicability and assumptions of the regression relationship between flow discharge and nitrogen concentrations for load estimation. Heliyon 10, e23603.

Srinivasan, R., Zhang, X., Arnold, J., 2010. SWAT ungauged: hydrological budget and crop yield predictions in the Upper Mississippi River Basin. Trans. ASABE 53, 1533-1546.

Srivastava, A., Kumari, N., Maza, M., 2020. Hydrological Response to Agricultural Land Use Heterogeneity Using Variable Infiltration Capacity Model. Water Resour. Manag. 34, 3779-3794.

Sun, W., Lv, Y., Li, G., Chen, Y., 2020. Modeling River Ice Breakup Dates by k-Nearest Neighbor Ensemble. Water 12(1), 222.

Tang, X., Zhang, J., Wang, G., Jin, J., Liu, C., Liu, Y., He, R., Bao, Z., 2021. Uncertainty Analysis of SWAT Modeling in the Lancang River Basin Using Four Different Algorithms. Water 13, 341.

Tao, J., Chen, W., Wang, B., Jiezhen, X., Nianzhi, J., Luo, T., 2008. Real-Time Red Tide Algae Classification Using Naive Bayes Classifier and SVM. 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, 2888-2891.

Tobin, K.J., Bennett, M.E., 2017. Constraining SWAT Calibration with Remotely Sensed Evapotranspiration Data. J. Am. Water Resour. Assoc. 53(3), 594-604.

Tosun, E., Aydin, K., Bilgili, M., 2016. Comparison of linear regression and artificial neural network model of a diesel engine gueled with biodiesel-alcohol mixture. Alex. Eng. J. 55, 3081-3089. Doi: 10.1016/j.aej.2016.08.011.

Vapnik, V., 1995. The nature of statistical learning theory. Berlin: Springer.

Vapnik, V., 1998. Statistical learning theory. New York: John Wiley & Sons.

Wainwright, J., Parsons, A.J., Cooper, J.R., Gao, P., Gillies, J.A., Mao, L., Orford, J.D. and Knight, P.G., 2015. The concept of transport capacity in geomorphology. Rev. Geophys. 53(4), 1155-1202.

Wang, Y., Xia, S.T., 2016. A novel feature subspace selection method in random forests for high dimensional data. 2016 International Joint Conference on Neural Networks (IJCNN), 4383-4389.

Wu, D., Cao, M., Gao, W., Cheng, G., Duan, Z., Hou, X., Zhang, Y., 2024. Spatial-temporal source apportionment of nitrogen and phosphorus in a high-flow variable river. J. Hydrol.: Reg. Stud. 53, 101839.

Xu, T., Liang, F., 2021. Machine learning for hydrologic sciences: An introductory overview. Wiley Interdisciplinary Reviews. Water 8, e1533.

Ye, Y., Wu, Q., Zhexue Huang, J., Ng, M.K., Li, X., 2013. Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recognit. 46, 769-787.

https://doi.org/10.5194/egusphere-2025-2036
Preprint. Discussion started: 2 June 2025

765 Yilmazkaya, E., Dagdelenler, G., Ozcelik, Y., Sonmez, H., 2018. Prediction of mono-wire cutting machine performance parameters using artificial neural network and regression models. Eng. Geol. 239, 96-108. Doi: 10.1016/j.enggeo.2018.03.009.

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

770 Zeiger, S., Hubbart, J.A., 2016. Quantifying suspended sediment flux in a mixed-land-use urbanizing watershed using a nested-scale study design. Sci. Total Environ. 542, 315-323.