Exploring Variable Synergy in Multi-Task Deep Learning for Hydrological Modeling

Wenyu Ouyang[1], Xuezhi Gu[1], Lei Ye[1], Xiaoning Liu[1], and Chi ZHANG[1]

[1]Dalian University of Technology

November 22, 2023

Abstract

Despite advances in hydrological Deep Learning (DL) models using Single Task Learning (STL), the intricate relationships among multiple hydrological components and model inputs might not be comprehensively encapsulated. This study employed a Long Short-Term Memory (LSTM) neural network and the CAMELS dataset to develop a Multi-Task Learning (MTL) model, predicting streamflow and evapotranspiration across multiple basins. An optimal multi-task loss weight ratio was determined manually during the validation phase for all 591 selected basins with streamflow data-gaps under 5%. During test period, MTL showed median Nash-Sutcliffe Efficiency predictions for streamflow and evapotranspiration at 0.69 and 0.92, consistent with two STL models. The MTL's strength appeared when predicting the non-target variable, surface soil moisture, using probes derived from LSTM cell states—representative of the internal DL model workings. This prediction showed a median correlation coefficient of 0.90, surpassing the 0.88 and 0.89 achieved by the streamflow and evapotranspiration STL models, respectively. This outcome suggests that MTL models could reveal additional rules aligned with hydrological processes through the inherent correlations among multiple hydrological variables, thereby enhancing their reliability. We termed this as “variable synergy,” where MTL can simultaneously predict varied targets with comparable STL performance, augmented by its robust internal representation. Harnessing this, MTL promises enhanced predictions for high-cost observational variables and a comprehensive hydrological model.

Hosted file

977935_0_art_file_11538232_s3bxwt.docx available at https://authorea.com/users/695536/articles/684181-exploring-variable-synergy-in-multi-task-deep-learning-for-hydrological-modeling

Hosted file

977935_0_supp_11538212_s3bwnn.docx available at https://authorea.com/users/695536/articles/684181-exploring-variable-synergy-in-multi-task-deep-learning-for-hydrological-modeling

Exploring Variable Synergy in Multi-Task Deep Learning for Hydrological Modeling

Wenyu Ouyang[1], Xuezhi Gu[1], Lei Ye[1], Xiaoning Liu[1], Chi Zhang[1]

[1] School of Hydraulic Engineering, Dalian University of Technology, Dalian, China

Corresponding author: Lei Ye (yelei@dlut.edu.cn)

Key Points:

- Multi-task learning matched single-task models in spatiotemporal extrapolation accuracy
- Reliability of multi-task learning is proven by enhanced correlation in probe predictions
- Termed "variable synergy" for multi-task learning, highlighting its superior modeling.

Abstract

Despite advances in hydrological Deep Learning (DL) models using Single Task Learning (STL), the intricate relationships among multiple hydrological components and model inputs might not be comprehensively encapsulated. This study employed a Long Short-Term Memory (LSTM) neural network and the CAMELS dataset to develop a Multi-Task Learning (MTL) model, predicting streamflow and evapotranspiration across multiple basins. An optimal multi-task loss weight ratio was determined manually during the validation phase for all 591 selected basins with streamflow data-gaps under 5%. During test period, MTL showed median Nash-Sutcliffe Efficiency predictions for streamflow and evapotranspiration at 0.69 and 0.92, consistent with two STL models. The MTL's strength appeared when predicting the non-target variable, surface soil moisture, using probes derived from LSTM cell states—representative of the internal DL model workings. This prediction showed a median correlation coefficient of 0.90, surpassing the 0.88 and 0.89 achieved by the streamflow and evapotranspiration STL models, respectively. This outcome suggests that MTL models could reveal additional rules aligned with hydrological processes through the inherent correlations among multiple hydrological variables, thereby enhancing their reliability. We termed this as "variable synergy," where MTL can simultaneously predict varied targets with comparable STL performance, augmented by its robust internal representation. Harnessing this, MTL promises enhanced predictions for high-cost observational variables and a comprehensive hydrological model.

1 Introduction

Deep learning (DL) models, specifically Long Short-Term Memory (LSTM) neural networks (Hochreiter & Schmidhuber, 1997), have exhibited notable proficiency for data integration and generalization in hydrological modeling (Feng et al., 2020; Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Ma et al., 2021). Their ability to efficiently leverage big data, discern high-dimensional relationships between variables and building general models, as posited by the Universal Approximation Theorem (Hornik et al., 1989), has been noteworthy in hydrology (Nearing et al., 2021; Shen, 2018). Consequently, they were widely employed in modeling and predicting a range of hydrological variables, such as streamflow, soil moisture, water temperature and dissolved oxygen (Liu et al., 2022; Nearing et al., 2021; Rahmani et al., 2021; Zhi et al., 2023). Despite these advancements, DL models might learn improper patterns in hydrological modeling, even in the presence of robust goodness-of-fit results (Yokoo et al., 2022). A possible reason for this could be the focus of many deep-learning-based hydrological models on univariate modeling, which means they center their simulations on a single variable. Such an approach might increase the risk of overfitting in single-variable modeling, leading to an inadequate representation of relationships between model inputs and various hydrological components.

Conventionally, Physically Based Hydrological Models (PBHM) are also calibrated primarily using single variable data, commonly streamflow (Herman et al., 2018). However, some research has emphasized that models calibrated exclusively with streamflow may generate inadequate simulations for other water balance components (Becker et al., 2019; Tobin & Bennett, 2017; Yassin et al., 2017). Given that the hydrological process encompasses a multitude of variables involved in complex physical subprocesses, including surface and subsurface streamflow, soil water, and evapotranspiration (Shah et al., 2021), it is reasonable to incorporate additional components in the calibration of hydrological models. This could aid in constraining the solution of model parameters within a more viable parameter space (Dembélé, Hrachowitz, et al., 2020). Previous studies in physics-based modeling have shown that by incorporating a more rational representation of hydrological processes and calibrating model parameters with multiple model outputs, the overall predictive accuracy of hydrological variables could be improved, both in temporal and spatial generalization (Dembélé, Ceperley, et al., 2020; Tong et al., 2021, 2022).

While PBHMs are often calibrated using single-variable data, it is essential to note that they inherently consider the physical mechanisms of all involved variables through meaningful equations. Therefore, despite potential imperfection, PBHMs generally exhibit a reduced tendency for overfitting. On the other hand, due to their layered design and flexible architecture, DL models are more vulnerable to overfitting for one target. For example, many PBHMs can reasonably estimate evapotranspiration (Dembélé, Hrachowitz, et al., 2020; Shah et al., 2021; Yeste et al., 2023), whereas DL models struggle to predict it without direct training. This underscores the importance of further research into multi-variable calibration. Moreover, the advancements in hydrological remote sensing have facilitated the accumulation of extensive remotely-sensed hydrological variable data (McCabe et al., 2017), forming a basis for the exploration and analysis of DL models with multiple interrelated outputs.

In machine learning, multi-task learning (MTL) is an approach that enables a model to simultaneously learn the relationships between inputs and outputs of multiple tasks (Zhang & Yang, 2022). To learn the shared information between different tasks, the model needs to establish connections between the parameter spaces of different tasks in the MTL model. Hard Parameter Sharing is a prevalent method for achieving MTL (Vandenhende et al., 2022). This approach allows multiple tasks to share some encoding layers, known as shared layers, along with different task-specific layers for decoding and output. This method allows the MTL model to simultaneously learn correlations between multiple tasks and the unique features intrinsic to each task. Shared layers minimize memory usage during operation and eliminate computational cost of features within the shared layers, thereby improving the efficiency of training and testing relative to multiple single-task learning (STL) models (Vandenhende et al., 2019). Furthermore, the complementary information shared among related tasks may enable the model to learn a more generalized function relationship (Standley et al., 2020), thereby reducing the risk of overfitting.

Given the intrinsic interconnectedness of hydrological variables within a water cycle process, it is plausible to introduce MTL to hydrological deep learning-based modeling. Several studies have started to investigate the efficacy of MTL in hydrological models. Initial studies in MTL hydrological modeling primarily focused on incorporating more components in water balance, especially at large scales with abundant data. These studies utilized the water balance equation as a physical constraint and ensure that multiple interrelated hydrological processes are jointly optimized (Kraft et al., 2020). At the basin scale, Sadler et al. (2022) undertook research on MTL modeling in daily streamflow and water temperature, revealing that for certain sites and some MTL settings (like the scaling factor, denoting the ratio of loss from different tasks), MTL could enhance prediction accuracy across multiple tasks. Li et al. (2023) improved streamflow modeling with spatiotemporal DL models and an MTL approach in three basins. Building on these advancements, MTL has been adapted for a variety of hydrological targets, such as soil moisture (satellite and local in situ) (Liu et al., 2023), satellite precipitation estimation (rain/no-rain classification and rain rate) (Bannai et al., 2023), and aquifer transmissivity and storativity (Vu & Jardani, 2022). However, as the trend of modeling multiple variables has emerged, the precise benefits of MTL and how it behaves in terms of temporal and spatial generalization still not be fully understanded, particularly in scenarios with large-sample basins.

Deep learning models have the potential to revolutionize our understanding of hydrological processes, but their reliability remains a topic of ongoing research. This study aimed not only to assess a model's potential for enhancing predictive performance but also to gauge its reliability by verifying if the input-output correlations learned by the model align with the established laws of hydrological processes. Various interpretative methods exist for hydrological deep learning models (Hu et al., 2021; Kratzert et al., 2021; Schmidt et al., 2020), but most are primarily used to analyze the attribution of input variables, not the internal states of DL models, thus posing challenges for our objectives. In natural language processing, supervised models known as “probes” have been devised to predict properties from representations (Belinkov et al., 2017; Hewitt & Liang, 2019), offering a way to inspect the learnt patterns of deep learning models. By leveraging such interpretability methods, we aimed to discern what the DL model truly learns for hydrological modeling. For example, Lees et al. (2022) adopted the probe method to analyze the relationship between the cell state and a non-target hydrological variable, thereby examining the plausibility of the processes LSTMs acquired.

The aim of this study was to construct MTL deep neural network models and conduct a comprehensive evaluation of their predictive performance in terms of both temporal and spatial generalizability across large-sample basins. We also investigated whether these models could learn more dependable correlations, potentially providing new "correct" insights that align with hydrological laws. Our approach integrated both MTL and STL techniques and evaluated their performance to examine the generalization capabilities and overall reliability of MTL.

2 Data and Methods

We utilized the Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS) dataset (Addor et al., 2017), comprising 671 relatively undisturbed basins across the contiguous United States (CONUS). This dataset was widely used as a benchmark dataset for hydrological deep-learning-based modeling (Fang et al., 2022; Feng et al., 2020; Jiang et al., 2020; Kratzert, Klotz, Shalev, et al., 2019; X. Li et al., 2022). Given that evapotranspiration is frequently observed via remote sensing (Xu et al., 2019) and can serve as an output variable in hydrological models (Zhao, 1992), we integrated remotely observed evapotranspiration with the CAMELS dataset and considered it along with ground-observed streamflow as MTL target. Initially, we evaluated the performance of the models using the CAMELS dataset. As LSTM hydrological models can store hidden information that represents hydrological knowledge, we employed physical interpretability methods (Lees et al., 2022) to compare the correlation relationships between internal states and outputs of MTL and STL models. This approach facilitated an understanding of the internal states of the trained models and provided evidence for varying degrees of reliability among the models.

2.1 Dataset

The development of MTL models that simultaneously consider the output of multiple hydrological variables necessitates the assembly of datasets incorporating several hydrological model output variables. In the CAMELS dataset, for general hydrological modeling outputs, only streamflow data are obtained from observations, with the other variables' outputs derived through hydrological simulation. Thus, to explore the potential for MTL based on the CAMELS dataset, we expanded the available data for the basins in CAMELS.

We retrieved the evapotranspiration data from the MOD16A2 data product (Running et al., 2017) from the Google Earth Engine (GEE) data catalog (Gorelick et al., 2017). This dataset, comprising an 8-day temporal resolution and a 500-meter spatial resolution, spans from 2001-01-01 to the present. However, it's noteworthy that while most data collection periods are 8 days, the final collection period of each year is adjusted to 5 days for non-leap years and 6 days for leap years. The algorithm behind the MOD16 data product employs the Penman-Monteith equation, which is supplemented with daily meteorological reanalysis data and other MODIS remote sensing data products. The output includes several raster data layers, including actual evapotranspiration (ET). The pixel values in the ET data layer denote the sum of daily values for each resolution period. In this study, ET was used as the output observation for model training. To derive the basin-mean daily time series of ET data, we used Map-Reduce functions in GEE (Gorelick et al., 2017). Specifically, each pixel from the gridded ET data was allocated to a specific region, leveraging weighted reducers to ensure accurate assignment.

The CAMELS data covers the period from 1980-01-01 to 2014-12-31, while the ET data is only available from 2001-01-01 onwards. To secure a sufficient data period, we extended each time series in the supplementary CAMELS dataset to 2021-09-30. As a result, the period considered for all models in this study was from 2001-01-01 to 2021-09-30. The NLDAS-2 (NASA, 2018), Phase 2 of the North American Land Data Assimilation System and one of the sources of meteorological data in CAMELS, was used as the basin meteorological forcing data. The daily time series basin-mean forcing data were obtained in GEE using the same method as for ET. Additionally, we supplemented the streamflow data from 2015-01-01 to 2021-09-30 from the U.S. Geological Survey (USGS) National Water Information System (NWIS) (USGS, 2019). To mitigate the influence of excessive missing data on the results, we selected basins with a streamflow data loss rate of less than 5% during the overall period analyzed. This resulted in the inclusion of 591 of the 671 CAMELS basins. Attributes related to soil, geology, topography, land use types, and climate from the CAMELS dataset were also used as inputs for all models in this study. For more information on these inputs, see Table 1.

This study used surface soil moisture (SSM) to assess the reliability of the STL and MTL models (details provided in section 2.5). The data source was the SMAP global SSM dataset (Mladenova et al., 2019) from GEE. The basin-averaged SMAP grid data was obtained using the same method as for ET data acquisition. Consequently, the daily time series data for the SSM of each basin were compiled.

Table 1. Hydrological variables selected as inputs and outputs to single- and multi-task deep learning models based on the augmented CAMELS dataset.

Variable Type		Variable Name	Description	Unit
Forcings		total_precipitation	Daily total precipitation	kg/m^2
		potential_evaporation	Potential evaporation	kg/m^2
		temperature	Air temperature at 2 meters above the surface	°C
		specific_humidity	Specific humidity at 2 meters above the surface	kg/kg
		shortwave_radiation	Surface downward shortwave radiation	W/m^2
		potential_energy	Convective available potential energy	J/kg
Attributes	Terrain	elev_mean	Basin mean elevation	m
		slope_mean	Basin mean slope	m/km
		area_gages2	Basin area	km^2
	Land Cover	frac_forest	Forest proportion	-
		lai_max	Maximum monthly mean of leaf area index	-
		lai_diff	Difference between the maximum and	-

			minimum monthly mean values of the leaf area index	
		dom_land_cover_frac	Proportion of major land cover types to watershed area	-
		dom_land_cover	Major land cover types	-
	Soil	root_depth_50	Average soil layer thickness containing the top 50% of the root system	m
		soil_depth_statgso	Soil depth	m
		soil_porosity	Soil porosity	-
		soil_conductivity	Saturated hydraulic conductivity	cm/hr
		max_water_content	Maximum soil water holding capacity	m
	Geology	geol_class_1st	Most common geological category in the watershed	-
		geol_class_2nd	Second most common geological category in the watershed	-
		geol_porosity	Subsurface porosity	-
		geol_permeability	Subsurface permeability	m^2
Model Outputs		streamflow	Daily streamflow in the outlet of a basin	m^3/s
		evaportranspiration	Basin mean daily actual evaportranspiration	mm/day

2.2 Multi-task LSTM

LSTMs have become a prevalent choice in hydrological modeling due to their ability to capture temporal sequences and intricate patterns in the data. In this framework, MTL simultaneously optimizes multiple related tasks, leveraging shared representations, while STL focuses on optimizing one specific task. In this study, both MTL and STL models incorporated LSTM structures. The LSTMs, analogous to those proposed in prior research (Feng et al., 2020; Ma et al., 2021), including our previous study (Ouyang et al., 2021), operated as N-to-N models. This N-to-N term indicates that for every N input sequences, N output sequences are generated. These models leverage meteorological forcings and static basin attributes to predict daily discharge in the CAMELS dataset. We developed MTL models using a hard parameter sharing architecture (Vandenhende et al., 2022). The STLs' structure was totally same with the MTL's except that they only calculated the loss of one output variable. With this setting, we controlled the varying factor and the difference between STL and MTL was only the output.

Figure 1 presents the structure of a single time-step unit in the MTL hydrological model. Central to this design is a shared layer, composed of a fully connected input layer and an LSTM unit. The input layer consisted of two layers, each containing 256 neurons. As data progresses through the LSTM, it attempts to express the intricate temporal patterns of hydrological processes. Emerging from this shared space are multiple, parallel fully connected output layers, each corresponding to a hydrological task. Each task-specific output's neurons were arranged in two layers, with 128 and 1 neurons, respectively. Both the input and output fully connected layers introduce non-linearity through the ReLU activation function. To summarize, during forward computation of the model, inputs pass through a shared layer, generating long sequential multiple feature variables. Then these variables are moved through different output layers to generate the corresponding multiple outputs.

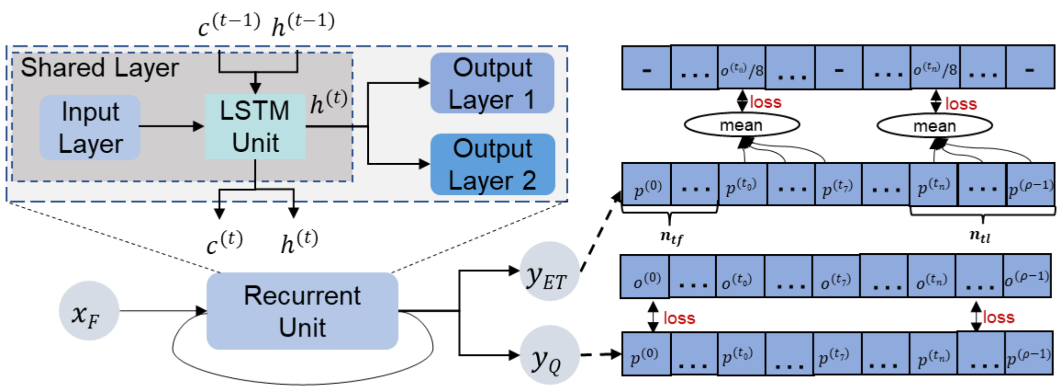


Figure 1. Illustration of the MTL hydrological model. The model inputs, $\boldsymbol{x_F}$, comprise a vector of raw meteorological forcing inputs, and outputs, $\boldsymbol{y_Q}$ and $\boldsymbol{y_{ET}}$ represent the streamflow and evapotranspiration, respectively. The LSTM's internal state in the t-th period is denoted by the cell state, $\boldsymbol{c_t}$ and hidden state, $\boldsymbol{h_t}$. In each period, $\boldsymbol{p_t}$ represents the prediction and $\boldsymbol{o_t}$ is the observed data. The missing data in a given period is indicated by "-". The symbol "mean" enclosed in a circle represents the mean value of selected periods.

Backpropagation in these models allows the independent updating of weight and bias parameters for task-specific output layers, based solely on the losses of the current layer and independent of other output losses. However, updates to the shared LSTM layer parameters depend on multiple outputs. The following equations illustrate these updates:

$$\theta_T(i+1) = \theta_T(i) - \alpha \nabla_{\theta_T} L\big(\theta_T(i), \theta_S(i)\big) \tag{1}$$

$$\theta_S(i+1) = \theta_S(i) - \alpha \sum_{j=1}^{n} \omega_j \left[\nabla_{\theta_S}^{(j)} L\left(\theta_S(i), \theta_T^{(j)}(i)\right)\right] \tag{2}$$

In these equations, θ denotes the weights and biases of the neural networks, with T and S representing the task-specific output and shared layers, respectively. The index I signifies the i-th training step, α denotes the learning rate, ∇ represents the gradient of the loss function relative to the weight parameter, and $L(\cdot)$ is the loss function itself. The j-th specific task is represented by j, ω_j signifies the weights and bias corresponding to the j-th specific task and n stand for the total number of tasks.

One of the challenges of constructing a multi-task learning model is balancing the loss from each task. This balance is crucial to avoid one task from dominating the model training and negatively impacting the learning of other tasks (Vandenhende et al., 2022). The MTL loss function, represented by equation (3), calculates the overall loss value for all tasks, where L_{MTL} signifies the overall loss value for all tasks, L_j represents the loss value of the j-th task, and other variables have the same meaning as in equation (2).

$$L_{\mathrm{MTL}} = \sum_{j=1}^{n} \omega_j \cdot L_j \tag{3}$$

$$\sum_{j=1}^{n} \omega_j = 1 \tag{4}$$

Balancing tasks can be achieved by setting task-specific weights, represented as ω_j, in the loss function. However, quantifying the weight of each task is challenging. Two usual approaches to task balancing exist: the uncertainty weighting method (Cipolla et al., 2018) and dynamic task prioritization (Guo et al., 2018). However, these methods adopt totally different views on the significance of tasks. The former balances task losses by considering homoscedastic uncertainty, assigning lesser weight to outputs with higher uncertainty and consequently higher weight to simpler tasks. But the latter prioritizes the learning of difficult tasks by assigning them higher task-specific weights.

A more direct and simpler approach is manual loss weight assignment, which was also used in some related studies (B. Li et al., 2023; Sadler et al., 2022). This paper defined the loss weight ratio λ as the ratio of evapotranspiration and streamflow variable loss weights, $\frac{\omega_{ET}}{\omega_Q}$. During the training period, multiple λ values were assigned, each corresponding to an MTL model trained for all basins simultaneously. The model demonstrating the best overall prediction performance during the validation period was chosen for testing.

The MTL model was designed to produce daily predictions for both streamflow and evapotranspiration. Although daily streamflow observation data is available, evapotranspiration observation data is cumulative and represents values over an 8-day interval. This interval is adjusted to 5 or 6 days in regular and leap years, respectively, to account for the final period of each year. Therefore, a specific design for the loss function calculation is necessary. As shown in Figure 1, the observed streamflow values were directly compared with predicted values. Meanwhile, predicted ET values were averaged over a period before being used to calculate the loss function. The first n_{tf} or last n_{tl} time-steps of the whole period could begin or end with a duration of less than 8 days. In such situations, we ignored the first n_{tf} non-value time-steps and multiplied the final observed value by $n_{tl}/8$, $n_{tl}/5$, or $n_{tl}/6$, depending on whether the last period was the final period in a regular or leap year. Throughout the model training phase, the root-mean-square error (RMSE) acted as the loss function. This same RMSE metric was applied to calculate the loss functions for each individual output under the MTL mode.

2.3 General settings

All models employed in this paper utilized the same input variables, including 6 meteorological forcing variables and 17 attribute variables pertinent to soil, geology, topography, land use type, and climate. The details are provided in Table 1. One of the distinct advantages of deep learning models is their ability to automatically extract input features from an end-to-end perspective, rather than manually analyzing and extracting features from multiple input variables. Hence, the basin attribute data were directly copied to each period and concatenated with the meteorological input, creating the model's input vector without necessitating manual selection.

The settings for data preprocess and model training aligned with our previous and related research (Ouyang et al., 2021; Rahmani et al., 2021) and were consistently applied to all models, including STL and MTL models, to ensure comparability.

Before model training, normalization of input and output data samples is essential for the efficient optimization of the neural network weight by the gradient descent algorithm during subsequent training. Test data also require normalization, and the statistical data used for normalization during testing is that used for training. After the model completed its predictions, the results are re-normalized back to their original dimension.

Consistent with our previous research, the Adadelta algorithm, an adaptive learning rate scheme (Zeiler, 2012), was chosen as the optimization method for performing stochastic gradient descent on the neural network model parameters. To mitigate overfitting, dropout regularization was implemented during the training of LSTM models. Dropout applies a fixed mask, meaning once a connection weight is set to zero, it stays at zero for the entire training process. The loss function was the root mean square error between the observed and predicted values. The hyperparameter settings of all models in this study were as follows: the mini-batch size was 100, the training sequence length was 365, the number of hidden units per layer was 256, and the LSTM dropout rate was 0.5.

In the evaluation phase, the Nash-Sutcliffe Efficiency (NSE) score was employed to assess streamflow and evapotranspiration prediction. NSE is a metric particularly suited to evaluate hydrological predictions. Additionally, other common metrics, such as the mean difference between modeled and observed values (Bias), RMSE, and Pearson's correlation (Corr), were also used to evaluate the models.

2.4 Experiments

This study devised two experiments to ascertain the conditions under which an MTL model could enhance the simultaneous prediction of each variable compared to STL models. In experiment A, we partitioned the dataset into training, validation, and test sets, with the validation data assisting in finding the optimal multi-task loss weight ratio λ. The evaluation metrics of the STL and MTL models for each output were compared in this experiment. In experiment B, we further investigated the temporal and spatial generalization capabilities of the STL and MTL models using scaling curves to gain a deeper understanding of their differences.

Experiment A: Comparison of STL and MTL models utilizing the entire dataset

We first constructed an MTL model that predicted both streamflow and evapotranspiration. To assess any potential improvement in predictive capability, the performance of this model was compared with that of two STL models; one predicting streamflow and the other predicting evapotranspiration. Notably, the STL model for streamflow did not encompass any input or output associated with evapotranspiration data, and vice versa.

Employing the multi-task balance strategy outlined in Section 2.2, the multi-task loss weight ratio λ was manually assigned. We chose five λ values (2, 1, 1/3, 1/8, and 1/24) to conduct prediction experiments with the MTL model. The optimal model for the testing period was identified by evaluating the NSE values achieved for each variable in the basins during the validation period. The training, validation, and test periods were from 2001-10-01 to 2011-09-30, 2011-10-01 to 2016-09-30, and 2016-10-01 to 2021-09-30, respectively.

Experiment B: Assessment of model temporal and spatial generalization

Generally, supplying more data for DL models often leads to superior model performance. As the number of basins expands, the temporal and spatial generalization of the models usually improve. Scaling curves, which depict the behavior of scaling relative to the amount of training data (Tsai et al., 2021), could be used to analyze how the models behave as the number of trained basins increases. By comparing the STL and MTL models, the conditions under which MTL models outperform STL models could be identified.

For all models, a percentage of basins were randomly chosen for training, with the remaining basins used for temporal and spatial generalization evaluation. We chose 11 percentage values: 5, 10, 20, 25, 33, 50, 66, 75, 80, 90 and 95. To mitigate geospatial bias, we ensured that each case included basins from every LEVEL-II ecoregion (Omernik & Griffith, 2014), rendering them representative of the entire group. When the number of basins was limited, the selection process could introduce bias. Hence, we employed cross-validation to randomly select basins from the entire dataset repeatedly and computed the average median metric value across all cases as the result. The training period was the same as in Experiment A (2001-10-01 to 2011-09-30). No specific validation period was assigned as it was determined based on the best multi-task loss weight ratio λ obtained from Experiment A. The test period was from 2011-10-01 to 2021-09-30.

2.5 Reliability Assessment

We evaluated the reliability of deep learning models by comparing the predictive capabilities of probes in STL and MTL models. Our hypothesis posited that if the probes in MTL models outperformed those in STL models in predicting non-target hydrological variables, then the MTL models were extracting more information, as the probes denoted the LSTM state vector's capacity to predict non-target variables. Such an outcome would further imply that MTL models were effectively identifying more credible correlations between inputs and multiple outputs. As DL models compress input information in their high-dimensional space based on the loss between observations and outputs, having more outputs implied that more information was encoded. Therefore, it was plausible to expect differences in the predictive performance of probes between STL and MTL models.

The implementation of the latent variable's probe is outlined in detail below. We began by training STL or MTL models, then we input the concatenated meteorological forcing data and attributes (XF) from the testing period into the trained models (as depicted in Figure 2). Next, we extracted cell states for all periods and use these to train a linear regression model. This model took 256 units from each sample over each period as input and generated a non-target variable as output. Subsequently, we produced predictions from the probe and compare them with observations. Notably, both the training and testing periods for the probe were included in the testing for both STL and MTL models. For this paper, we adopted a 4:1 ratio for the training-to-testing periods for the probes.

In both STL and MTL models used for streamflow (Q), evapotranspiration (ET), or both, surface soil moisture (SSM) was the non-target variable for STL or MTL models and serves as the target variable for the probe. In the STL model of streamflow, ET was the non-target variable, and for the STL model of ET, Q was the non-target variable. Even though the probe was typically used for non-target variables in deep learning models, we used it to probe both Q and ET in all STL and MTL models to thoroughly examine the differences between STL and MTL models. We evaluated the correlation of the predictive performance of probes in both STL and MTL models. The probes for ET or Q could assist us in understanding how probes behave with target variables. A stronger correlation for the SSM probe in the MTL model could suggest that the implicit information in the MTL model aligns more closely with the actual hydrological process.

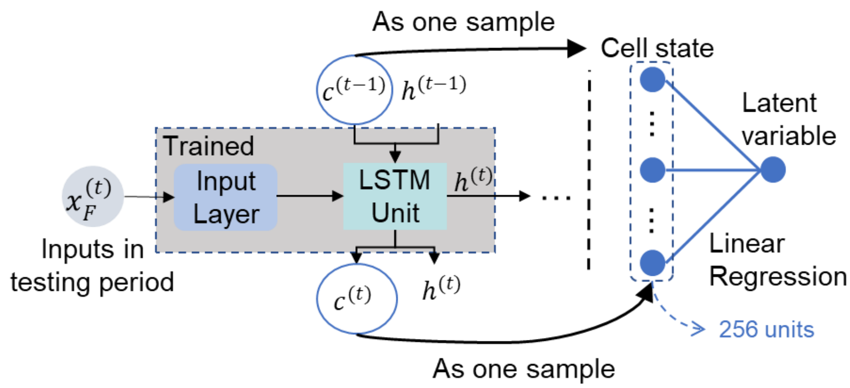


Figure 2. Illustration of the training process of a latent variable's probe. Each cell state in one period is considered a single sample for the linear regressor. The input for the regressor matches the size of the cell state (256 units), while the output size is 1, representing a single latent variable.

3 Results

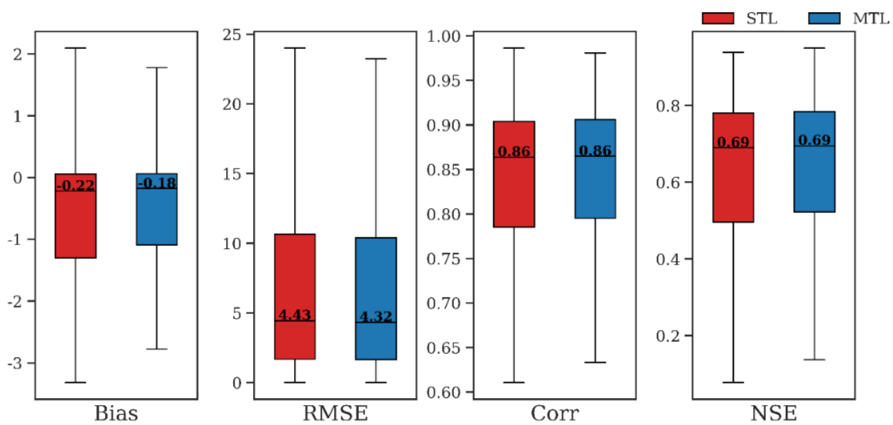
3.1 Prediction performance of MTL and STL models

Upon examining the performance of the two variables during the validation period, λ=1/3 was chosen for the MTL model evaluation in testing period. Further details can be found in Supporting Information Figure S1. In this section, we focus on the results for testing period.

As depicted in Figure 3, the performance metrics of the MTL and STL models for both streamflow and evapotranspiration prediction are relatively similar. For streamflow prediction, the median value of the RMSE of the MTL model is 4.32 (m^3/s), marginally lower than that of the STL model. The Correlation and NSE median values are almost identical to those of the STL model, at 0.86 and 0.69, respectively, albeit with slightly superior upper and lower boundaries of the box plot. The Bias of the MTL model for streamflow prediction is nearer to 0 than that of the STL model, showing an improvement of approximately 18%. The results for evapotranspiration prediction, as displayed in Figure 3(b), follow a similar pattern with the RMSE, Correlation, and NSE of the MTL model being -0.04 (mm/day), 0.96, and 0.92, respectively. These values are equivalent to those of the STL model.

Previous studies calibrating physics-based hydrological models with multiple hydrological output variables (Dembélé, Ceperley, et al., 2020; Dembélé, Hrachowitz, et al., 2020; Tong et al., 2021) found that while using multiple outputs enhanced the simulation accuracy for variables other than streamflow, the accuracy of the streamflow simulation itself decreased. In contrast, our research highlighted that, when looking at the collective performance across all basins, a deep-learning-based MTL model that not only slightly improved the prediction performance of streamflow but also maintained the accuracy for evapotranspiration prediction. Significantly, MTL models could simply consider multiple process and simultaneously output multiple variables. Hence, in scenarios where various processes should be considered, it was more reasonable to construct an MTL model rather than using multiple STL models that only modeled one hydrological output variable at a time.

(a)



(b)

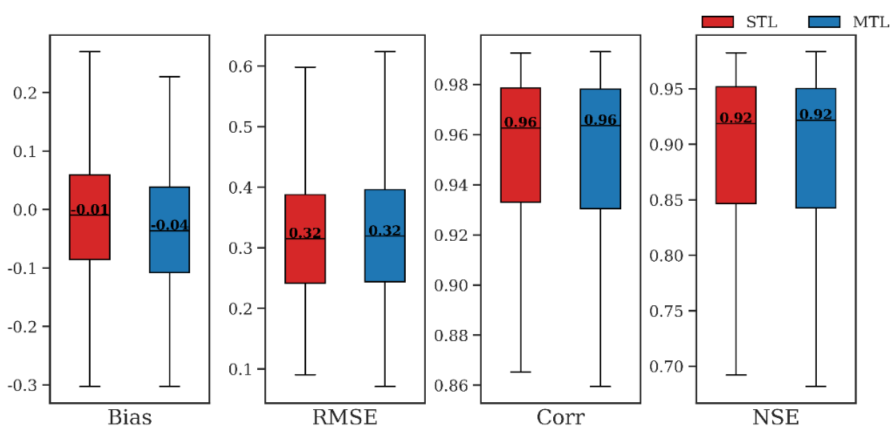


Figure 3. Statistical indicators of the streamflow and evapotranspiration prediction results of each STL model during the testing period and the MTL model under the λ=1/3 scheme, which include Bias, RMSE, Corr, and NSE.

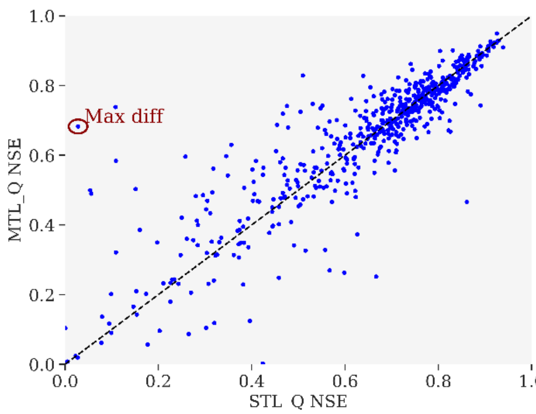
Figure 4 contrasts the performance of STL and MTL models for streamflow and evapotranspiration prediction across various basins. The NSE values for the STL and MTL models vary significantly among different basins. As illustrated in Figure 4(a), the integration of evapotranspiration into the MTL model doesn’t invariably enhance streamflow prediction for all basins. About half of the basins (280) display superior streamflow predictions with the STL model, while the other 311 basins demonstrate improved predictions with the MTL model. Most basins are situated near the 1:1 line of equality, suggesting that the added dimension of evapotranspiration often led to subtle variations in prediction results in most basins. Figure 4(b) reveals that evapotranspiration prediction shows analogous patterns with minor differences between the STL and MTL models across most basins. Some basins demonstrate superior predictions with the MTL model, while others with the STL model.

In some basins, particularly displayed in the top left corner of Figure 4(a), shown in the circle with label “Max diff”, the differences in NSE for streamflow prediction are strikingly large. The MTL model exhibits a considerably higher NSE value for streamflow prediction of 0.68 compared to the STL model's 0.03. However, the STL model performs more effectively in forecasting evapotranspiration in this basin, shown in Figure 4(b) with a circle label, with NSE values for STL and MTL models being 0.76 and 0.72, respectively.

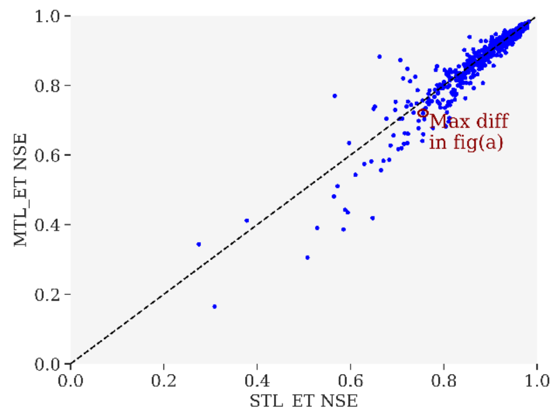
Figure 4(c) and (d) feature a comparison between the streamflow and evapotranspiration prediction results of STL and MTL models, along with the observational data, in the basin where streamflow prediction saw the most significant improvement. The MTL model's streamflow prediction accurately sidesteps unrealistic high flow rates that don't align with observations, thereby enhancing the predictive performance. In terms of evapotranspiration prediction values, the MTL model is slightly lower than the STL model and exhibits limited consistency with the observation values.

These results suggested a competitive relationship among various output variables in MTL, where improving the prediction performance of one variable might lead to a decline in the performance for another. This phenomenon tied back to the concept of attaining a Pareto frontier in multi-objective optimization. Hydrological variables should ideally adhere to a single physical law, regardless of apparent competitive relationships between multiple objectives, such as flood control and power generation in reservoir operations. This competitive relationship indicated the presence of latent variables influencing the formation processes of evapotranspiration and streamflow that the current input failed to capture. Therefore, reaching the Pareto frontier of the MTL process became crucial (Sener & Koltun, 2018).

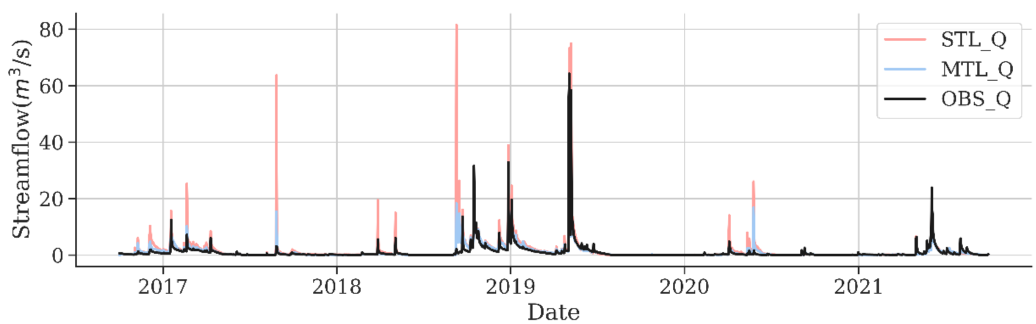
(a)



(b)



(c)



(d)

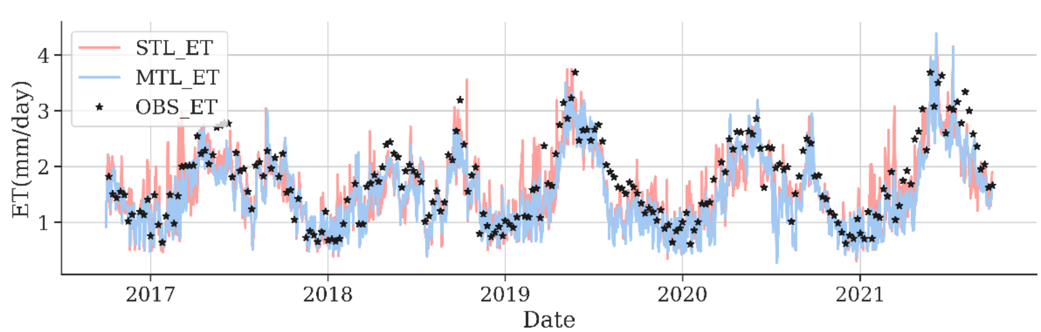


Figure 4. Streamflow and evapotranspiration predictions of STL and MTL models. During the test period in all basins and the time series data for streamflow and evapotranspiration predictions and observations of the STL and MTL models in the basin exhibiting the most considerable improvement in streamflow prediction. The 1:1 line is represented as a black dashed line in both (a) and (b), where points above the line denote a higher NSE using the MTL model compared to the STL model. In (c) and (d), the evapotranspiration observational data is showcased as a scatterplot, with observations gathered at eight-day intervals, while the streamflow observational data and predicted values for streamflow and evapotranspiration variables are provided daily.

In conclusion, the prediction performance of MTL and STL models was generally comparable. In most instances, the prediction of each variable was either nearly equal to or slightly superior to that of the STL models. Under certain loss weight configurations, the MTL model might exhibit marginally superior performance. Furthermore, instead of training and deploying multiple STL models, it was more efficient to select a unified loss weight and utilize a single MTL model to simulate the multi-variate hydrological process and predict multiple outputs concurrently.

3.2 Temporal and spatial generalization of MTL and STL models

We extended the comparison of STL and MTL models to examine how temporal generalizability evolved with an increased number of trained basins and assessed spatial generalizability through a PUB test. Figure 5 depicts the scaling curves of both models. Due to the spatial extrapolation, blue lines in Figures 5 generally display lower NSE values than red lines.

(a) streamflow (b) evapotranspiration

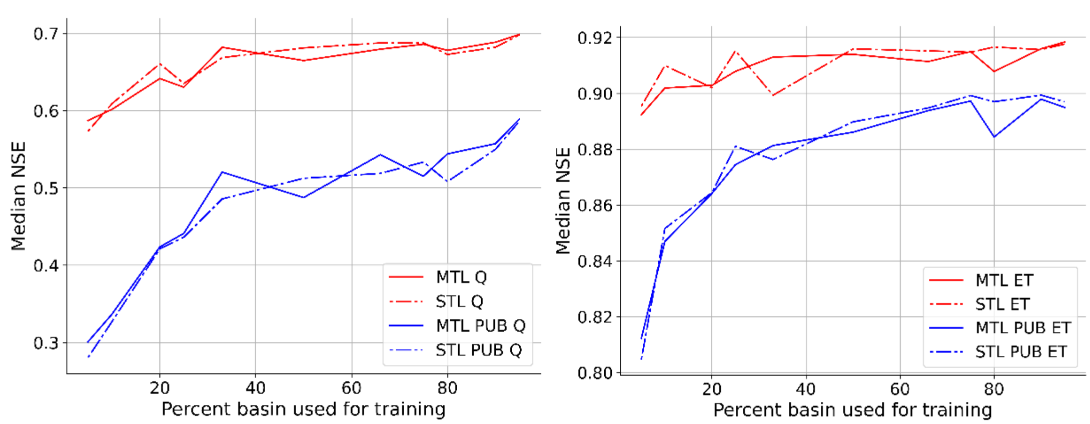


Figure 5. Scaling curves for MTL and STL models. The STL models are indicated by dash lines, while the MTL models are represented by solid lines. Temporal testing results are represented by

red lines and spatial testing results are shown in blue lines. Figures (a) show the predictions of streamflow, while Figure(b) outline the predictions of evapotranspiration. The y-axis signifies the median NSE, reflecting the mean value of median NSEs across all folds in a particular setting. The x-axis represents the percentage of basins used for model training in each setting. We established 11 scenarios for training, which encompass 5%, 10%, 20%, 25%, 33%, 50%, 66%, 75%, 80%, 90% and 95% of the basins.

A consistent trend observed across all subplots was the enhancement in median NSE value with the rising percentage of basins utilized for training, visible in both temporal and spatial generalization tests. This pattern indicated that an augmented dataset enhanced the generalization prowess of DL models in hydrological contexts. Fang et al. (2022) linked this phenomenon to a data synergy effect, suggesting that accumulating and training more heterogeneous data enabled DL models to generate better predictions. Our spatial generalization test affirmed this, highlighting that even in a PUB scenario, the diversity of basins could enhance the prediction accuracy for all hydrological outputs.

Moreover, it became evident that spatial generalizability appeared to improve more markedly than temporal generalizability. For example, considering streamflow (Q) in Figure 5(a), the median NSE values ranged from approximately 0.58 to 0.70 as the basin training percentage progressed from 5% to 95%. In contrast, in PUB contexts, this range expanded from about 0.30 to roughly 0.60, reflecting an enhancement of almost 100%. Similar trends are evident for evapotranspiration (ET) predictions. These observations suggested that heterogeneous data provides more substantial benefits for PUB, whereas local data is generally sufficient for local predictions.

In comparison to STL models, MTL models exhibited varying performances in temporal and spatial generalization tests, predominantly in three distinct patterns. In one scenario, for both streamflow and evapotranspiration, MTL models marginally underperformed. For example, a 2-fold cross-validation (equivalent to training with 50% of the basins) assessing the PUB performance of STL and MTL could prematurely suggest that MTL has weaker spatial generalization capabilities than STL. In another scenario, a trade-off between streamflow and evapotranspiration resulted in MTL outperforming STL for one variable while underperforming for the other, as observed in the 80% training data scenario. Yet, there were instances where MTL models showcased superior prediction, such as when 33% of basins were used for training, outperforming in both variable predictions. Considering we only chose one ratio for MTL's different task loss weight, there should be some randomness in the results, but after comparing MTL with STL in these different scenarios, it could be inferred that MTL model won't be worse in both temporal and spatial generalization than multiple STL models.

3.3 Reliability assessment via analysis of internal states in MTL and STL models

The cell state of the LSTM model serves as a vital tool for retaining and transmitting information throughout time series, encapsulating the long-term dependencies observed in sequences. This characteristic facilitates a deeper understanding of the learning process in hydrological phenomena (Lees et al., 2022). Before diving into the probe analysis, it's helpful to examine the direct correlation of the internal states with outputs. Figure 6 offers an illustrative representation of the correlation coefficients between the hidden layer cell states of the LSTM in the MTL model compared to two STL models, set against observed evapotranspiration data. The LSTM hidden layer comprises 256 cell state units, across 591 basins.

The most prominent correlation between the LSTM cell state and evapotranspiration is observed in the STL model for evapotranspiration (STL-ET), followed closely by the MTL model. Conversely, the STL model for streamflow (STL-Q), which excludes evapotranspiration from its output, exhibits the least correlation. Figure 6(a) indicates that many basins, especially around cell numbers 0, 50, and 100, have correlation coefficients approaching 1 or -1. Meanwhile, Figure 6(b) suggests that the MTL model's LSTM cell state maintains a strong correlation with evapotranspiration, albeit marginally weaker than the STL-ET model. This difference can be attributed to the shared layer structure of the MTL model compared to the specialized nature of the STL-ET model. Figure 6(c) emphasizes the subdued correlation between the STL-Q model's LSTM cell state and evapotranspiration. However, specific cells, such as those between 128 to 132, display discernible correlation patterns.

We also calculated the maximum absolute value of correlation between cell state and observation data of evapotranspiration and get the median value of the maximum values for all basins (median-max-corr). It showed that the values of STL-ET, MTL and STL-Q models are 0.93, 0.93 and 0.85, respectively. Corresponding analyses for the relationship between LSTM cell states and streamflow are presented in Figure S2 in Supporting Information, where the median-max-corr values for streamflow in STL-ET, MTL, and STL-Q models are 0.50, 0.71, and 0.71, respectively.

The shared-layer LSTM in the MTL model adeptly captured the intricacies of both streamflow and evapotranspiration. Although its correlations with individual variables might not match those of specialized STL models, its multi-variable proficiency was commendable. STL models inherently focus on singular variables, but due to the interrelated nature of hydrological components, they might inadvertently capture patterns from non-target variables. In contrast, an MTL model, trained on multiple variables, offered a well-rounded correlation pattern with each. Simply put, while individual models might discern patterns of related variables due to inherent hydrological links, models tailored for multi-variable predictions better comprehend the complex interrelationships, even if their correlations appear slightly less intense than singular-focused models.

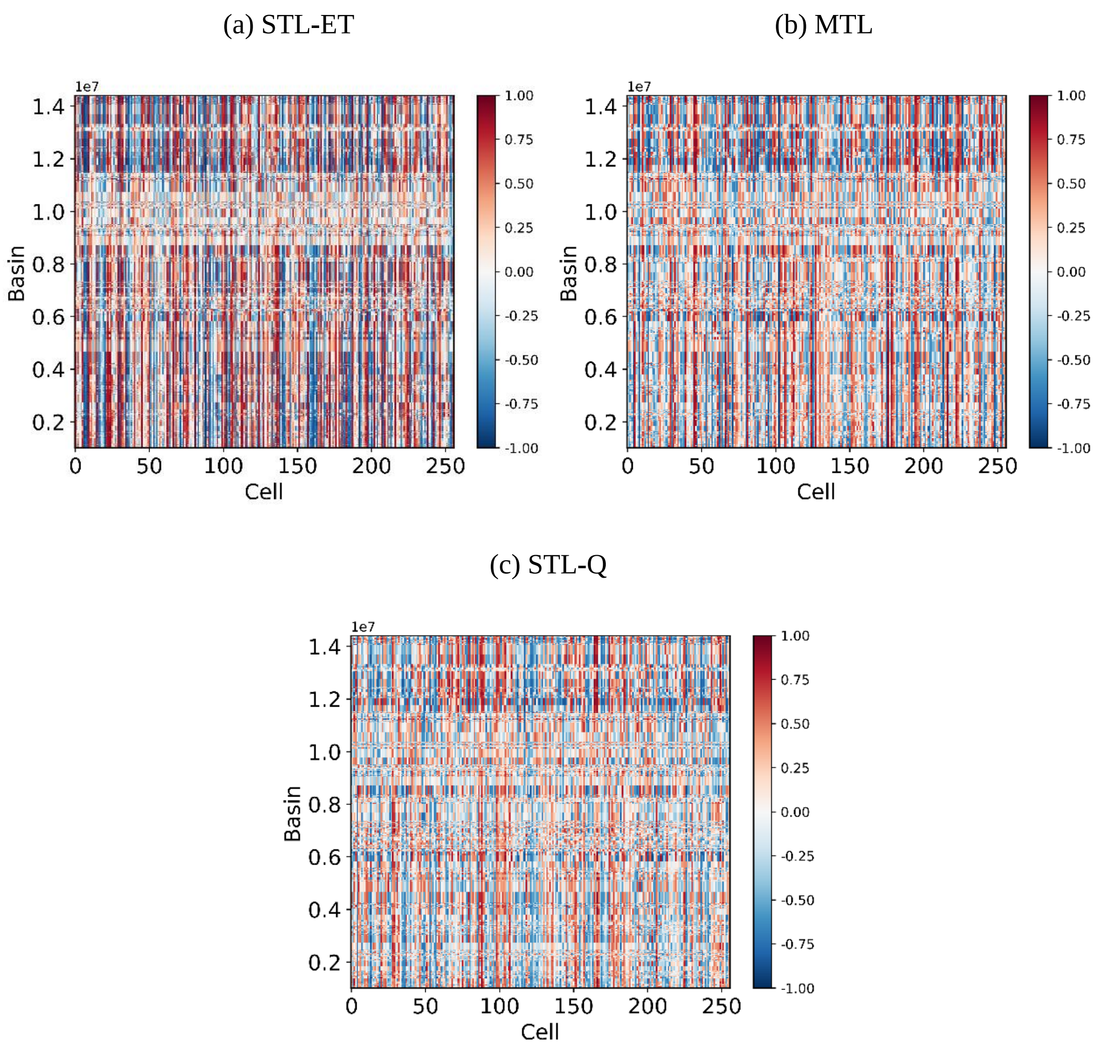
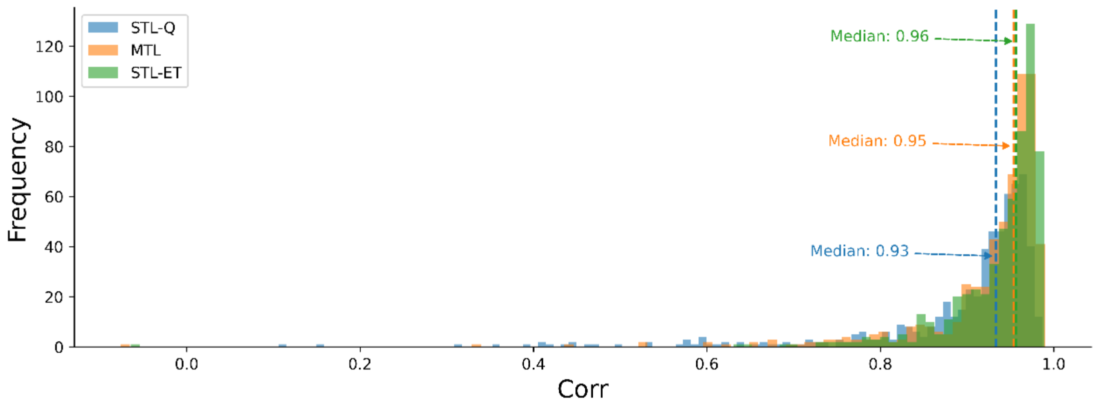


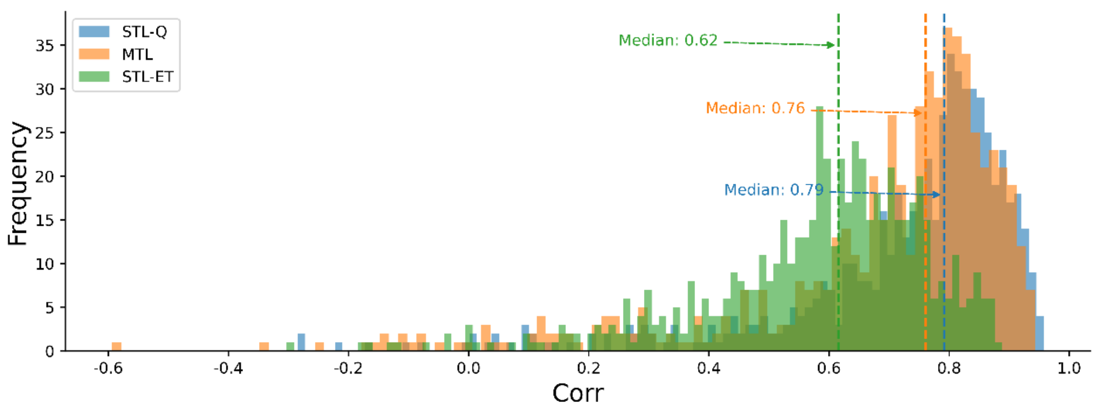
Figure 6. Correlations between the trained LSTM's cell states during the testing period and evapotranspiration in different models for each basin. Panels (a), (b), and (c) correspond to the

STL-ET, MTL, and STL-Q models, respectively. Basins on the y-axis are identified by their 8-digit ID from the CAMELS dataset, where notation such as "1e7" represents "10^7", and "0.2" corresponds to "02000000". The x-axis cell labels represent the index of the cell unit within the LSTM's cell state.

(a) Corr of evapotranspiration probe's prediction



(b) Corr of streamflow probe's prediction



(c) Corr of surface soil moisture probe's prediction

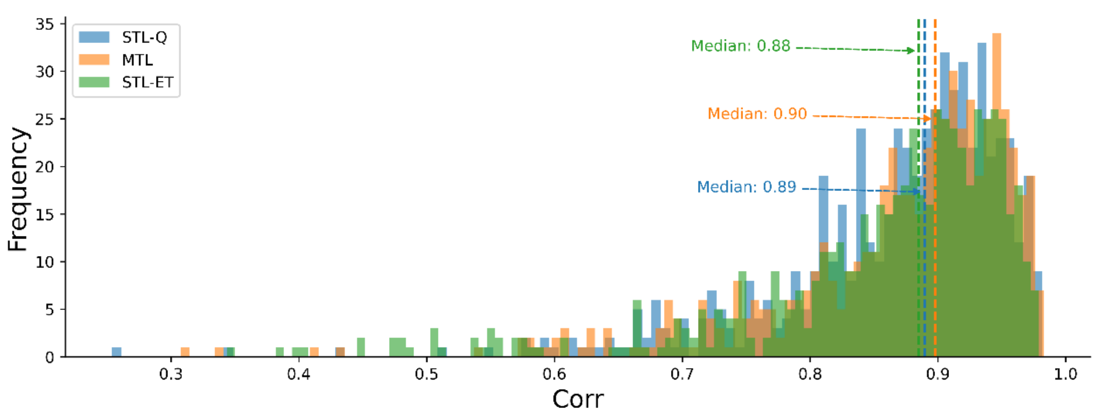


Figure 7. A comparison of the correlation coefficients of (a) evapotranspiration (ET), (b) streamflow (Q), and (c) surface soil moisture (SSM) probes across different models. The blue, orange, and green bars respectively represent STL-Q, MTL, and STL-ET models. Given that the correlation serves as a performance indicator, it assumes only positive values, unlike the correlation between cell states and target variables, which can take negative values.

Figures 7(a) and 7(b) illustrate histograms of prediction correlation coefficients for evapotranspiration and streamflow probes, respectively, across three DL models: STL-Q, MTL, and STL-ET. The median value of 591 basins for the evapotranspiration probe are approximately 0.93, 0.95, and 0.96 for STL-Q, MTL, and STL-ET models, respectively. For the streamflow probe, these values are approximately 0.79, 0.76, and 0.62 across the respective models. Evidently, from the perspective of probe prediction, the LSTM cell state of the STL-ET/STL-Q model exhibits the strongest correlation with evapotranspiration/streamflow, followed by the MTL model. In contrast, the LSTM cell state of the STL-Q/STL-ET model shows the weakest correlation with evapotranspiration/streamflow.

The use of cell states to predict a non-target variable was significantly less effective. While Lees et al. (2022) referred to the performance as “Hydrological Concept Formation” of DL models and deemed it acceptable, in the absence of constraints imposed by multiple outputs, the probe’s performance of the STL model might not match that of the MTL model. One interesting phenomenon was that the highest correlation did not originate from the MTL model. A linear probe finding the correlation between all cell states and the probe's target variable did not equate to the highest correlation from one cell state.

Figure 7(c) illustrates the correlation coefficients between the predicted and observed values obtained from the surface soil moisture probe. The MTL model achieves the highest correlation coefficient for probe prediction results, with a median value for all basins approximating 0.90. In contrast, the STL-Q and STL-ET models yield correlation coefficients of approximately 0.89 and 0.88, respectively. This suggested that the shared LSTM layer in the MTL model, by factoring in input-output correlations for multiple variables, could effectively learn hydrological processes relevant to non-target variables. Combining all these results, we proposed that this layer should not simply be considered a trade-off mechanism for multiple variables. Instead, the learned correlations were more closely aligned with hydrological processes, thereby enhancing the reliability of the MTL model.

Sections 3.1 and 3.2 highlighted that the MTL model, when predicting multiple variables, was not inferior to the two STL models with large datasets. In fact, it might slightly surpass them under specific loss weight ratios. Moreover, the MTL model can output multiple variables concurrently, whereas multiple models would need to be constructed for STL. The results from Figures 6 and 7 indicated that while the shared-layer LSTM in the MTL model effectively learned patterns from multiple variables, its individual correlation with each variable might not be as strong as the dedicated focus each STL model has on its specific target variable. From these findings, we inferred that the shared LSTM layer in the MTL model exceled at discerning input-output relationships across multiple variables and delving into variable-specific input-output correlations. This strengthened the reliability of the correlation rules, a phenomenon we termed the 'variable synergy effect' within the MTL model framework.

4 Discussion

The "variable synergy" effect, inherent to the MTL model, goes beyond just predicting multiple targets within a single framework. Generally, MTL models show generalization capabilities comparable to those achieved by using multiple STL models. The combined layers in the multi-task neural network often yield a more reliable internal representation compared to the STL models. The implications of this synergy effect could be interpreted in some way and found resemblance with the principle of multi-objective optimization (MOP). Such an effect also held the promise to advance hydrological modeling. We would further explore potential improvements from MTL models and address the limitations of this study in the subsequent section.

4.1 Trade-off or synergy with multiple outputs in MTL

Both STL and MTL models employ deep learning as a universal approximator to capture the intricate, high-dimensional relationships between inputs and outputs. However, MTL distinguishes itself by utilizing the relationships between multiple outputs. Through assimilating loss from these interrelated outputs and updating the shared layers of the neural network, MTL can aggregate and leverage shared information across tasks. Then, the internal states of an MTL neural network show a stronger correlation with third-party water balance components, indicating a more comprehensive representation of basin hydrological processes compared to STL models.

In certain MTL studies, the input data itself (Le et al., 2018) or specific noise within the input data (Pironkov et al., 2017) can serve as auxiliary tasks, effectively acting as regularization methods. These can improve the generalization performance of the main target prediction. Hence, in MTL modeling, the inclusion of other tasks can be seen as a regularization method that reduces overfitting.

According to results in Figure 4 in section 3.1, we could find that in MTL modeling, the notion of trade-offs is salient; bolstered prediction performance for one variable might entail minor setbacks for another. This dynamic resembles the trade-offs seen in MOP, typical of reservoir operations. MTL inherently involves an MOP process, and the strategy employed in this study for MOP involves using weights to convert multi-target objectives into single-target objectives. A more nuanced or flexible approach could involve strategies like NSGA-II to create a Pareto front, providing a clear visualization of the competitive relationship between different targets. Yet, for about 10^5 parameters, traditional evolutionary algorithms fall short. This indicates a prospect for investigating MOP-adapted stochastic gradient descent algorithms in upcoming studies.

Interestingly, evapotranspiration (ET) and streamflow (Q) are not as conflicting as water supply and flood control in reservoir operations (Castelletti et al., 2013). This insight offers deep learning researchers a unique lens. They might probe deeper, analyzing and gleaning input data for latent variables from extant hydrological process insights. Gathering these resources could enable a Pareto improvement for the multi-variable learning process, potentially enhancing overall model performance.

4.2 Potential and limitations of MTL models

MTL models prove advantageous in modeling variables that incur significant observational costs, especially when paired with more affordably observed variables within one model. For instance, Surface Soil Moisture (SSM) typically has a shorter observation period compared to streamflow, which prompts the exploration of integrating the longer streamflow data into the multi-task learning model to enhance data effectiveness. This approach explores the leverage of long-term data to forecast short-term data within the MTL paradigm, termed as "data-augmentation with variable synergy". Evapotranspiration was not included in this exploration because, based on our preliminary analysis, its predictive performance was already outstanding, leaving minimal room for enhancement through this technique.

Initially, we pretrained the MTL model using only the streamflow data from 2005-04-01 to 2015-03-31, a period without any SSM data records. This treated the MTL model as an STL model. To ensure the model focused solely on the streamflow, the non-shared fully connected layer dedicated to the SSM task was intentionally ignored, and its associated loss weight was set to zero. Subsequently, the MTL model underwent further training using both streamflow and SSM data from 2015-04-01 to 2020-09-30. We termed model with this training strategy as MTL_Pretrained, as depicted in Figure 8. It was then compared to a standard MTL model, which was trained on both streamflow and SSM data over the same period without any pretraining, as well as the STL model for SSM. The outcomes of these comparisons are presented in Figure 8.

As Figure 8a illustrates, modeling SSM over short durations with limited datasets poses challenges. However, using an MTL modeling framework can significantly improve the prediction of SSM, utilizing the synergy effect from streamflow and SSM. Through pretraining, the LSTM weights and bias are first calibrated guided by streamflow data, circumventing the commencement from entirely random states. This pretrained model, when retrained, could lead to slightly better prediction performance. Therefore, even if there is no observation for the data-scarce variable, it is recommended to use a trained model for another data-rich variable as the pretrained model, rather than random initialization of bias and weights. Figure 8b confirms that the pretrained MTL model outperforms in the majority of basins.

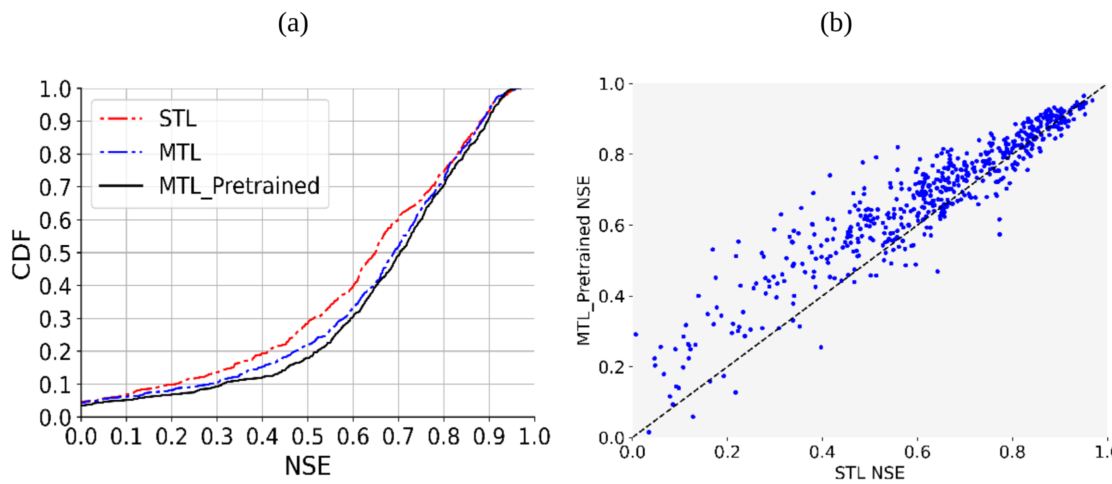


Figure 8. Demonstrating the augment effect for the data-scarce variable from the data-rich variable within the MTL modeling framework. Figure 8a is the empirical cumulative density function plot for three models: an STL of SSM (STL), an MTL for SSM and Q (MTL) and the MTL_Pretrained model. Figure 8b demonstrates the comparison of NSEs between the STL model and the MTL_Pretrained model. The black line represents a 1:1 line. Points above this line indicate that the MTL_Pretrained NSE is superior.

This study proposes an empirical rule that an increased number of observed variables can potentially enhance the prediction of less-observed variables within an MTL model. This finding is particularly beneficial for predicting hydrological variables with fewer observations, such as groundwater streamflow. Under this circumstance, for the high-cost observations, we could use more weak-labeled data such as crowdsource data, they can be involved in the multi-task modeling framework and provide more information to calibrate the model, which is very difficult in traditional modeling methods.

Although multiple output variables can bring about predictive refinements, realizing substantial advancements without additional input remains a challenge. Sadler et al. (2022) suggested that optimizing the loss weight for each variable on a basin-specific basis could further improve the prediction within the MTL model framework, but the observed improvements were still not significant. This limitation stems from reaching a local optimal point in the feasible region of the high-dimensional parameter space without additional information. Since the predictions did not improve considerably, the impact on the long-term water balance was minor, even though more water balance components were included in the outputs.

In hydrological deep learning models, it becomes pivotal to integrate deeper insights into the rainfall-runoff dynamics, like pre-event soil moisture. By integrating this additional information, we can gain a more comprehensive understanding of hydrological processes, which can, in turn, improve the accuracy of predictions.

5 Conclusions

This paper explored the role of multi-task deep learning in hydrological modeling across 591 catchments in the CAMELS dataset, using remote sensing observations of actual evapotranspiration and ground-based streamflow data. An MTL model, rooted in the LSTM neural network architecture, was developed. We evaluated each variable's predictive performance of the MTL model by contrasting it with those of two STL models in terms of both temporal and spatial generalizability. The correlation coefficients between the LSTM cell states of each model and their corresponding output variables were further investigated. Then a surface soil moisture probe, which enabled an examination of the neural network's ability to extract internal representations for the hydrological process was also constructed.

Our findings demonstrate that the MTL model, designed for simultaneous predictions of multiple outputs, consistently matched the performance metrics of its STL counterparts. In contrast, STL models are restricted to predicting a single output variable, limiting their ability to capture associations between hydrologic variables. Moreover, in both temporal and spatial generalization contexts, the MTL model exhibited performance comparable with STL models, regardless of the dataset size. This highlights the robustness of MTL within hydrological modeling frameworks. underscoring the resilience of multi-task learning in hydrological modeling. As a result, the MTL model emerges as a promising deep learning instrument for further hydrological process exploration, and may soon become the preferred approach in hydrological modeling over STL.

Regarding model reliability, the MTL model mines the relevance of multiple variables without a marked bias towards any single target, unlike the STL model. Though the MTL's shared-layer LSTM might have a marginally reduced correlation for individual variables compared to STL models, it still upholds a reasonable correlation with observations for various variables. On the other hand, the STL model's correlation with non-target variable observations is notably weaker. Additionally, the LSTM cell states of the MTL model align more closely with hydrological processes than those of the STL models. A probe designed for SSM using LSTM cell states—excluded from all model training—highlighted a superior prediction correlation in the MTL model. This suggests that MTL models better bridge inputs with multiple outputs, while STL models concentrate mainly on specific target variables.

The MTL model also showcased its potential as a regular deep learning method, especially when faced with limited data observations for certain variables. Its adaptability is particularly beneficial for bridging data gaps. Nevertheless, a deeper exploration into the connection between MTL and multi-objective optimization is required. Leveraging gradient-based multi-objective optimization methods to identify the Pareto frontier could push the frontiers of MTL in hydrological modeling. Another critical challenge for deep learning in hydrology remains the need for comprehensive data. It's crucial to understand that hydrological processes extend beyond just meteorological influences. Incorporating a wider range of ground-based hydrological time-series data, including pre-event soil moisture, can refine our understanding of hydrological patterns, driving more precise predictions. In summary, the future of hydrological modeling will benefit from blending deep learning with multi-objective optimization techniques, leveraging vast and diverse datasets for richer insights.

Acknowledgments

W. Ouyang and L. Ye were primarily supported by the General Program (Grant No. 5217090835) from the National Natural Science Foundation of China (NSFC). W. Ouyang was also supported by the Youth Science Foundation Project (Grant No. 52309010) from NSFC. All authors contributed to writing-reviewing and editing, with primary contributions by W. Ouyang. The contact author has declared that none of the authors has any competing interests.

Data Availability Statement

All data used in this study are available from public sources. The NLDAS-II dataset can be downloaded from the website (http://dx.doi.org/10.5067/THUF4J1RLSYG), which originally obtained the dataset from the NOAA/NCEP; The basin attribute data can be downloaded from the CAMELS website (http://dx.doi.org/10.5065/D6G73C3Q) provided by the U.S. National Center for Atmospheric Research; The SMAP surface soil moisture dataset is available at (https://doi.org/10.5067/ZX7YX2Y2LHEB); The streamflow data can be obtained from USGS

Water Data for the Nation website (http://dx.doi.org/10.5066/F7P55KJN). The code used in this study are available in the open-source repository (https://doi.org/10.5281/zenodo.10024012)

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Bannai, T., Xu, H., Utsumi, N., Koo, E., Lu, K., & Kim, H. (2023). Multi-Task Learning for Simultaneous Retrievals of Passive Microwave Precipitation Estimates and Rain/No-Rain Classification. *Geophysical Research Letters*, *50*(7), e2022GL102283. https://doi.org/10.1029/2022GL102283

Becker, R., Koppa, A., Schulz, S., Usman, M., Beek, T. aus der, & Schüth, C. (2019). Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived ET data. *Journal of Hydrology*, *577*, 123944. https://doi.org/10.1016/j.jhydrol.2019.123944

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017). What do Neural Machine Translation Models Learn about Morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–872. https://doi.org/10.18653/v1/P17-1080

Castelletti, A., Pianosi, F., & Restelli, M. (2013). A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run. *Water Resources Research*, *49*(6), 3476–3486. https://doi.org/10.1002/wrcr.20295

Cipolla, R., Gal, Y., & Kendall, A. (2018). Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7482–7491. https://doi.org/10.1109/CVPR.2018.00781

Dembélé, M., Ceperley, N., Zwart, S. J., Salvadore, E., Mariethoz, G., & Schaefli, B. (2020). Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, *143*, 103667. https://doi.org/10.1016/j.advwatres.2020.103667

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets. *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019WR026085

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology. *Water Resources Research*, *58*(4), e2021WR029583. https://doi.org/10.1029/2021WR029583

Feng, D., Fang, K., & Shen, C. (2020). Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/10.1029/2019WR026793

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Big Remotely Sensed Data: Tools, Applications and Experiences*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Guo, M., Haque, A., Huang, D.-A., Yeung, S., & Fei-Fei, L. (2018, September). Dynamic Task Prioritization for Multitask Learning. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., Anderson, M. C., Sadeghi, A. M., Hain, C. R., & Sharifi, A. (2018). Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology*, *556*, 39–49. https://doi.org/10.1016/j.jhydrol.2017.11.009

Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743. https://doi.org/10.18653/v1/D19-1275

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hu, X., Shi, L., Lin, G., & Lin, L. (2021). Comparison of physical-based, data-driven and hybrid modeling approaches for evapotranspiration estimation. *Journal of Hydrology*, *601*, 126592. https://doi.org/10.1016/j.jhydrol.2021.126592

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning.

Geophysical Research Letters, *47*(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Kraft, B., Jung, M., Körner, M., & Reichstein, M. (2020). Hybrid modeling: Fusion of a deep learning approach and a physics-based model for global hydrological modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B2-2020*, 1537–1544. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

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. *Hydrology and Earth System Sciences*, *25*(5), 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Le, L., Patterson, A., & White, M. (2018). Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information*

Processing Systems (Vol. 31). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/2a38a4a9316c49e5a833517c45d31070-Paper.pdf

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., & Dadson, S. J. (2022). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *26*(12), 3079–3101. https://doi.org/10.5194/hess-26-3079-2022

Li, B., Li, R., Sun, T., Gong, A., Tian, F., Khan, M. Y. A., & Ni, G. (2023). Improving LSTM hydrological modeling with spatiotemporal deep learning and multi-task learning: A case study of three mountainous areas on the Tibetan Plateau. *Journal of Hydrology*, *620*, 129401. https://doi.org/10.1016/j.jhydrol.2023.129401

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., & Kumar, V. (2022). Regionalization in a Global Hydrologic Deep Learning Model: From Physical Descriptors to Random Vectors. *Water Resources Research*, *58*(8), e2021WR031794. https://doi.org/10.1029/2021WR031794

Liu, J., Hughes, D., Rahmani, F., Lawson, K., & Shen, C. (2023). Evaluating a global soil moisture dataset from a multitask model (GSM3 v1.0) with potential applications for crop threats. *Geoscientific Model Development*, *16*(5), 1553–1567. https://doi.org/10.5194/gmd-16-1553-2023

Liu, J., Rahmani, F., Lawson, K., & Shen, C. (2022). A Multiscale Deep Learning Model for Soil Moisture Integrating Satellite and In Situ Data. *Geophysical Research Letters*, *49*(7), e2021GL096847. https://doi.org/10.1029/2021GL096847

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., & Shen, C. (2021). Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions. *Water Resources Research*, *57*(5), e2020WR028600. https://doi.org/10.1029/2020WR028600

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N. E. C., Franz, T. E., Shi, J., Gao, H., & Wood, E. F. (2017). The future of Earth observation in hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3879–3914. https://doi.org/10.5194/hess-21-3879-2017

Mladenova, I. E., Bolten, J. D., Crow, W. T., Sazib, N., Cosh, M. H., Tucker, C. J., & Reynolds, C. (2019). Evaluating the Operational Application of SMAP for Global Agricultural Drought Monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(9), 3387–3397. https://doi.org/10.1109/JSTARS.2019.2923555

NASA. (2018). Forcing Data for Phase 2 of the North American Land Data Assimilation System (NLDAS-2). In *National Aeronautics and Space Administration (NASA)*. https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php#AppendixC

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020WR028091

Omernik, J. M., & Griffith, G. E. (2014). Ecoregions of the conterminous United States: Evolution of a hierarchical spatial framework. *Environmental Management*, *54*(6), 1249–1266. https://doi.org/10.1007/s00267-014-0364-1

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of Hydrology*, *599*, 126455. https://doi.org/10.1016/j.jhydrol.2021.126455

Pironkov, G., Dupont, S., Wood, S. U. N., & Dutoit, T. (2017). Noise and Speech Estimation as Auxiliary Tasks for Robust Speech Recognition. In N. Camelin, Y. Estève, & C. Martín-Vide (Eds.), *Statistical Language and Speech Processing* (pp. 181–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-68456-7_15

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*(2), 024025. https://doi.org/10.1088/1748-9326/abd501

Running, S., Mu, Q., & Zhao, M. (2017). *MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006* [dataset]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD16A2.006

Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., Zwart, J. A., & Kumar, V. (2022). Multi-Task Deep Learning of Daily Streamflow and Water Temperature. *Water Resources Research*, *58*(4), e2021WR030138. https://doi.org/10.1029/2021WR030138

Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany. *Water Resources Research*, *56*(5), e2019WR025924. https://doi.org/10.1029/2019WR025924

Sener, O., & Koltun, V. (2018). Multi-Task Learning as Multi-Objective Optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/432aca3a1e345e339f35a30c8f65edce-Paper.pdf

Shah, S., Duan, Z., Song, X., Li, R., Mao, H., Liu, J., Ma, T., & Wang, M. (2021). Evaluating the added value of multi-variable calibration of SWAT with remotely sensed evapotranspiration data for improving hydrological modeling. *Journal of Hydrology*, *603*, 127046. https://doi.org/10.1016/j.jhydrol.2021.127046

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020). Which Tasks Should Be Learned Together in Multi-task Learning? In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 9120–9132). PMLR. https://proceedings.mlr.press/v119/standley20a.html

Tobin, K. J., & Bennett, M. E. (2017). Constraining SWAT Calibration with Remotely Sensed Evapotranspiration Data. *JAWRA Journal of the American Water Resources Association*, *53*(3), 593–604. https://doi.org/10.1111/1752-1688.12516

Tong, R., Parajka, J., Salentinig, A., Pfeil, I., Komma, J., Széles, B., Kubáň, M., Valent, P., Vreugdenhil, M., Wagner, W., & Blöschl, G. (2021). The value of ASCAT soil moisture and MODIS snow cover data for calibrating a conceptual hydrologic model. *Hydrology* *and Earth System Sciences*, *25*(3), 1389–1410. https://doi.org/10.5194/hess-25-1389-2021

Tong, R., Parajka, J., Széles, B., Greimeister-Pfeil, I., Vreugdenhil, M., Komma, J., Valent, P., & Blöschl, G. (2022). The value of satellite soil moisture and snow cover data for the transfer of hydrological model parameters to ungauged sites. *Hydrology and Earth System Sciences*, *26*(7), 1779–1799. https://doi.org/10.5194/hess-26-1779-2022

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., & Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

USGS. (2019). National Water Information System: Web interface. In *United States Geological Survey*. https://waterdata.usgs.gov/nwis?

Vandenhende, S., Georgoulis, S., De Brabandere, B., & Van Gool, L. (2019). *Branched Multi-Task Networks: Deciding What Layers To Share* (1904.02920v5). arXiv. https://doi.org/10.48550/arXiv.1904.02920

Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., & Van Gool, L. (2022). Multi-Task Learning for Dense Prediction Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3614–3633. https://doi.org/10.1109/TPAMI.2021.3054719

Vu, M. T., & Jardani, A. (2022). Multi-task neural network in hydrological tomography to map the transmissivity and storativity simultaneously: HT-XNET. *Journal of Hydrology*, *612*, 128167. https://doi.org/10.1016/j.jhydrol.2022.128167

Xu, T., Guo, Z., Xia, Y., Ferreira, V. G., Liu, S., Wang, K., Yao, Y., Zhang, X., & Zhao, C. (2019). Evaluation of twelve evapotranspiration products from machine learning, remote sensing and land surface models over conterminous United States. *Journal of Hydrology*, *578*, 124105. https://doi.org/10.1016/j.jhydrol.2019.124105

Yassin, F., Razavi, S., Wheater, H., Sapriza-Azuri, G., Davison, B., & Pietroniro, A. (2017). Enhanced identification of a hydrologic model using streamflow and satellite water storage data: A multicriteria sensitivity analysis and optimization approach. *Hydrological Processes*, *31*(19), 3320–3333. https://doi.org/10.1002/hyp.11267

Yeste, P., Melsen, L. A., & Garc, M. (2023). A Pareto-based sensitivity analysis and multi-objective calibration approach for integrating streamflow and evaporation data. *Water Resources Research*, 1–29. https://doi.org/10.1029/2022WR033235

Yokoo, K., Ishida, K., Ercan, A., Tu, T., Nagasato, T., Kiyama, M., & Amagasaki, M. (2022). Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM. *Science of The Total Environment*, *802*, 149876. https://doi.org/10.1016/j.scitotenv.2021.149876

Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. *arXiv Preprint arXiv:1212.5701*.

Zhang, Y., & Yang, Q. (2022). A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203

Zhao, R.-J. (1992). The Xinanjiang model applied in China. *Journal of Hydrology*, *135*(1), 371–381. https://doi.org/10.1016/0022-1694(92)90096-E

Zhi, W., Ouyang, W., Shen, C., & Li, L. (2023). Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. *Nature Water*, *1*(3), 249–260. https://doi.org/10.1038/s44221-023-00038-z