Supporting Information for "Leveraging Contextual Cues from a Conceptual Model with Predictive Skills of Machine Learning for Improved Predictability and Interpretability in the Hydrological Processes"

Pravin Vasudev Bhasme [1], Udit Bhatia [1]

[1]Civil Engineering Discipline, Indian Institute of Technology Gandhinagar, Palaj, Gandhinagar, 382355, Gujarat, India

Contents of this file

Text S1: Review of the SIMHYD model

The SIMHYD model is lumped conceptual hydrological model that works at daily time-step (Chiew et al., 2002). It is widely applied for various hydrological studies, including hypothesis testing (Vaze et al., 2010), the understanding impact of land-use change on catchment hydrology (Siriwardena et al., 2006), analysis of climate change impact on runoff (Mpelasoka & Chiew, 2009; Chiew et al., 2010), runoff predictions in ungauged catchments (F. Li et al., 2014), analyzing grid-based regionalization in data-sparse region (H. Li & Zhang, 2017). The model consists of seven parameters and requires daily precipitation and potential evapotranspiration (PET) as input. Additionally, two parameters

(DELAY and X) for the Muskingum routing method (McCarthy, 1938) are used. The interception store in the SIMHYD model first intercepts the precipitation (RAIN). The maximum interception (IMAX) (Eq. 1) is the minimum of interception store capacity (INSC) and potential evapotranspiration. Thus, interception (INT) (Eq. 2) will be the minimum of maximum interception and precipitation. The infiltration function handles the precipitation excess of interception. The precipitation that reaches the ground (Eq. 3) that exceeds the infiltration capacity becomes part of streamflow as infiltration excess runoff (IRUN) (Eq. 5). The soil moisture function governed the infiltrated water. It is divided into three parts saturation excess runoff (SRUN) (Eq. 6), soil moisture (SMF) (Eq. 8) in soil moisture store (SMS), and groundwater store (GW) through recharge (REC). The SRUN and REC are linearly dependent on the ratio of SMS and SMSC. The evapotranspiration (ETS) (Eq. 10) from soil moisture store is also a function of the ratio of SMS and soil moisture store capacity (SMSC), but it is limited to the potential rate (POT) (Eq. 9). The actual evapotranspiration (ET) is calculated with the sum of ETS and INT (Zhang et al., 2009). The excess of SMSC joins the GW as a recharge. The baseflow (GD) (Eq. 11) is derived from GW through a linear relationship. The SRUN and IRUN together form direct runoff (DR) (Eq. 12). The GD and DR collectively generate the runoff (Eq. 13). Later this runoff is routed using the Muskingum routing method (Eq. 14 - 17), and the final streamflow (Q) is obtained.

$$IMAX = min\{INSC, PET\} \tag{1}$$

$$INT = min\{IMAX, RAIN\} \tag{2}$$

$$INR = RAIN - INT \tag{3}$$

$$RMO = min\{COEFF * e^{-SQ*SMS/SMSC}, INR\} \tag{4}$$

$$IRUN = INR - RMO \tag{5}$$

$$SRUN = SUB * RMO * SMS/SMSC \tag{6}$$

$$REC = CRAK * (RMO - SRUN) * SMS/SMSC \tag{7}$$

$$SMF = RMO - SRUN - REC \tag{8}$$

$$POT = PET - INT \tag{9}$$

$$ETS = min\{10 * SMS/SMSC, POT\} \tag{10}$$

$$GD = K * GW \tag{11}$$

$$DR = SRUN + IRUN \tag{12}$$

$$RUNOFF = GD + DR \tag{13}$$

$$O_t = C_1 * I_t + C_2 * I_{t-\Delta t} + C_3 * O_{t-\Delta t} \tag{14}$$

$$C_1 = \frac{0.5 * \Delta t - DELAY * x}{(1-x) * DELAY + 0.5 * \Delta t} \tag{15}$$

$$C_2 = \frac{DELAY * x + 0.5 * \Delta t}{(1-x) * DELAY + 0.5 * \Delta t} \tag{16}$$

$$C_3 = \frac{-0.5 * \Delta t + (1-x) * DELAY}{(1-x) * DELAY + 0.5 * \Delta t} \tag{17}$$

Where O_t and I_t are the inflow and outflow at time t. The $DELAY$ and x are the storage constant and dimensionless weighing factor respectively, two parameters used in the Muskingum routing method and C_1, C_2 and C_3 are routing coefficients. The $DELAY$ depicts approximate time taken required for flow travel in the given reach of the river (O'Sullivan et al., 2012).

Text S2: Review of the LSTM model: The Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) is applied widely in time series modeling due to its ability to learn long-term information. It has been applied successfully in various hydrological studies, including post-processing of physics-based model outputs (Frame et al., 2021), prediction of extreme events (Frame et al., 2022), leverage synergy when multiple datasets are used for given variable (Kratzert et al., 2021), flood forecasting (Nevo et al., 2022; Feng et al., 2020), improvement in the streamflow predictions of ungauged basins (Kratzert et al., 2019), streamflow prediction for multiple timescales (Gauch et al., 2021). The LSTM is a special type of Recurrent Neural Network (RNN) in which the vanishing or exploding gradient issue of RNN is solved by incorporating gates and memory cells. The flow of information to the memory cells is controlled by gates. The w_i, w_f, w_c, w_o, U_i, U_f, U_c, and U_o denotes weights associated with the layers and b_i, b_f, b_c, b_o depicts the biases. The forget gate decides the amount of information retained by the cell state. The process of storing new information in the cell state is carried out in two parts, includes information that can be updated in the cell state is decided by the input gate, and the tanh layer generates a new candidate value that is further added to the state then the cell state gets updated. Later, the output gate controls the passage of information from the cell state to the new hidden state, which is obtained by multiplying a *tanh* function of the cell state by the output from the output gate.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{18}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{19}$$

$$\tilde{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{20}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{21}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{22}$$

$$h_t = o_t \times tanh(C_t) \tag{23}$$

Text S3: Performance evaluation metrics: The model performance is evaluated with Nash-Sutcliffe Efficiency (NSE) (Nash & Sutcliffe, 1970), Root Mean Square Error (RMSE), and Percent Bias (PBIAS) widely applied in the field of hydrology (Swain & Patra, 2017; Paul et al., 2019; Wagena et al., 2020). The value of NSE (Eq. 24) ranges from $-\infty$ to 1.0. When NSE is 1, it shows that both simulated and observed data perfectly match each other. The RMSE (Eq. 25) is used to measure the error in the model predictions where its value ranges from 0 to ∞. The PBIAS shows model behavior in estimating the average magnitude of model output. Its optimal value is 0 while having a range of $-\infty$ to ∞. The positive and negative values of PBIAS show underestimation and overestimation of average modeled output, respectively.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \tag{24}$$

where S_i, O_i, and $\bar{O}$ are model output, observed data, and mean of observed data, respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - S_i)^2}{n}} \tag{25}$$

$$PBIAS = \frac{\sum_{i=1}^{n}(O_i - S_i)}{\sum_{i=1}^{n} O_i} \times 100 \tag{26}$$

References

Chiew, F., Kirono, D., Kent, D., Frost, A., Charles, S., Timbal, B., ... Fu, G. (2010). Comparison of runoff modelled using rainfall from different downscaling methods for historical and future climates. *Journal of Hydrology*, *387*(1-2), 10–23.

Chiew, F., Peel, M., Western, A., et al. (2002). Application and testing of the simple rainfall-runoff model simhyd. *Mathematical models of small watershed hydrology and applications*, 335–367.

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.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., . . . Nearing, G. S. (2022). Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, *26*(13), 3377–3392.

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, *25*(4), 2045–2062.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

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.

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.

Li, F., Zhang, Y., Xu, Z., Liu, C., Zhou, Y., & Liu, W. (2014). Runoff predictions in ungauged catchments in southeast tibetan plateau. *Journal of Hydrology*, *511*, 28–38.

Li, H., & Zhang, Y. (2017). Regionalising rainfall-runoff modelling for predicting daily runoff: Comparing gridded spatial proximity and gridded integrated similarity approaches against their lumped counterparts. *Journal of Hydrology*, *550*, 279–293.

McCarthy, G. T. (1938). The unit hydrograph and flood routing. In *proceedings of conference of north atlantic division, us army corps of engineers, 1938* (pp. 608–609).

Mpelasoka, F. S., & Chiew, F. H. (2009). Influence of rainfall scenario construction methods on runoff projections. *Journal of Hydrometeorology*, *10*(5), 1168–1183.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, *10*(3), 282–290.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., ... others (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, *26*(15), 4013–4032.

O'Sullivan, J., Ahilan, S., & Bruen, M. (2012). A modified muskingum routing approach for floodplain flows: theory and practice. *Journal of Hydrology*, *470*, 239–254.

Paul, P. K., Gaur, S., Kumari, B., Panigrahy, N., Mishra, A., & Singh, R. (2019). Diagnosing credibility of a large-scale conceptual hydrological model in simulating streamflow. *Journal of Hydrologic Engineering*, *24*(4), 04019004.

Siriwardena, L., Finlayson, B., & McMahon, T. (2006). The impact of land use change on catchment hydrology in large catchments: The comet river, central queensland, australia. *Journal of Hydrology*, *326*(1-4), 199–214.

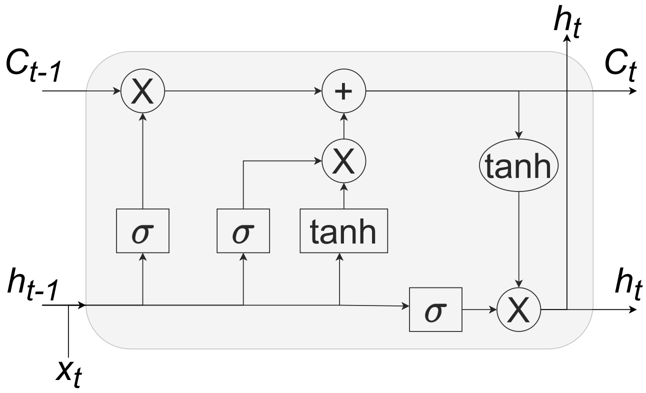

Figure S1. LSTM structure details

Table S1. SIMHYD model parameters for lumped model case.

Catchment	INSC	COEFF	SQ	SMSC	SUB	CRAK	K	DELAY (days)	x
Anandpur	1.033	377.131	2.225	270.96	0.231	0.477	0.031	1.245	0.064
Kantamal	1.282	196.433	2.847	499.201	0.75	1.0	0.003	1.635	0.091
Keesara	0.717	394.968	4.035	489.227	0.38	0.856	0.003	1.762	0.0001

Swain, J. B., & Patra, K. C. (2017). Streamflow estimation in ungauged catchments using regionalization techniques. *Journal of Hydrology*, *554*, 420–433.

Vaze, J., Post, D., Chiew, F., Perraud, J.-M., Viney, N., & Teng, J. (2010). Climate non-stationarity–validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, *394*(3-4), 447–457.

Wagena, M. B., Goering, D., Collick, A. S., Bock, E., Fuka, D. R., Buda, A., & Easton, Z. M. (2020). Comparison of short-term streamflow forecasting using stochastic time series, neural networks, process-based, and bayesian models. *Environmental Modelling & Software*, *126*, 104669.

Zhang, Y., Chiew, F. H., Zhang, L., & Li, H. (2009). Use of remotely sensed actual evapotranspiration to improve rainfall–runoff modeling in southeast australia. *Journal of Hydrometeorology*, *10*(4), 969–980.

Table S2. SIMHYD model parameters for semi-distributed without reservoir model case.

Catchment	Subcatchment	INSC	COEFF	SQ	SMSC	SUB	CRAK	K	DELAY (days)	x
Anandpur	Champua - Anandpur	1.283	106.799	0.545	439.362	0.266	0.932	0.076	0.871	0.149
	Champua	1.283	106.799	0.545	439.362	0.266	0.932	0.076	0.500	0.0004
Kantamal	Kesinga - Kantamal	1.327	134.364	1.624	499.388	0.702	0.955	0.038	0.619	0.188
	Kesinga	1.327	134.364	1.624	499.388	0.702	0.955	0.038	1.185	0.087
Keesara	Madhira - Keesara	1.154	188.348	1.729	355.487	0.363	0.546	0.007	0.999	0.105
	Madhira	1.154	188.348	1.729	355.487	0.363	0.546	0.007	0.543	0.006

Table S3. SIMHYD model parameters for semi-distributed with reservoir model case.

Catchment	Subcatchment	INSC	COEFF	SQ	SMSC	SUB	CRAK	K	DELAY (days)	x	k_r
Brady	d/s of Brady reservoir	1.777	251.159	0.574	276.010	0.011	0.005	0.25	1.289	0.438	-
	Brady reservoir	1.777	251.159	0.574	276.010	0.011	0.005	0.25	0.228	0.462	0.014
Canyon	d/s of Canyon lake	0.841	347.774	1.093	118.446	0.050	0.359	0.003	0.040	0.207	-
	Canyon lake	0.841	347.774	1.093	118.446	0.050	0.359	0.003	1.435	0.359	0.011

Table S4. PIML model hyperparameters for lumped model case.

Catchment	Variable	Dropout rate	Epochs	Units	Batch size	Model
Anandpur	ET_t	0.2	600	100	32	ML - 1
	Q_t	0.2	200	60	32	ML - 2
Kantamal	ET_t	0.1	800	90	64	ML - 1
	Q_t	0.3	300	60	32	ML - 2
Keesara	ET_t	0.1	1000	90	64	ML - 1
	Q_t	0.4	600	100	64	ML - 2

Table S5. ML model hyperparameters for prediction of streamflow in lumped model case.

Catchment	Variable	Dropout rate	Epochs	Units	Batch size
Anandpur	Q_t	0.1	400	90	128
Kantamal	Q_t	0.3	500	10	32
Keesara	Q_t	0.1	400	90	128

Table S6. PIML model hyperparameters for semi-distributed without reservoir model case.

Catchment	Subcatchment	Variable	Dropout rate	Epochs	Units	Batch size	Model
Anandpur	Champua - Anandpur	ETd/s_t	0.3	600	80	32	ML - 2
		Qd/s_t	0.3	900	100	360	$ML - 3$
	Champua	ETu/s_t	0.4	1000	50	32	ML - 1
		Qu/s_t	0.3	900	100	360	$ML - 3$
Kantamal	Kesinga - Kantamal	ETd/s_t	0.1	800	90	32	ML - 2
		Qd/s_t	0.4	300	70	360	$ML - 3$
	Kesinga	ETu/s_t	0.4	1000	40	64	ML - 1
		Qu/s_t	0.4	300	70	360	$ML - 3$
Keesara	Madhira - Keesara	ETd/s_t	0.1	900	100	32	ML - 2
		Qd/s_t	0.2	800	90	360	$ML - 3$
	Madhira	ETu/s_t	0.3	600	100	32	ML - 1
		Qu/s_t	0.2	800	90	360	$ML - 3$

Table S7. PIML model hyperparameters for semi-distributed with reservoir model case.

Catchment	Subcatchment	Variable	Dropout rate	Epochs	Units	Batch size	Model
Brady	d/s of Brady reservoir	ETd/s_t	0.1	1000	90	32	ML – 2
		Qd/s_t	0.3	900	80	256	ML – 5
	Brady reservoir	ETu/s_t	0.2	900	40	32	ML – 1
		R_t	0.4	300	10	256	ML – 3
		S_t	0.4	100	90	360	ML – 4
Canyon	d/s of Canyon lake	ETd/s_t	0.1	500	40	32	ML – 2
		Qd/s_t	0.2	200	90	32	ML – 5
	Canyon lake	ETu/s_t	0.4	200	100	64	ML – 1
		R_t	0.2	1000	90	32	ML – 3
		S_t	0.2	900	90	360	ML – 4