

AI-Augmented Catchment-Scale Hydrological Model Significantly Improves Streamflow Estimation in Arid Regions

M Jawad¹, HV Gupta¹, YH Wang^{1,2}, MA Farmani¹, A Behrangi¹, and GY Niu¹

¹Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA, 85721.

Key Points:

Physically consistent AI model: Combines MCP with a multi-state structure to enforce mass conservation while learning dynamic hydrological processes.

Robust across regimes: Delivers strong and stable streamflow predictions, with major gains in arid and water-limited catchments.

Process & ET matter: Explicit process representation and flexible ET formulations significantly improve accuracy and reliability.

Abstract

Accurate streamflow estimation at the catchment scale remains a persistent challenge, particularly in arid and water-limited regions where hydrological processes are highly nonlinear and difficult to represent. This study presents a novel AI-augmented hydrological modeling framework based on the mass-conserving perceptron (MCP), which integrates physical consistency with data-driven flexibility. The proposed multi-state model explicitly represents key hydrological processes, including canopy interception, soil moisture dynamics, snow accumulation and melt, and baseflow generation, while enforcing mass conservation through adaptive gating mechanisms. The model is evaluated across 15 catchments spanning diverse hydroclimatic regimes and compared against ablated physical models and purely data-driven LSTM benchmarks. Results demonstrate that the proposed framework achieves robust and consistent streamflow predictions, with particularly strong improvements in arid environments. Analysis further highlights the critical role of explicit process representation and flexible evapotranspiration formulations in enhancing model accuracy, variability, and temporal correlation. Overall, the study underscores the value of combining physically grounded modeling with machine learning to improve predictive performance, interpretability, and generalizability in hydrological applications.

1. Introduction and Motivation

The Accurate estimation of streamflow at the catchment scale is a central objective in hydrology, underpinning water resources management, flood forecasting, and climate impact assessments. Despite decades of development, existing modeling approaches continue to face important limitations. Traditional physical and conceptual rainfall–runoff models are grounded in conservation principles and offer interpretability, yet they often rely on simplified representations of complex hydrological processes and require extensive calibration of fixed parameters. These limitations become particularly evident when attempting to represent nonlinear, time-varying interactions among key processes such as precipitation partitioning, evapotranspiration (ET), soil moisture dynamics, snow accumulation and melt, and baseflow generation. On the other hand, recent advances in machine learning, particularly recurrent neural networks such as long short-term memory (LSTM) models, have demonstrated strong predictive skill by learning complex input–output relationships directly from data. However, these purely data-driven models typically lack physical interpretability, may violate mass conservation, and often exhibit reduced robustness and generalizability outside the conditions represented in the training data. This dichotomy between physically based and data-driven approaches highlights the need for modeling frameworks that can simultaneously ensure physical consistency while retaining the flexibility required to capture complex hydrological dynamics.

The motivation for this study arises from the need to address these challenges, with a particular focus on hydrological modeling in arid and water-limited regions, where process representation is especially difficult. In such environments, strong soil moisture constraints, highly variable precipitation, and dominant evapotranspiration losses lead to complex and nonlinear system behavior that is not adequately captured by conventional models. Recent studies have also shown that widely used ET formulations, including remote sensing-based products and traditional Penman–Monteith approaches, tend to introduce systematic biases in dryland regions, further degrading streamflow predictions. To overcome these limitations, this study builds upon the mass-conserving perceptron (MCP) framework, which provides a novel way to embed conservation laws directly within a machine learning architecture. Using this foundation, we develop a multi-state, process-informed catchment-scale model that explicitly represents key hydrological storages and flux pathways, including canopy interception, soil moisture, snow dynamics, and baseflow contributions. The proposed framework allows model parameters and flux-partitioning mechanisms to evolve dynamically as functions of meteorological conditions and system states, enabling a more realistic and adaptive representation of hydrological processes. Furthermore, by systematically evaluating alternative formulations of evapotranspiration and comparing the full model against ablated versions and purely data-driven benchmarks, this work aims to quantify the value of physical process representation in improving predictive accuracy, robustness, and interpretability. Ultimately, the goal is to advance a new class of AI-augmented hydrological models that are both physically grounded and capable of learning complex system behavior, with particular benefits for challenging environments such as arid catchments.

2. Model Architecture and System Equations Based on the Mass-Conserving Perceptron

2.1. The Mass Conserving Perceptron (MCP) Unit and its Desirable Properties

[1] ([Wang & Gupta, 2024a](#)) introduced the mass-conserving perceptron (MCP), a machine learning–based computational unit designed to explicitly enforce physical consistency. The MCP is structurally analogous to a single unit within a gated recurrent neural network, which allows it to embed conservation laws directly into data-driven modeling approaches. At the node level, it describes system dynamics through a discrete-time mass balance:

$$X_{t+1} = X_t + U_t - O_t - L_t$$

[2] In this formulation, the system state at the next time step is determined by adding incoming mass and subtracting outgoing fluxes over the current interval. When applied to lumped catchment-scale rainfall–runoff modeling, the input term represents precipitation, while the output and loss terms correspond to streamflow and evapotranspiration, respectively. These outgoing fluxes are assumed to scale with the current system state according to: $O_t = G_t^O X_t$; $L_t = G_t^L X_t$, where G_t^O and G_t^L are time-dependent gating coefficients that regulate how much mass leaves the system through each pathway. Substituting these relationships into the mass balance yields:

$$X_{t+1} = X_t + U_t - G_t^O X_t - G_t^L X_t$$

which can be rearranged into a more compact form:

$$X_{t+1} = G_t^R X_t + U_t$$

[3] Here, G_t^R denotes a retention (or “remember”) factor that determines the fraction of the state preserved between time steps. These gating coefficients are not fixed; instead, they evolve dynamically in response to the system’s broader hydrologic conditions, such as soil moisture patterns and other

environmental variables. This context dependence allows the model to capture complex, time-varying behavior beyond what static parameterizations can represent. To maintain physical validity, each gating coefficient is constrained to lie within the interval $[0, 1]$. Moreover, conservation of mass requires that all fractions sum to unity: $G_t^R + G_t^O + G_t^L = 1$. This constraint ensures that, at any time step, the entire system state is partitioned into retained, released, or lost components. As a result, the remember gate can be expressed as: $G_t^R = 1 - G_t^O - G_t^L$ which tightly couples the three gating terms. Starting from an initial condition and a sequence of inputs, the system state can be updated iteratively using this framework, provided the time-varying gating functions are specified. Because these functions are generally unknown, [Wang and Gupta \(2024a\)](#) proposed estimating them directly from data using machine learning models, enabling flexible yet physically consistent representations of system dynamics.

[4] Moreover, as described by ([Wang & Gupta, 2024a, 2024b, 2025](#)) the MCP unit possesses several key characteristics that make it well suited for interpretable, physically grounded conceptual modeling of dynamical systems, such as those encountered in hydrology (e.g., rainfall–runoff processes). These include:

- 1) A recurrent structure that captures the temporal evolution of the system state, allowing memory effects to be explicitly represented.
- 2) The ability to enforce conservation laws directly at the node level, ensuring physically consistent system behavior.
- 3) The capacity to account for and learn unobserved mass exchanges, including gains and losses that are not directly measured.
- 4) Flexibility to learn context-dependent process relationships, whereby the gating functions governing system dynamics adapt based on current and antecedent conditions.
- 5) Straightforward implementation and efficient training using gradient-based optimization (backpropagation), supported by standard machine learning frameworks such as PyTorch.

[5] In the following section, we describe how the MCP framework can be used to develop a simplified yet process-representative, physical-conceptual rainfall–runoff model at the catchment scale, incorporating representations of the key processes of the hydrologic cycle including evapotranspiration (ET), snow water equivalent (SWE), and streamflow.

2.2. Model Architecture and System Equations

[6] [Wang and Gupta \(2024b\)](#) proposed the use of up to five state variables to better represent multiple hydrologic storages and capture multi-timescale rainfall–runoff dynamics. Our proposed five-state architecture is structured around three principal land-surface process domains: canopy interception storage, upper soil layer dynamics, and baseflow generation mechanisms.

[7] In the supplementary material section S1 we explain the different components of the model from precipitation partition to intermediate flow dynamics to stream flow generation along with system equations. The model architecture is shown in [Figure 1](#).

Alternate Formulations of ET

[8] In the MA_5 model, evapotranspiration (ET) losses are represented through three state variables, each corresponding to a distinct hydrological component: canopy storage, soil surface storage, and the lower soil bucket. These are denoted generically as X_t , with associated losses L_t representing canopy interception loss, soil evaporation loss, and ET loss from the lower soil layer, respectively. ET loss is formulated as a gated process, where the loss at time t is computed as $L_t = G_t \cdot X_t$, with G_t representing

a nonlinear gating function. In the baseline configuration, this gating function is defined using a sigmoid transformation applied to a combination of meteorological drivers and system states, including air temperature (T), downwelling shortwave radiation (DSW), wind speed (u), vapor pressure deficit (VPD), the slope of the saturation vapor pressure curve (Δ), and the current state X_t .

To evaluate alternative formulations, several variants of the gating function were tested (See [Table 1](#)):

- **S_1 : Penman-Monteith (PM) with traditional resistances:** The gating function incorporates ET estimates derived from the Penman–Monteith (PM) equation, where aerodynamic resistance (r_a) and surface resistance (r_s) are computed using standard empirical formulations (e.g., as described in [Jawad et al. \(2025\)](#)). The sigmoid function is applied to a linear combination of X_t and PM_t^{Trad} .
- **S_2 : Penman-Monteith with neural network-based surface resistance:** This formulation retains the Penman–Monteith structure but replaces the empirical surface resistance (r_s) with a neural network-based estimate such that the NN takes input the following variables: DSW_t , T_t , VPD_t , Δ_t , and u_t . The resulting ET estimate (PM_t^{NN}) is combined with X_t and passed through the sigmoid gating function, allowing for greater flexibility in representing surface controls.
- **S_3 : Direct meteorological forcing formulation:** The gating function is constructed by directly combining the state variable X_t with meteorological inputs, including net DSW_t , T_t , VPD_t , Δ_t , and u_t followed by a sigmoid transformation. This approach removes reliance on intermediate ET equations and directly links environmental drivers to ET response.
- **S_4 : Fully data-driven neural network formulation:** In this approach, all input variables (as stated in S_3) are passed through a feedforward neural network to produce a learned latent representation (FNN_t). This representation is then combined with the state variable X_t and passed through the sigmoid function. This formulation maximizes flexibility by allowing the model to learn complex, nonlinear interactions among inputs without imposing explicit physical structure.

3. Experimental Setup and Methods

3.1. Selection of Study Catchments

[9] For this analysis, we applied the behavioral classification approach introduced by Jiang et al. (2022) to identify 15 representative catchments from the CAMELS dataset. The selected set spans a diverse range of hydro-meteorological conditions and runoff generation processes (see [Table 2](#) and [Figure 2](#)). Of these, twelve catchments fall into four primary streamflow response categories: (i) systems in the western United States dominated by recent precipitation, where discharge mainly reflects rainfall from the same or previous day; (ii) similar rainfall-driven systems located in the eastern United States; (iii) systems influenced by antecedent precipitation, where streamflow integrates rainfall over periods from several days to weeks; and (iv) snowmelt-controlled systems, where temperature plays the dominant role in regulating discharge. The remaining three catchments correspond to arid regions (aridity index $AI > 1$), allowing the analysis to include water-limited conditions.

3.2. Input Forcing and Output (Target) Data

[10] To force the models, we used the latest near-surface atmospheric forcing dataset produced by the *National Oceanic and Atmospheric Administration* (NOAA), known as the *Analysis of Record for Calibration* (AORC). This dataset provides hourly meteorological variables at a spatial resolution of 800m ([Fall et al., 2023](#)). The variables—including air temperature, precipitation, incoming shortwave and

longwave radiation, surface pressure, wind speed, and specific humidity—were obtained from AORC version 1.1 (October 1, 2003, to September 30, 2023) and aggregated to daily values to match the temporal resolution required by the models. Leaf Area Index (LAI) was taken from the MODIS MOD15A2H (Level 4), a global dataset with 500 m spatial resolution and an 8-day temporal frequency (Myneni et al., 2015). This product has been shown to provide reliable LAI estimates across a wide range of vegetation types (Yan et al. (2016)). Surface emissivity data, available at 1 km resolution and 8-day intervals, were obtained from the MODIS MOD21A2 (Hulley, 2017). In addition, hourly albedo data at 9 km resolution were sourced from the ERA5-Land reanalysis (Muñoz-Sabater et al., 2021), developed by the European Centre for Medium-Range Weather Forecasts. All variables were spatially averaged over each catchment prior to analysis.

Observed streamflow, used as the target variable, was obtained for all 15 catchment outlets from the *United States Geological Survey* (USGS) via its online data portal (USGS, 2025). These records are reported in cubic feet per second (ft³/s) as mean daily discharge at each gauging station. For modeling purposes, discharge was converted to millimeters per day by transforming flow into daily volume and normalizing by the corresponding catchment area, resulting in area-standardized streamflow suitable for analysis.

3.3. Metrics Used for Training and Performance Assessment

[11] The metric used in this study is *Jawad-Kling-Gupta Efficiency* ($JKGE_{ss}$) developed by Jawad et al. (2026) wherein they explicitly accounted for non-stationarity in the mean of streamflow generating process as opposed to *Nash–Sutcliffe Efficiency* (Nash & Sutcliffe, 1970) and classical *Kling–Gupta Efficiency* (Gupta et al., 2009; Gupta & Kling, 2011) or in particular the skill-score adjusted version of the latter (KGE_{ss} ; (Knoben et al., 2019)). The metric ($JKGE_{ss}$) has four components M, M^*, V^*, C^* (Eqns 1-5)

$$JKGE_{ss} = 1 - \sqrt{(M + M^* + V^* + R^*)/2} \quad (\text{Eq. 1})$$

$$M = (1 - \beta_{so})^2 \quad (\text{Eq. 2})$$

$$M^* = \frac{1}{N} \sum_{t=1}^N \left(1 - \frac{b_t^s}{b_t^o}\right)^2 \quad (\text{Eq. 3})$$

$$V^* = (1 - \alpha_{so}^*)^2 \quad (\text{Eq. 4})$$

$$C^* = (1 - \rho_{so}^*)^2 \quad (\text{Eq. 5})$$

Where

- M is time-stationary long-term mass balance term: $\beta_{so} = \frac{\mu_s}{\mu_o}$ such that μ_s and μ_o are long-term means of simulated and observed timeseries,
- M^* time non-stationary mass balance term: $b_t^o = f_b(t|\bar{O})$ represents a time-varying benchmark timeseries value obtained by applying some pre-specified functional operation, represented symbolically by $f_b(\cdot)$, to the observed signal $\bar{O}$ and the corresponding model-simulated analog of this observed benchmark time-series can be computed as $b_t^s = f_b(t|\bar{S})$ by applying the same functional operation $f_b(\cdot)$ to the simulated signal $\bar{S}$.
- V^* is variance term of the metric such that $\alpha_{so}^* = \frac{\psi_s}{\psi_o} \rightarrow \psi_s = \sqrt{\frac{1}{N} \sum_{t=1}^N (s_t - b_t^s)^2}$ and $\psi_o = \sqrt{\frac{1}{N} \sum_{t=1}^N (o_t - b_t^o)^2}$ where ψ_o represents the standard deviation of the benchmark-centered “observed anomaly” time-series (obtained by subtracting the time-non-stationary benchmark) and ψ_s represent the analogous standard deviation of the “observed anomaly” time-series.

- C^* is the cross correlation term where $\rho_{so}^* = \frac{1}{N} \sum_{t=1}^N \left(\frac{s_t - b_t^s}{\psi_s} \right) \cdot \left(\frac{o_t - b_t^o}{\psi_o} \right)$.

[12] To select a benchmark timeseries (b_t^o) they proposed through their detailed analysis of CONUS wide catchments belonging to different hydro-climatic regimes to use the benchmark value b_t^o at each time step as being a “sectional mean” of some fixed length N_s such that $b_t^o = \frac{1}{N_s} \sum_{i \in G(t)} O_i$ where $G(t)$ indicates the non-overlapping section to which the particular time step t belongs, and O_i indicates the observed timeseries within that section and hence for the simulated signal $b_t^s = \frac{1}{N_s} \sum_{i \in G(t)} S_i$ and their recommended value of $N_s=30$ days which was used in this study. Therefore, $JKGE_{ss}$ metric was used to train and evaluate all the models used in this study.

3.4. Benchmark Models

3.4.1. Ablated Versions of Physical-Conceptual Model (MA5)

[13] We evaluated three of ablated versions of the MA5 model to assess the contribution of individual physical processes to overall model performance (Figure 3). Specifically, key components of the full physical-conceptual structure are selectively removed, and the resulting degradation in hydrological performance is analyzed across multiple catchments. This approach allows us to quantify the relative importance of different process representations and to identify which mechanisms are most critical for maintaining robust and consistent model behavior under varying hydro-climatic conditions.

- MA₂**: This is a two-state architecture that can be interpreted as representing the temporal dynamics of snow water equivalent ($X_t^{S_{snow}}$) and soil storage (X_t^S). The model incorporates three parallel flow pathways that collectively contribute to the total system outflow (O_t): one for infiltration excess ($P_t^{S2Excess}$), another for soil saturation excess ($O1_t^S$), and the third one for state dependent outflow flux ($O2_t^S$). Conceptually, this architecture captures the most basic key hydrological processes of snow accumulation/melt and soil water dynamics, while allowing runoff generation through both infiltration-driven and saturation-driven mechanisms.
- MA₃^{NC}**: This three-state architecture is built upon MA₂ by adding a baseflow storage (X_t^B) and a seepage flux that connects the soil state (X_t^S) to the baseflow state (X_t^B) through a one-directional flow path. This additional pathway enables a clearer separation between the quicker surface and near-surface flow runoff processes and slower subsurface drainage. The inclusion of baseflow storage allows MA₃ to better represent delayed flow contributions and to improve hydrograph recession behavior, while still maintaining a relatively parsimonious representation.
- MA₃^{NS}**: This three-state architecture is built upon MA₃ by removing snow state variable (X_t^S) and adding canopy water storage state (X_t^C). This modification alters the internal process representation by excluding snow accumulation and melt dynamics, while introducing canopy interception and storage processes. The architecture is implemented to systematically evaluate how the absence of an explicit snow component affects model behavior and predictive performance, particularly in terms of its ability to capture seasonal variability, runoff generation, and overall hydrologic response across different catchment conditions.

3.4.2. Purely ML Based Upper Benchmark Models

[14] To assess and compare the performance of the newly developed AICRR (MA_5) model against purely machine learning (ML) models, we employed long short-term memory (LSTM) networks of varying complexity following the work by (Wang & Gupta, 2024a, 2024b). LSTM networks are a gated form of recurrent neural networks (RNNs) and have been shown to be highly effective for data-driven modeling of dynamical systems (Kratzert et al., 2018). Their strength lies in the cell-state structure, which is regulated by three context-dependent gating mechanisms that enable learning of long-term dependencies (Hochreiter & Schmidhuber, 1997). Single-layer LSTM architectures with 2 to 5 hidden nodes were tested in this study.

3.5. Data Splitting Method used for Model Development

[15] Robust model development requires dividing the available dataset D into separate subsets for calibration and evaluation, ensuring that model evaluation is conducted on independent data. Following the approach introduced by (Zheng et al., 2022), the 20-year record was partitioned into training (D_{Train}) and evaluation (D_{Eval}) datasets using a year-based stratified sampling strategy. In this approach, all years were first ranked according to their total annual streamflow. The most extreme years—representing the wettest and driest conditions—were initially assigned to the training set, while the next pair of extreme years was allocated to the evaluation set. After removing these selected years, the procedure was repeated iteratively until all years were assigned. Ultimately, 60% of the year pairs (12 water years) were included in D_{Train} , with the remaining 40% assigned to D_{Eval} . This stratification method ensures that both subsets encompass a comparable spread of hydroclimatic variability, thereby reducing the risk of bias and improving the representativeness of model training and evaluation.

3.6. Hyperparameter and Training Procedure

[16] The training procedure for both physical–conceptual and machine learning models generally followed the framework outlined by Wang and Gupta (2024a). To mitigate sensitivity to uncertain initial conditions in physical–conceptual models, a three-year spin-up period was introduced by repeating the first water year (WY 2004) three times at the start of the simulation. This approach reduces the impact of incorrect state initialization, as discussed by De la Fuente et al. (2023). All models were implemented in PyTorch, with parameter optimization carried out using the Adam optimizer (Kingma, 2014). The loss function and corresponding gradients were computed over the training subset. For each catchment, and for every model architecture, models were trained using 10 different random seeds to account for variability in parameter initialization. Each training run was conducted for 1500 epochs using full-batch updates. A learning rate of 1.0×10^{-1} was adopted, and all trainable parameters were initialized from a zero-mean standard normal distribution ($\theta \sim \mathcal{N}(0,1)$) to encourage stable optimization in the early stages of training. The final model configuration was selected based on the seed that achieved the best average loss across both D_{Train} and D_{Eval} . Each model architecture was trained separately for all 15 catchments, allowing for catchment-specific characterization of input–state–output relationships. This design enabled a systematic evaluation of model performance and sensitivity across architectures of varying complexity.

4. Experimental Results

4.1. Performance Assessment of MA_5

4.1.1. Performance Assessment via Evaluation Metrics

[17] **Figure 4** presents the performance of the MA_5 model across five hydroclimatic regimes using box-and-whisker plots. It can be clearly seen that model performance varies across the regimes, with consistently higher and less variable values observed for the snowmelt- and historical rainfall-dominated catchments compared to the rainfall-driven regimes. Overall, the model performs best in snowmelt-dominated catchments and poorest in arid catchments. The figure consists of multiple panels, each representing a different performance metric, including JKGEss, KGEss, M' (long-term mass balance), $M^{*'} (time non-stationary mass balance), $V^{*'} (variability), and ρ^* (cross-correlation). Within each panel, five box plots correspond to the five hydroclimatic regimes using the color codes defined in **Figure 2**. Each box plot summarizes the distribution of yearly metric values computed across all water years for the catchments belonging to that regime.$$

- **JKGEss:** exhibits notable variation across the five hydroclimatic regimes. The highest median values are observed for snowmelt-dominated catchments, followed by the historical rainfall regime, indicating relatively better model performance in these regions. In contrast, the arid regime shows the lowest median values, while the recent rainfall regimes (RR (W) and RR (E)) display intermediate medians with greater dispersion. The spread of values is also smallest for snowmelt catchments and largest for the arid regime, which is further characterized by several outliers, reflecting higher uncertainty and reduced model skill. The maximum yearly JKGE_{ss} values are comparable across most regimes, reaching approximately 0.83, whereas slightly lower peak values (around 0.78) are observed for the arid regime.
- **KGEss:** A similar pattern is observed for KGEss, where snowmelt and historical rainfall regimes demonstrate higher median values and relatively compact interquartile ranges, suggesting stable performance. In contrast, RR (W) and RR (E) exhibit moderate medians with wider dispersion, indicating spatial variability in performance. The arid regime again shows the lowest median and the greatest spread, reflecting inconsistent model behavior under arid conditions. However, it can also be seen that KGEss values are clearly higher than JKGEss values for all categories.
- **M' :** For the long-term mass balance component (M'), the median values remain consistently high across all hydroclimatic regimes, accompanied by relatively narrow interquartile ranges, indicating that the model effectively captures the overall water balance irrespective of hydroclimatic conditions. The limited spread and the presence of only a few lower outliers further suggest robust and stable performance for this component. Among the regimes, the recent rainfall (eastern), snowmelt-dominated, and historical rainfall-dominated catchments exhibit slightly higher performance compared to the arid and recent rainfall (western) regimes.
- **$M^{*'}:$** The time non-stationary mass balance component ($M^{*'} (M') exhibits broadly similar behavior across all hydroclimatic regimes, with comparable median values and variability. A slightly improved performance is observed for snowmelt-dominated catchments; however, differences among regimes remain relatively small. Overall, the results suggest that the model is able to capture the non-stationary characteristics of the water balance reasonably well across all hydroclimatic conditions, with only minor variations in performance.$
- **$V^{*'} and ρ^* :$** The variability component ($V^{*'} (V^{*'} and cross-correlation term (ρ^*) further highlight differences in model performance across regimes. For $V^{*'} (V^{*'} , higher median values and tighter distributions are observed in snowmelt and historical rainfall regimes, whereas arid and recent rainfall regimes show lower medians and larger variability, indicating difficulty in capturing flow variability. Similarly, ρ^* shows the highest median and least spread for snowmelt catchments, reflecting strong temporal agreement, while reduced medians and increased dispersion in RR (W), RR (E), and arid regimes suggest weaker correlation and timing errors in model simulations.$$

4.1.2. Performance in Reproduction of Streamflow Hydrographs

[18] We further examined the model performance by visually examining the streamflow hydrographs (Figure 5). For this purpose, we selected five representative catchments, each corresponding to a distinct hydroclimatic regime, for both wet (panel a) and dry (panel b) years. Overall, the model demonstrates a strong ability to capture the timing and magnitude of streamflow dynamics across most regimes, with notable differences in performance depending on hydroclimatic conditions. In wet years, the model accurately reproduces peak flows and seasonal patterns in snowmelt-dominated (black) and historical rainfall-dominated (green) catchments, with simulated flows closely matching observations in both magnitude and timing. The recent rainfall regimes, particularly RR (East) (blue), also show reasonable agreement, although some deviations are observed during high-flow events. In contrast, the RR (West) (red) and arid (orange) catchments exhibit some discrepancies, especially in peak flow representation, indicating challenges in capturing rapid runoff responses.

[19] During dry years, the model performance remains comparable to that observed in wet years across most hydroclimatic regimes, indicating its robustness in reproducing streamflow dynamics under varying conditions. The snowmelt- and historical rainfall-dominated catchments continue to show good agreement between observed and simulated flows in terms of both timing and magnitude. Similarly, the recent rainfall regimes maintain a reasonable level of accuracy, although minor deviations persist during certain flow events. However, in arid catchments, some peak flows are overestimated, and discrepancies in representing low-flow conditions are more evident. Despite these limitations, the overall performance during dry years remains consistent with that observed during wet years, highlighting the model's ability to generalize across contrasting hydrological conditions.

4.1.3. Functional Forms of Activation Gating Functions

[20] The model architecture MA5 contains various gating functions to regulate the fluxes from different state variables. In Figure 6, we plot the functional forms of these gates:

- $G^{S_{melt}}$ (Panel a): Controls snowmelt activation as a function of temperature and incoming shortwave radiation (DSWR). All three snowmelt dominated catchments show a sharp sigmoid monotonic increasing response to temperature as compared to DSWR, showing steeper, threshold-like activation near 0-5 °C.
- $G^{f_{rain}}$ (Panel b): Controls the fraction of precipitation falling as rain based on air temperature. Rainfall-dominated catchments (East) show an earlier (-5 °C) and smoother transition, indicating rain occurs even at relatively low temperatures. In contrast, snowmelt-dominated basins exhibit a delayed (around -2 °C) and steeper switch, implying a narrow temperature band where phase change occurs. Arid systems with snowfall tend to resemble rainfall-dominated behavior but with slightly sharper thresholds.
- $G^{S_{Excess}}$ (Panel c): The softplus type infiltration excess gate which operates when the effective precipitation P^{S3} exceeds a threshold limit (P_{max}). It regulates how excess water contributes to runoff/storage. The thin lines are shown of individual catchments, and thick lines show average of function across each category. Wet (historical rainfall) catchments activate early and rise smoothly, meaning even small excess generates response. Arid catchments show delayed, more gradual increases, requiring larger excess before activation. Snowmelt-dominated basins have the slowest rise, indicating buffering/storage effects.
- G^{S01} (Panel d): The softplus type saturation excess gate which operates when the state variable X^S exceeds a threshold limit (S_{max}). The thin lines are shown of individual catchments, and thick

lines show average of function across each category. Arid catchments show early but smoother activation, suggesting quick response once minimal storage is reached. Rainfall-dominated catchments (West/East) display steeper mid-range increases, indicating strong nonlinear response once soils wet up. Snowmelt-dominated systems activate later and more gradually, reflecting higher storage capacity and slower release.

- **G^{S02} (Panel e):** Represents state dependent outflow flux, which increases as the state X^S increases. Rainfall-dominated (West) catchments show a sharp, threshold-like increase, suggesting activation beyond a certain storage level. Snowmelt-dominated basins increase more gradually, while arid and historical rainfall systems remain low and slowly varying, indicating weaker sensitivity.
- **G^{S03} (Panel f):** Represents seepage from upper soil storage to lower soil storage as a function of upper soil moisture (X^S). The gate is sigmoidal nonlinear, increasing from near zero at low storage to higher values as the soil wets up. Snowmelt-dominated catchments and rainfall-dominated systems (West) activate earlier (lower storage thresholds) with a steeper rise, indicating quicker vertical connectivity. Arid catchments often show limited and gradual activation, suggesting weaker or more restricted seepage pathways.
- **G^{B0} (Panel g):** Controls baseflow (groundwater outflow) as a function of deeper storage (X^B). Snowmelt-dominated basins show a steady, near-linear increase, reflecting sustained groundwater contribution. Arid catchments exhibit a delayed but eventually stronger nonlinear rise, while rainfall-dominated systems (West) remain lower and increase more gradually, indicating weaker baseflow response. Historical rainfall catchments show moderate, smooth growth.

4.1.4. Dominant Flow Path:

[21] There are four primary flow pathways through which water contributes to total streamflow (O_t), as expressed in Eq. X. These pathways collectively represent different runoff generation mechanisms. The first component, $P_t^{S3\text{Excess}}$, corresponds to infiltration-excess runoff, generated when rainfall intensity exceeds the soil's infiltration capacity. The second component, $O1_t^S$, represents saturation-excess runoff, which occurs when the soil becomes fully saturated and any additional input contributes directly to streamflow. The third term, $O2_t^S$, is a state-dependent outflow flux from the soil, capturing lateral or near-surface flow processes that depend on the current soil moisture storage. Finally, O_t^B denotes baseflow, which originates from deeper groundwater storage and contributes more gradually to streamflow.

[22] While all four pathways contribute to total streamflow, their relative importance varies across catchments. In some systems, fast-response mechanisms such as infiltration- or saturation-excess runoff dominate, particularly in wetter or more responsive landscapes. In contrast, other catchments, especially those with higher storage capacity or permeability, may be dominated by slower pathways such as state-dependent subsurface flow or baseflow. This variability reflects differences in hydroclimatic conditions, soil properties, and catchment structure.

Figure 7 illustrates the partitioning of streamflow into different flow pathways for five representative catchments under (a) wet years and (b) dry years, along with (c) the frequency of activation of each pathway across all 15 catchments.

[23] In panels (a) and (b), the stacked hydrographs show the temporal contribution of the four flow components to total streamflow. In recent rainfall dominated and historical rainfall dominated catchments, infiltration-excess runoff ($P_t^{S3\text{Excess}}$, magenta) dominates during high-flow events, particularly in wet years, where sharp peaks indicate rapid runoff generation driven by intense precipitation. In

contrast, during dry years, peak magnitudes are reduced and events are less frequent, leading to a more subdued contribution from this pathway. In snowmelt dominated and arid catchment this pathway was not adopted.

[24] The saturation-excess runoff ($O1_t^S$, cyan) contributes significantly during moderate flow conditions, especially in wetter catchments where soil saturation is more frequently reached. This component shows a more sustained contribution compared to infiltration-excess runoff, particularly during prolonged wet periods. Again, this pathway was not dominant in snowmelt dominated catchment.

[25] The state-dependent soil outflow ($O2_t^S$, green) exhibits a delayed and smoother response, dominating the recession limbs of hydrographs. This behavior is especially evident in snowmelt-dominated and wetter systems, where stored soil water is gradually released over time.

[26] The baseflow component (O_t^B , brown) remains relatively small but persistent, providing a continuous background contribution to streamflow. Its role becomes more apparent during low-flow periods and in dry years, where it sustains flow in the absence of significant precipitation inputs.

[27] Panel (c) summarizes the frequency (percent of timesteps) of activation of each pathway, defined as occurrences exceeding 0.01 mm over 7300 timesteps (20 years). The results show that baseflow and soil outflow pathways (O_t^B and $O2_t^S$) are active for a large fraction of time, particularly in wetter and snowmelt-dominated catchments, indicating their role as persistent background processes. In contrast, infiltration-excess runoff ($P_t^{S3Excess}$) is activated less frequently but contributes disproportionately during peak events, especially in rainfall-dominated systems. Saturation-excess runoff ($O1_t^S$) shows intermediate activation frequencies, with higher occurrence in wetter regimes and lower in arid catchments.

4.1.5. Comparison of Model Generated Snow Water Equivalent (SWE) with UA-SWE:

[28] We compared the model-generated snow water equivalent (SWE) with an independent observational dataset to evaluate the model's ability to represent snow accumulation and melt dynamics. In the model, SWE is computed as the sum of snow water stored at the canopy and soil levels: $SWE_t = X_t^{C_{snow}} + X_t^{S_{snow}}$ where $X_t^{C_{snow}}$ represents snow stored in the canopy, and $X_t^{S_{snow}}$ represents snow stored at the ground/soil surface.

[29] The analysis includes catchments where snowfall contributes either as a primary or secondary source of precipitation. Among the selected catchments, three are classified as snowmelt-dominated systems, where seasonal snow accumulation and melt strongly control hydrological response. The remaining catchments, including arid systems, receive snowfall intermittently and exhibit weaker snow influence. For evaluation, we used the University of Arizona Snow Water Equivalent (UA-SWE; [Broxton et al. \(2019\)](#)) dataset, a gridded SWE product derived from remote sensing and data assimilation techniques, providing spatially and temporally continuous estimates of snow water storage. This dataset is widely used for large-scale hydrological and climate studies due to its consistency and coverage.

[30] **Figure 8** shows a comparison between model-simulated SWE (blue solid lines) and UA-SWE (red dashed lines) for five representative catchments for seven water years (WY-2004 to WY-2010). The top three panels correspond to snowmelt-dominated catchments, while the bottom two panels represent arid catchments with intermittent snow occurrence.

[31] Overall, the model demonstrates strong skill in reproducing SWE dynamics, despite SWE being an untrained variable in the modeling framework. The model is able to capture the seasonal accumulation and melt patterns, as well as the general timing and magnitude of SWE across different catchment types. Some discrepancies (in arid catchments) in peak values and melt rates are observed, which may arise from uncertainties in precipitation phase partitioning, temperature thresholds controlling melt, and limitations in representing spatial heterogeneity within the catchments. Nonetheless, the results highlight the

model's ability to realistically represent SWE dynamics across a wide range of hydroclimatic conditions, even without direct calibration to SWE observations.

4.1.6. Impact of Various ET Formulation on Model Performance

[32] We evaluated the impact of different evapotranspiration (ET) formulations S_1 to S_4 (Section 2.3) on model performance in simulating streamflow across 15 catchments representing diverse hydroclimatic regimes (Figure 9). Model performance is assessed using yearly boxplots of the five metrics KGE_{ss} , $JKGE_{ss}$, M' (long-term mass balance), M^* (time non-stationary mass balance), V^* (variability), and ρ^* (cross-correlation).

- In rainfall-dominated systems (RR-W and RR-E), all formulations perform similarly across all metrics. The median values of KGE_{ss} and $JKGE_{ss}$ remain consistently high, with only minor variations between formulations. The bias-related metrics (M' , M^*) are generally close to optimal values, while variability (V^*) and correlation (ρ^*) are also well captured. This indicates that improvements from S_1 to S_4 are marginal, and all formulations are able to reproduce streamflow dynamics with comparable skill.
- In snowmelt-dominated catchments, the median performance remains nearly unchanged across all formulations for all five metrics. However, a clear improvement is observed in the lower range of performance, with minimum values increasing from S_1 to S_4 . This trend is also reflected across the individual components, where variability and correlation show fewer low-performing cases in the more flexible formulations. Overall, while the central tendency remains stable, model robustness improves with increasing formulation complexity.
- In historical rainfall-dominated catchments, S_1 shows noticeably poorer performance compared to the other formulations. This is evident across all metrics, including lower median KGE_{ss} and $JKGE_{ss}$, as well as larger spread in bias and variability components. In contrast, S_2 , S_3 , and S_4 show consistently higher median values and reduced spread, indicating improved and more stable performance.
- The largest differences are observed in arid catchments, where performance improves substantially from S_1 to S_4 . For S_1 , KGE_{ss} and $JKGE_{ss}$ values are generally low with large variability, indicating poor and inconsistent performance. As we move toward S_4 , there is a clear increase in median values along with a reduction in spread. Improvements are also evident in bias, variability, and correlation metrics, suggesting a more balanced and consistent representation of streamflow. This highlights the strong sensitivity of arid systems to ET formulation.

[33] It can also be observed that the traditional PM-based formulation (S_1) consistently performs the worst across all hydroclimatic regimes, with particularly poor performance in arid and historical rainfall-dominated systems. In contrast, the more flexible formulations show clear improvements. Formulations S_3 and S_4 perform very similarly across most metrics and catchment types, indicating that both direct incorporation of meteorological drivers and fully data-driven approaches yield comparable gains in performance. Furthermore, replacing the empirical formulation of surface resistance (r_s) with a neural network-based function in S_2 leads to a noticeable improvement over S_1 . This suggests that empirical resistance parameterizations limit model performance, and introducing flexibility in representing surface controls on evapotranspiration enhances overall model skill. Overall, these results highlight that increasing flexibility in ET representation, whether through modified physical formulations or data-driven approaches, results in consistent improvements in model performance across hydroclimatic regimes.

4.2. Comparison of MA₅ with Ablated Versions and Pure Data Based Models

[34] The model was also compared with its ablated versions and purely ML-based models to check how the newly proposed model architecture performs in comparison with its ablated version if we remove some processes as explained in section 3.4.1.

4.2.1. Comparison using Metrics Box Plots

[35] **Figure 10** compares the performance of the full MA₅ model with its ablated versions (MA_2 , MA_3^{NC} , MA_3^{NS}) and purely data-driven models (LSTM variants) across different hydroclimatic regimes using multiple evaluation metrics. Across all catchments, the full MA₅ architecture consistently shows strong and stable performance, with high median values and relatively low spread across all metrics. This indicates that the inclusion of all key physical processes including canopy processes, snow dynamics, soil storage, seepage, and baseflow contributes to a balanced and robust representation of hydrological behavior.

[36] The ablated models show clear degradation in performance, highlighting the importance of the removed processes. The simplest architecture, MA_2 , generally exhibits lower median performance and larger variability, particularly in snowmelt-dominated and arid catchments. This suggests that excluding baseflow and deeper storage processes limits the model's ability to capture delayed flow dynamics and recession behavior. However, its performance remains comparable to the full model in rainfall-dominated catchments, where fast runoff processes dominate and the role of deeper storage is less critical.

[37] The addition of baseflow storage in MA_3^{NC} leads to noticeable improvements over MA_2 , particularly in terms of variability and correlation, indicating a better representation of slow-flow components. However, performance still remains slightly below that of MA₅ in arid and snowmelt-dominated catchments, although it is largely comparable. This suggests that while incorporating baseflow improves model behavior, additional processes are required to achieve optimal performance across all hydroclimatic regimes.

[38] In contrast, MA_3^{NS} , which excludes the snow state, shows substantial performance degradation in snowmelt-dominated catchments, with lower median values and larger spread across all metrics. This clearly demonstrates the critical role of explicitly representing snow accumulation and melt processes in such regions. In rainfall-dominated and arid systems, the impact of removing the snow component is less severe but still noticeable.

[39] The purely data-driven models (LSTM variants) show mixed performance. In recent rainfall- and historical rainfall-dominated catchments, they achieve performance slightly better than MA₅ in terms of median values of metrics. However, they exhibit larger variability and lower minimum values, indicating less robustness and consistency across arid and snowmelt dominated catchments, where physical constraints appear to provide important guidance to the model.

[40] Across hydroclimatic regimes, snowmelt-dominated catchments show the strongest sensitivity to model structure, with clear performance drops when snow processes are excluded. Arid catchments also show strong sensitivity, particularly to the inclusion of baseflow and soil-atmosphere interactions, leading to improved performance in MA₅ compared to simpler models. In contrast, rainfall-dominated catchments show relatively smaller differences across model structures, although MA₅ still provides more consistent performance.

[41] Overall, the figure demonstrates that including key physical processes in the MA₅ architecture leads to improved accuracy, stability, and robustness, while ablation of these processes results in systematic performance degradation. Additionally, while purely data-driven models can achieve competitive performance in some cases, they lack the consistency and generalizability provided by the physically informed structure of MA₅.

4.2.2. Comparison using Streamflow Hydrographs

[42] We also evaluated the performance of the MA_5 model and its ablated versions in reproducing streamflow hydrographs (Figure 11) across five catchments spanning distinct hydroclimatic regimes, considering both wet (a) and dry (b) years for each system.

[43] In the recent rainfall-dominated catchment (14222500), model performance is strong during wet conditions, with all architectures capturing the timing and magnitude of peaks reasonably well. However, even in wet years, MA_2 exhibits faster recession and underestimation of low flows due to the absence of a baseflow component. These deficiencies become substantially more pronounced during dry years, where MA_2 fails to sustain inter-event flows, leading to near-zero discharge (in log-space). MA_5 consistently provides the best representation across both conditions, while MA_3^{NC} and MA_3^{NS} perform similarly, indicating a limited role of canopy and snow processes relative to groundwater storage in this regime.

[44] In the snowmelt-dominated catchment (9035900), MA_5 accurately captures the seasonal hydrograph in both wet and dry years, including the timing and magnitude of melt-driven peaks. MA_2 , MA_3^{NC} perform comparably, confirming the secondary importance of baseflow, and canopy storage. In contrast, MA_3^{NS} shows substantial errors in both years, failing to reproduce the snowmelt dynamics altogether. This catchment clearly highlights the combined importance of both snow and groundwater processes.

[45] For the historically rainfall-dominated catchment (3331500), all models reproduce the general hydrograph structure in wet years, with MA_5 providing the closest agreement across peaks and recessions. MA_3^{NC} and MA_3^{NS} show only minor deviations, suggesting that neither canopy nor snow processes dominate system behavior. However, differences become clearer in dry years, where MA_2 fails to maintain baseflow and shows rapid depletion of flow, while MA_3 variants retain more realistic low-flow dynamics due to the inclusion of groundwater storage. MA_5 again, demonstrates the most consistent performance across both conditions.

[46] In the rainfall-runoff eastern regime (1667500), characterized by flashy and frequent events, MA_5 best captures the variability and magnitude of peaks in both wet and dry years. During wet conditions, ablated models show increasing smoothing of peaks, with MA_2 exhibiting the largest attenuation. In dry years, the absence of a baseflow tank in MA_2 leads to pronounced underestimation of low flows and loss of inter-event persistence. MA_3^{NC} and MA_3^{NS} perform better due to groundwater contributions, though they still underestimate variability compared to MA_5 , indicating the importance of full process coupling.

[47] In the arid catchment (13161500), differences among models are most pronounced. During wet years, MA_5 captures both episodic peaks and recession behavior more accurately, while MA_2 shows rapid post-event decline. In dry years, when flows are predominantly sustained by storage, MA_2 fails almost entirely to represent low-flow persistence, producing near-zero discharge for extended periods. In contrast, MA_3^{NC} and MA_3^{NS} maintain some level of baseflow due to the groundwater tank, though still with reduced accuracy compared to MA_5 . This regime most clearly demonstrates the critical role of baseflow storage under water-limited conditions.

[48] Overall, analyzing wet and dry years jointly within each catchment reveals that while simplified model structures may perform adequately under high-flow (wet) conditions, their deficiencies become evident under dry conditions, particularly in representing flow persistence and recession dynamics. The presence of a baseflow tank in MA_3 and MA_5 is essential for maintaining realistic low flows across all regimes, while snow representation is critical in snowmelt-dominated systems. The full model (MA_5) consistently provides the most robust and transferable performance across both hydroclimatic conditions and regimes.

4.2.3. Comparison using Flow Duration Curves (FDCs)

[49] We also evaluated the performance of the four model architectures using flow duration curves (FDCs) (Figure 12). Panel (a) shows the high-flow segment (exceedance probability, EP = 0-5%), panel (b) presents the full FDC (EP = 0-100%), and panel (c) focuses on the low-flow segment (EP = 80-100%), allowing a detailed assessment of model behavior across the entire flow regime.

[50] For high flows (panel a), all models generally reproduce the overall shape of the observed FDCs (red dashed line) across catchments, indicating their ability to capture the upper tail behavior. However, systematic differences are evident. The full model (MA_5) consistently provides the closest agreement with observations, particularly in flashy and recent rainfall-dominated catchments (14222500 and 1667500), where it better captures the steepness and curvature of the high-flow tail. The ablated models show progressive deviations: MA_3^{NC} performs comparably to MA_5 in most cases, suggesting a limited role of canopy storage in extreme flows, whereas MA_3^{NS} significantly underestimates high flows in the snowmelt-dominated catchment (9035900), reflecting the absence of snow processes. MA_2 shows mixed behavior, often underestimating the highest flows or producing overly steep declines, indicating limitations in representing rapid runoff generation with a simplified structure.

[51] Across the full FDCs (panel b), differences among model structures become more pronounced, particularly in the mid- and low-flow ranges. MA_5 consistently tracks the observed curves across all exceedance probabilities, indicating a balanced representation of both fast and slow flow components. MA_3^{NC} and MA_3^{NS} generally follow similar trends but exhibit biases depending on the dominant processes: MA_3^{NS} deviates substantially in the snowmelt-dominated catchment, while MA_3^{NC} shows minor differences in rainfall-driven systems. In contrast, MA_2 systematically departs from observations in the lower half of the FDC, with steeper slopes and earlier drop-offs, reflecting its inability to sustain flows due to the absence of a baseflow storage component.

[52] The low-flow segment (panel c) highlights the most significant structural differences among models. MA_5 provides the best agreement with observed low flows across all catchments, maintaining realistic curvature and magnitude in the high exceedance range. MA_3^{NC} and MA_3^{NS} also perform reasonably well, benefiting from the inclusion of a groundwater/baseflow tank that supports flow persistence. However, MA_3^{NS} shows additional bias in snow-influenced catchments, where missing snow dynamics propagate into low-flow errors. In contrast, MA_2 exhibits a pronounced underestimation of low flows across all regimes, with a sharp downturn in the tail of the FDC. This behavior is particularly evident in arid (13161500) and rainfall-driven catchments, where sustained baseflow dominates streamflow during dry periods. The absence of a baseflow reservoir in MA_2 leads to rapid depletion of storage and an inability to reproduce observed flow persistence.

Conclusions

[53] This study demonstrates that integrating physical principles with data-driven learning through the mass-conserving perceptron (MCP) framework provides a powerful and balanced approach for catchment-scale hydrological modeling. The proposed multi-state AI-augmented model consistently achieves strong performance across diverse hydroclimatic regimes, with particularly notable improvements in challenging environments such as arid catchments. By explicitly enforcing mass conservation and embedding key process representations—including canopy interception, soil moisture dynamics, snow processes, and baseflow generation—the model not only improves predictive accuracy but also enhances robustness and interpretability relative to purely data-driven approaches. Systematic comparisons with ablated model structures highlight the critical role of individual processes, especially baseflow and snow dynamics, in sustaining realistic streamflow behavior across varying conditions.

Furthermore, the evaluation of alternative evapotranspiration formulations reveals that increasing flexibility—particularly through hybrid or fully data-driven representations—leads to substantial gains in model performance, especially in water-limited regions where traditional formulations tend to fail. Overall, the results underscore the value of combining physical consistency with machine learning flexibility, advancing a new generation of hydrological models that are both scientifically interpretable and operationally reliable.

References

- Broxton, P., Zeng, X., & Dawson, N. (2019). *Daily 4 km Gridded SWE and Snow Depth from Assimilated In-Situ and Modeled Data over the Conterminous US, Version 1* (NASA National Snow and Ice Data Center Distributed Active Archive Center. <https://doi.org/10.5067/0GGPB220EX6A>
- De la Fuente, L. A., Gupta, H. V., & Condon, L. E. (2023). Toward a multi-representational approach to prediction and understanding, in support of discovery in hydrology. *Water Resources Research*, 59(1), e2021WR031548.
- Fall, G., Kitzmiller, D., Pavlovic, S., Zhang, Z., Patrick, N., St. Laurent, M., Trypaluk, C., Wu, W., & Miller, D. (2023). The Office of Water Prediction's Analysis of Record for Calibration, version 1.1: Dataset description and precipitation evaluation. *JAWRA Journal of the American Water Resources Association*, 59(6), 1246-1272.
- Gupta, H. V., & Kling, H. (2011). On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research*, 47(10).
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2), 80-91.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Hulley, G. (2017). MYD21A2 MODIS/Aqua Land Surface Temperature/3-Band Emissivity 8-Day L3 Global 1km SIN Grid V006. *NASA EOSDIS Land Processes DAAC: Oak Ridge, TN, USA*.
- Jawad, M., Behrang, A., Farmani, M. A., Qiu, Y., Sohi, H. Y., Gupta, A., & Niu, G.-Y. (2025). Improved evapotranspiration estimation using the Penman-Monteith equation with a deep learning (DNN) model over the dry southwestern US: Comparison with ECOSTRESS, MODIS, and OpenET. *Journal of Hydrology*, 133460.
- Jiang, S., Zheng, Y., Wang, C., & Babovic, V. (2022). Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resources Research*, 58(1), e2021WR030185.
- Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10), 4323-4331.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005-6022.
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., & Hersbach, H. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9), 4349-4383.
- Myneni, R., Knyazikhin, Y., & Park, T. (2015). MOD15A2H MODIS/Terra leaf area Index/FPAR 8-Day L4 global 500m SIN grid V006. *NASA EOSDIS Land Processes DAAC*.
- 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.
- USGS. (2025). *United States Geological Survey - Daily Streamflow Data*. United States Geological Survey. <https://waterdata.usgs.gov/>
- Wang, Y. H., & Gupta, H. V. (2024a). A mass-conserving-perceptron for machine-learning-based modeling of geoscientific systems. *Water Resources Research*, 60(4), e2023WR036461.
- Wang, Y. H., & Gupta, H. V. (2024b). Towards interpretable physical-conceptual catchment-scale hydrological modeling using the mass-conserving-perceptron. *Water Resources Research*, 60(10), e2024WR037224.
- Wang, Y. H., & Gupta, H. V. (2025). Using machine learning to discover parsimonious and physically-interpretable representations of catchment-scale rainfall-runoff dynamics. *Water Resources Research*, 61(12), e2025WR040178.
- Yan, K., Park, T., Yan, G., Chen, C., Yang, B., Liu, Z., Nemani, R. R., Knyazikhin, Y., & Myneni, R. B. (2016). Evaluation of MODIS LAI/FPAR product collection 6. Part 1: Consistency and improvements. *Remote Sensing*, 8(5), 359.
- Zheng, F., Chen, J., Maier, H. R., & Gupta, H. (2022). Achieving robust and transferable performance for conservation-based models of dynamical physical systems. *Water Resources Research*, 58(5), e2021WR031818.

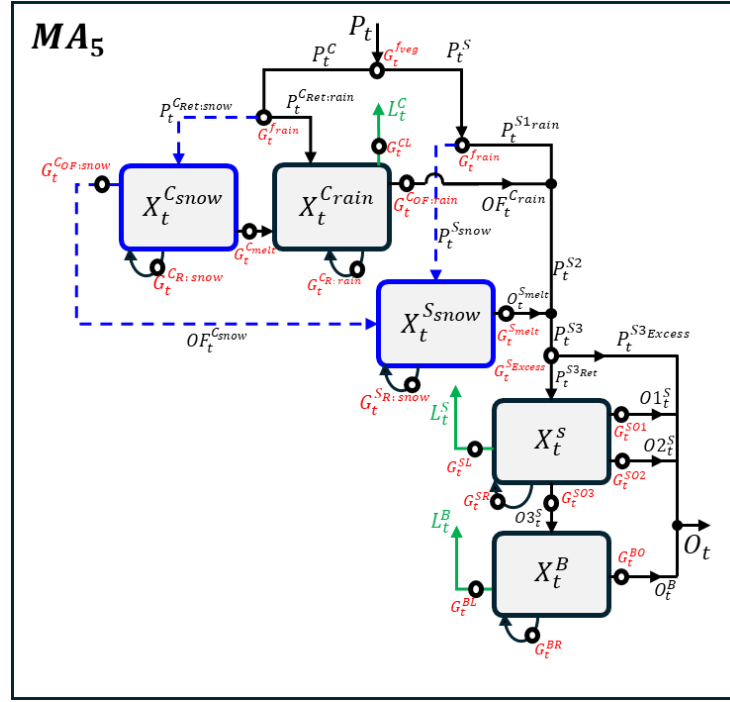


Figure 1: AI-Augmented Hydrological Catchment Model Architecture

Table 1: Different ET formulations tested in this study

Formulation	Equation of Gate	Architecture Identifier
S_1	$\sigma(a \cdot X_t + b \cdot PM_t^{Trad} + c)$	$MA_5^{PM:Trad}$
S_2	$\sigma(a \cdot X_t + b \cdot PM_t^{NN} + c)$	$MA_5^{PM:NN}$
S_3	$\sigma(a \cdot X_t + b \cdot DSW_t + c \cdot VPD_t + d \cdot T_t + e \cdot \Delta_t + f \cdot u_t + g)$	$MA_5^{Sigmoid}$
S_4	$\sigma(a \cdot X_t + b \cdot FNN_t + c)$	MA_5^{NN}

Table 2: Different Catchments tested in this study

ID	Name	Latitude	Longitude	Area (km ²)	Criteria
11523200	Trinity River above Coffee Creek, CA	41.11	-122.71	382.94	Recent rainfall-dominated (West)
11473900	Middle Fork Eel River, near Dos Rios, CA	39.71	-123.33	1925.01	
14222500	East Fork Lewis River, near Heisson, WA	45.81	-122.29	329.06	
9223000	Hams Fork below Pole Creek, near Frontier, WY	42.11	-110.71	333.15	Snowmelt-dominated
9035900	South Fork of Williams Fork, near Leal, CO	39.80	-106.03	72.84	
12358500	Middle Fork Flathead River near West Glacier, MT	48.33	-113.53	2917.03	
2472000	Leaf River, near Collins, MS	31.71	-89.41	1927.13	Historical rainfall-dominated
4185000	Tiffin River at Stryker, OH	41.74	-84.34	1064.23	
3331500	Tippecanoe River near Ora, IN	41.16	-86.56	2248.65	
3173000	Walker Creek at Bane, VA	37.27	-80.71	773.32	Recent rainfall-dominated (East)
1539000	Fishing Creek, near Bloomsburg, PA	41.08	-76.43	701.78	
1667500	Rapidan River near Culpeper, VA	38.36	-78.27	1210.77	
9505800	West Clear Creek, near Camp Verde AZ	34.60	-111.43	624.73	Arid Catchment
8324000	Jemez River near Jemez, NM	35.89	-106.67	1222.22	
13161500	Bruneau River at Rowland, NV	41.70	-115.56	990.36	

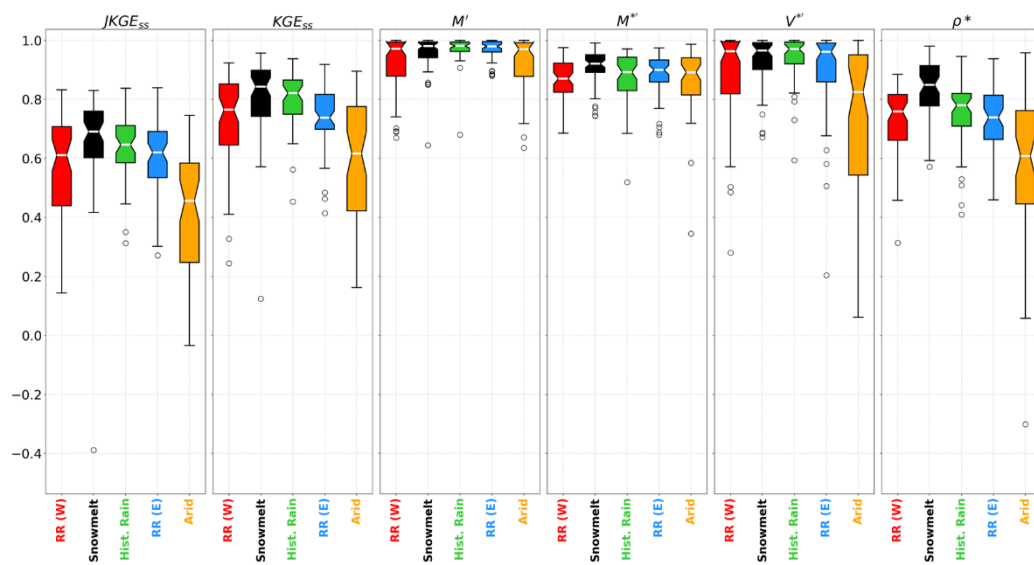


Figure 4: Yearly metrics values for different catchment regimes

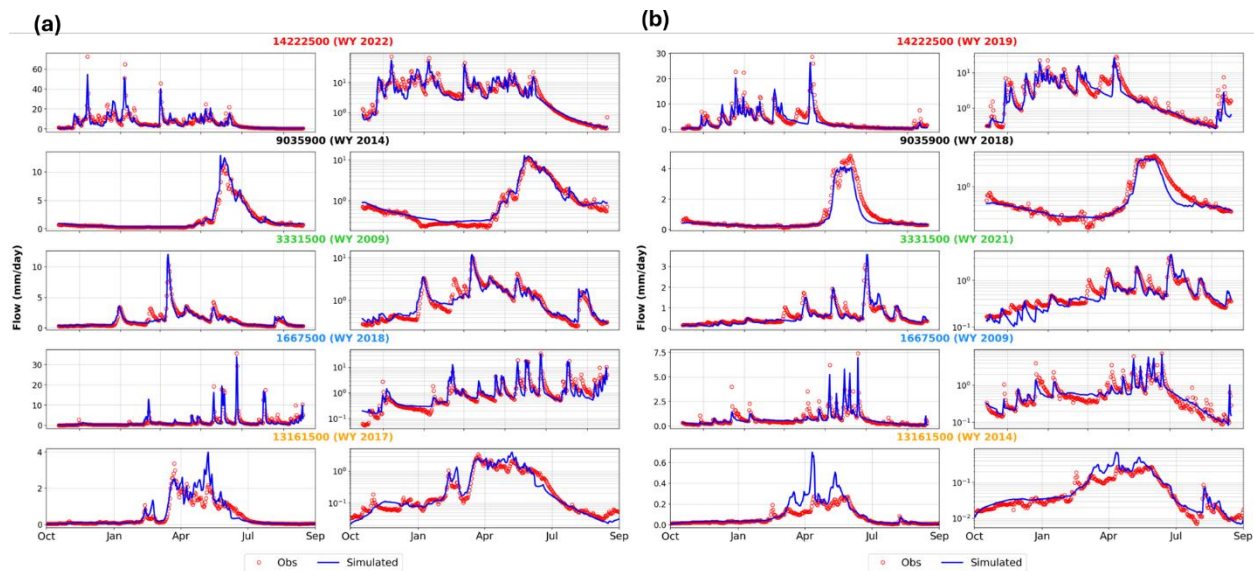


Figure 5: Model Performance Evaluation in Streamflow Hydrographs

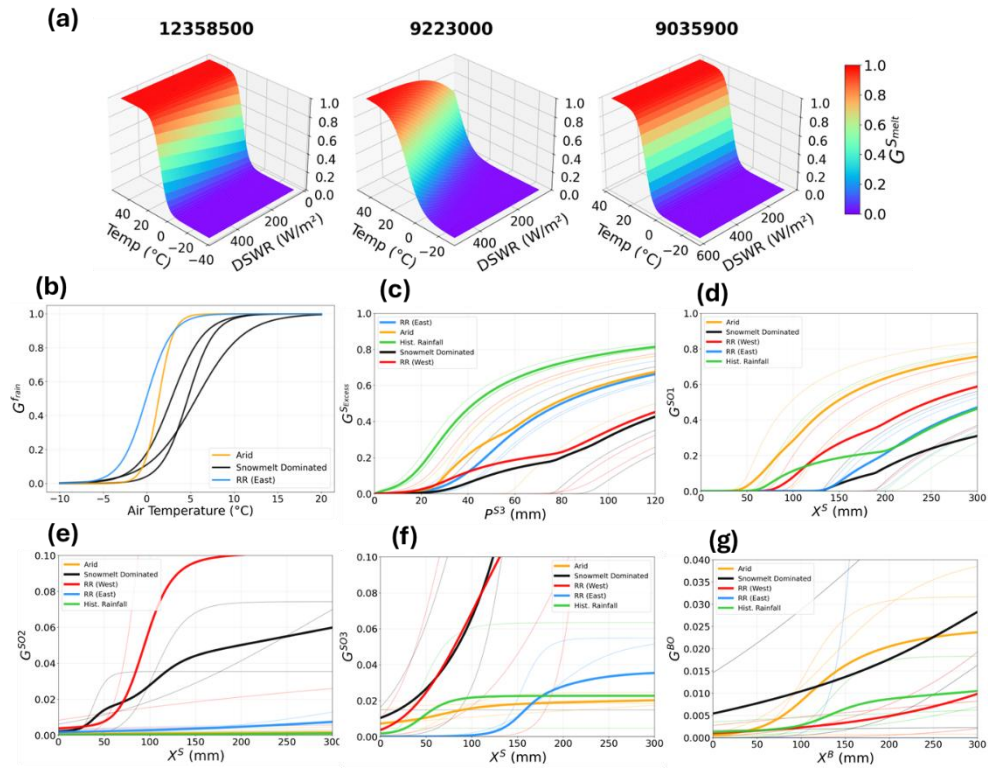


Figure 6: Operational Importance of Different Gating Functions Followed in different Types of Catchments

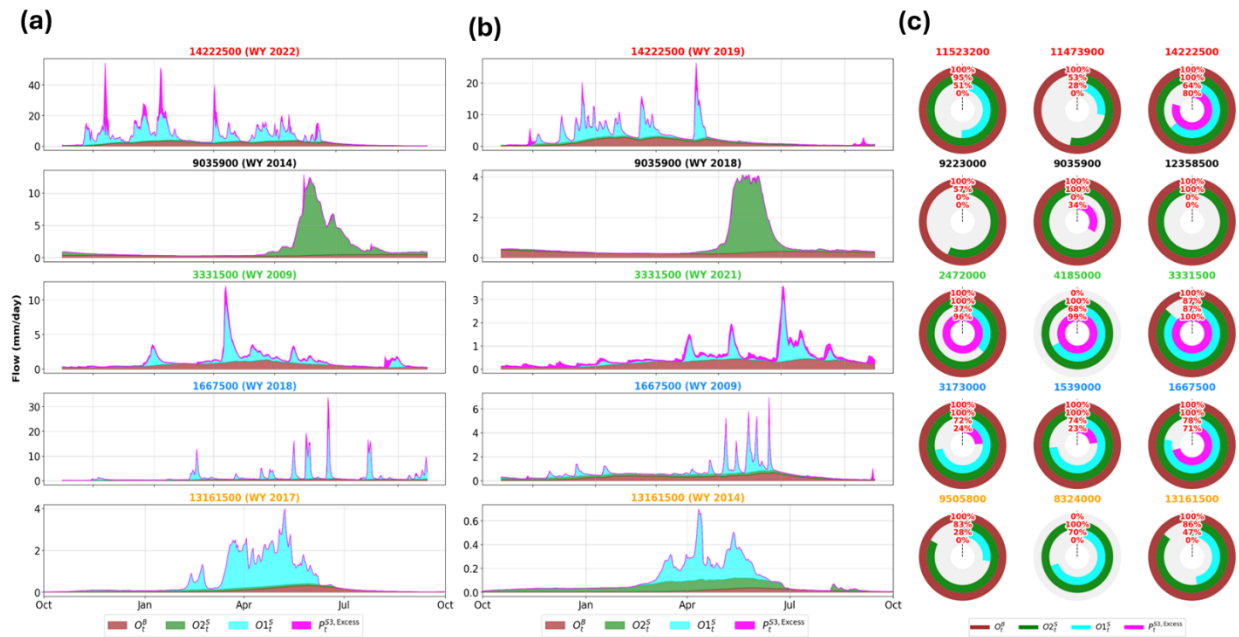


Figure 7: Operational Importance of Different Pathways Followed in different Types of Catchments

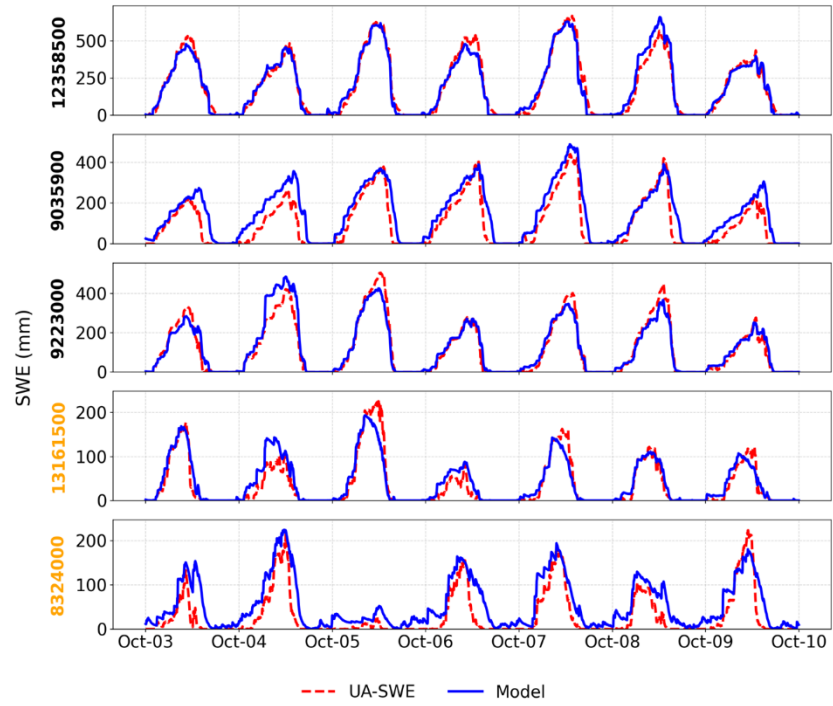


Figure 8: Evaluation of Model Performance in reproducing SWE on Different Types of Catchments

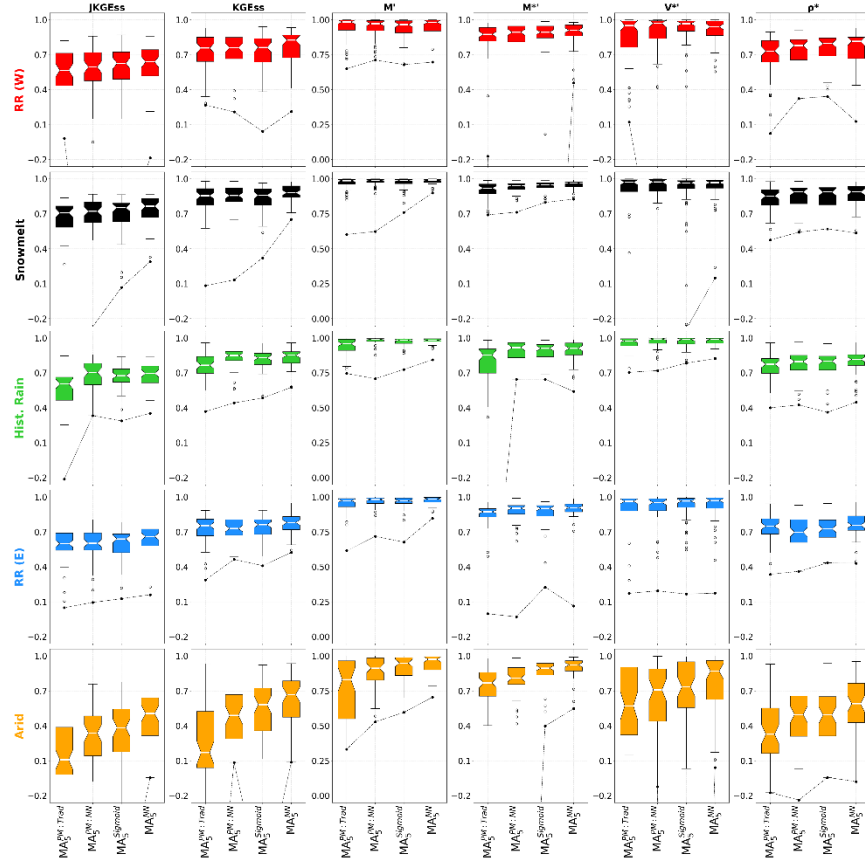


Figure 9: Evaluation of Model Performance on using Different ET Formulations

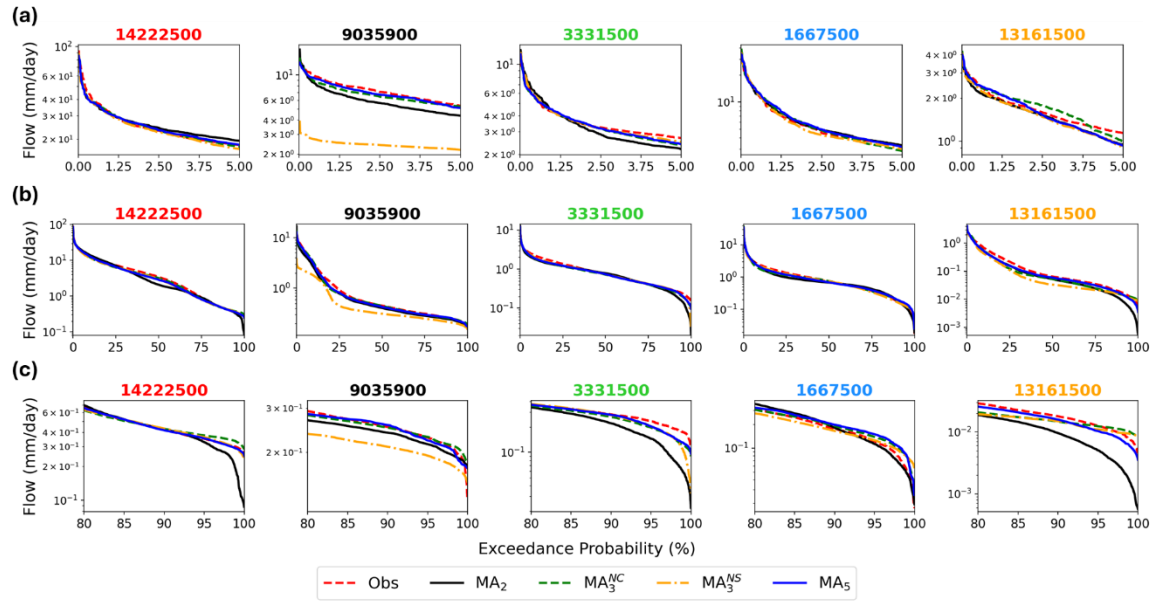


Figure 12: Evaluation of Different Ablated Versions of Model in Flow Duration Curves