Context-Aware Concept Distillation for Trustworthy Flood Prediction

Eli Levinkopf[1] , **Efrat Morin**[2] and **Claudia V. Goldman**[3]
[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel
[2]Institute of Earth Sciences, The Hebrew University of Jerusalem, Israel
[3]Hebrew University Business School, The Hebrew University of Jerusalem, Israel
{eli.levinkopf, efrat.morin, claudia.goldman}@mail.huji.ac.il

Abstract

Effective flood risk management relies on accurate forecasting, yet the "black box" nature of state-of-the-art Deep Learning models creates a barrier to trust and accountability in high-stakes public safety decisions. While existing Explainable AI (XAI) methods offer local attributions, they fail to provide the verifiable, operationally meaningful causal narratives required by disaster response authorities. To address this societal challenge, we propose **Context-Aware Concept Distillation (CACD)**, a framework developed in collaboration with domain experts to distill opaque LSTMs into interpretable, hydrology-aware surrogate models. We introduce an unsupervised pipeline to discover a "Hydrological Language" and a Residual Hypernetwork that dynamically modulates these concepts based on static basin characteristics. Evaluated on 5,203 basins globally, our model achieves high fidelity (Median NSE 0.70), significantly outperforming black-box baselines (e.g., Multi Layer Perceptrons) on unseen future data. By demonstrating that human-interpretable concepts are sufficient to reconstruct flood dynamics, this work balances AI accuracy with the transparency required for responsible environmental decision-making.

1 Introduction

AI systems for flood forecasting have reached unprecedented scale and accuracy, exemplified by recent global initiatives for ungauged watersheds [Nearing *et al.*, 2024]. Such systems are outstanding in making remote data accessible and facilitating early warnings, effectively democratizing access to predictive information. However, authorities and emergency responders managing flooded areas need to trust these predictions and understand the outputs beyond a raw number. Systems that lack transparency face significant hurdles in adoption for decision support and public communication [Dikshit *et al.*, 2024]. Therefore, explainable AI (XAI) capabilities are essential to transform these powerful computational systems into trustworthy and interpretable tools [Visave, 2025; Vu *et al.*, 2025].

Currently, public authorities routinely accompany flood forecasts with causal narratives, such as "intense short-duration rainfall" or "prolonged precipitation over saturated soils", to enhance credibility. While recent studies have employed interpretable machine-learning approaches to identify flood-generating mechanisms [Jiang *et al.*, 2022], such efforts often remain qualitative and do not demonstrate whether the extracted explanations are operationally meaningful. This limits their utility for accountability. To bridge this gap, we introduce **Context-Aware Concept Distillation (CACD)**, a framework that represents flood dynamics through interpretable concepts derived from SHAP-based clustering. Unlike traditional methods, we explicitly validate usability by reconstructing flood magnitude using an encoder–decoder architecture. By demonstrating that these concepts are not only interpretable but also sufficient to re-simulate flood behavior, we move beyond descriptive explainability toward verifiable, decision-relevant understanding.

Our framework relies on a fundamental insight: hydrological reasoning is intrinsically low-dimensional but context-dependent. A complex flood event can be described by a small vocabulary of dynamic drivers (e.g., precipitation), but the impact of these drivers on streamflow is strictly modulated by the static physical attributes of the basin (e.g., slope, soil permeability). We thus train a surrogate model governed by a Residual Hypernetwork, decoupling the prediction into global hydrological laws and local physiographic adaptations.

Societal Problem: We address the "Black Box Trust Gap" in flood risk management (UN SDG 13). Decision-makers cannot rely on opaque AI predictions for life-critical early warnings without verifiable justification. **Stakeholder Roles:** The hydrological domain expert co-authoring this paper provided the data and the original model for streamflow prediction. She was actively involved in the XAI development, contributing domain expertise to assess the explanations according to their use in real-world hydrological practice. The work is further informed by her sustained, close collaboration with national-level water management and watershed authorities involved in flood prediction and risk mitigation, ensuring that the proposed methods address operational needs and support trustworthy deployment in public safety contexts. **Impact Path:** The proposed methodology bridges machine learning forecasting systems and trustworthy decision-making, enabling authorities to audit AI predictions against established

hydrological knowledge before issuing public alerts. The main impact lies in transforming powerful predictive models into reliable AI decision-support tools.

Our Contributions: (1) **Unsupervised Concept Language:** We propose a high-fidelity feature engineering pipeline using log-temporal windowing to extract semantic concepts from raw attribution scores, eliminating the need for manual concept labeling. (2) **Hydrology-Aware Architecture:** We introduce a Context-Aware Decoder based on a Residual Hypernetwork. This design imposes a structural inductive bias that models streamflow as the product of temporal forcing and spatial sensitivity. (3) **Superior Generalization via Structural Bias:** We demonstrate that our interpretable surrogate significantly outperforms a standard black-box decoder (Deep MLP) on unseen future data (Median Fidelity NSE 0.70 vs. 0.60). This suggests that the structural constraints imposed by our architecture act as a powerful regularizer, capturing the teacher's reasoning rules better than a generic non-linear model. **Code & Reproducibility:** The complete codebase, pre-trained surrogate models, and the extracted concept dataset are open-sourced at https://github.com/eli-levinkopf/cacd-flood.

2 Related Work

2.1 Deep Learning in Hydrology

The hydrological modeling community has seen a paradigm shift toward Deep Learning, driven by the availability of large-scale datasets such as CAMELS and Caravan [Kratzert *et al.*, 2023]. In particular, Long Short-Term Memory (LSTM) networks have emerged as the state-of-the-art for rainfall-runoff modeling, consistently outperforming traditional process-based models in regional benchmarks [Kratzert *et al.*, 2019] and operational forecasting frameworks [Nevo *et al.*, 2022]. While our distillation framework is architecture-agnostic, we employ an LSTM teacher because recent literature demonstrates they still outperform Transformers in hydrological simulations [Frame *et al.*, 2022; Liu *et al.*, 2024].

Kratzert et al. [2019] integrated static basin attributes to modulate LSTM processing. While our decoder also uses static features for conditioning, our goals differ fundamentally: their approach learns an opaque, high-dimensional embedding ($\mathbb{R}^{256}$) optimized solely for performance. In contrast, we enforce a low-dimensional semantic bottleneck ($\mathbb{R}^{6}$) designed explicitly for interpretability.

2.2 Explainability in Earth Systems

The deployment of AI for high-stakes climate and weather challenges requires more than just accuracy; it demands interpretability to calibrate user trust and facilitate scientific discovery [Mamalakis *et al.*, 2022]. This is particularly critical for modeling extreme events such as flash floods, where understanding the physical controls—such as runoff-contributing areas [Rinat *et al.*, 2018] and radar rainfall patterns [Morin *et al.*, 2009]—is essential for risk assessment.

However, reviews of XAI applications in hydrology [Bramm *et al.*, 2025] indicate that the field relies predominantly on post-hoc feature attribution. As surveyed by Rojat et al. [2021], these methods fall into two main categories. First, Backpropagation-based methods (e.g., Saliency Maps [Simonyan *et al.*, 2014], Grad-CAM [Selvaraju *et al.*, 2017]) compute gradients of the output with respect to input time steps. Second, Perturbation-based methods measure prediction changes when inputs are masked (e.g., SHAP [Lundberg and Lee, 2017], LIME [Ribeiro *et al.*, 2016]). These methods suffer from a lack of semantic interpretability and global consistency. A key limitation of LIME is that it implicitly trains thousands of disjoint local linear models, offering no coherent global logic. Similarly, raw SHAP values provide local attribution but fail to distinguish between different hydrological mechanisms (e.g., "Snowmelt" vs. "Flash Flood") across the entire dataset. Our work bridges this gap by aggregating atomic SHAP values into higher-level semantic concepts.

2.3 Concept-Based Interpretability

To move beyond low-level features (pixels or time steps), recent work has proposed analyzing models in terms of human-understandable concepts. Kim et al. [2018] introduced Testing with Concept Activation Vectors (TCAV), demonstrating that high-dimensional internal states of neural networks can be mapped to linear concept vectors.

This line of reasoning led to Concept Bottleneck Models (CBMs) [Koh *et al.*, 2020], which enforce a strict separation where the model first predicts concepts (e.g., "Wing Color") and then predicts the target via a linear function. However, standard CBMs suffer from two limitations in the hydrological context: **Supervision Bottleneck:** They require dense concept labels (e.g., expert annotation for every time step), which do not exist at the scale of the Caravan dataset (spanning millions of daily time steps across thousands of basins). Beyond this vast scale, it is cognitively infeasible for a human expert to manually process 65 interacting sequential parameters to label even a single event. Lacking pre-existing labels, unsupervised clustering combined with post-hoc expert validation is the only practical approach. **Static Mappings:** They typically assume a fixed relationship between concepts and targets. However, in hydrology, the relationship is context-dependent: the concept "Heavy Rain" leads to different prediction outcomes when referring to data in arid bare-soil versus humid forested basins.

Our *Context-Aware Concept Distillation* framework overcomes these limitations by discovering concepts in an unsupervised manner (via SHAP clustering) and utilizing a Residual Hypernetwork to dynamically adapt the concept-to-prediction mapping based on static basin attributes.

3 Methodology

We propose the *Context-Aware Concept Distillation* (CACD) framework, designed to interpret the reasoning of black-box sequence-based models in the physical sciences (demonstrated here on hydrological LSTMs).

Our approach is built on the hypothesis that the reasoning process of complex models is intrinsically low-dimensional but context-dependent. That is, we assume that the high-dimensional internal representations of a trained LSTM can be compressed into a small vocabulary of dynamic concepts

(e.g., "Snowmelt", "Flash Flood"), provided that the mapping from these concepts to streamflow predictions is modulated by the static physical attributes of the specific basin.

The framework operates in three stages (Figure 1): (1) Unsupervised discovery of a "Hydrological Language" via high-fidelity feature engineering and hierarchical clustering; (2) Distilling the LSTM's internal representations into these concepts via a Concept Encoder; and (3) Reconstructing streamflow predictions via a Context-Aware Decoder that utilizes a residual hypernetwork to enforce physical consistency.

3.1 Unsupervised Concept Discovery

To characterize the model's behavior, we utilize SHAP [Lundberg and Lee, 2017], which decomposes a prediction into additive contributions from each input. For each sample, this generates an attribution matrix $\Phi \in \mathbb{R}^{T \times F}$, where $T = 365$ days and F represents the set of dynamic and static input features.

We elect to perform clustering in this explanation space rather than in the raw input space. This distinction is fundamental: two basins may experience identical heavy rainfall (similar inputs), yet the LSTM may predict a flood for one (impermeable urban soil) and a negligible flow response for another (permeable sandy soil). By clustering SHAP values, we group events based on the model's internal reasoning mechanism, the specific triggers it utilized to reach a decision, rather than merely correlating weather patterns. However, the raw attribution matrix Φ is high-dimensional and sparse. To extract meaningful semantic patterns from this dense signal, we apply a specialized temporal abstraction.

Log-Temporal Windowing

The raw attribution matrix Φ is extremely high-dimensional; specifically, the dynamic subset alone contains over 5,000 values per sample (365 days $\times$ 14 dynamic features). Directly clustering this sparse time series is computationally intractable. To distill the relevant signals, we must aggregate the temporal dimension for the dynamic features.

Hydrological systems exhibit a fading memory: model sensitivity is acute for recent events but diffuse for distant ones. We replace uniform aggregation with a log-temporal windowing scheme $\mathcal{W} = \{W_0, \dots, W_6\}$. The windows cover days [0-2], [3-7], [8-14], [15-30], [31-60], [61-180], and [181-364]. These bins approximate logarithmic spacing to capture hydrologic memory decay: narrow early bins isolate acute responses (e.g., 0–2 days for surface runoff, 3–7 days for interflow), while wider bins track slow long-term dynamics (e.g., 15–30 days for soil moisture, 181–364 days for deep groundwater). For the 51 static basin attributes, temporal windowing is unnecessary; their SHAP values are appended directly to the feature vector.

Robust Statistical Features

Naïve aggregation strategies reduce a temporal window to a single sum, discarding the internal structure of the model's reasoning. To preserve the signal's morphology, we instead extract four robust statistics for each feature f and window W (where $\phi_{f,t}$ denotes the SHAP attribution of feature f at time step t): **Net Influence (I):** The signed sum of SHAP values $\sum_{t \in W} \phi_{f,t}$, capturing the net directional push (positive or negative) on the prediction. **Magnitude (M):** The sum of absolute SHAP values $\sum_{t \in W} |\phi_{f,t}|$, measuring the total attention paid to the feature regardless of direction. **Soft Peak Lag ($\tilde{t}$):** A differentiable temporal center-of-mass indicating effectively the weighted timing of the signal within the window, making this statistic stable against noise.

$$\tilde{t}_{f,W} = \frac{\sum_{t \in W} t \cdot |\phi_{f,t}|^2}{\sum_{t \in W} |\phi_{f,t}|^2} \tag{1}$$

Concentration (C^*): To distinguish between sustained pressures (e.g., Snowmelt) and acute shocks (e.g., Flash Floods), we compute a rescaled Herfindahl-Hirschman Index (HHI). Let $p_{f,t,W} = |\phi_{f,t}|/M_{f,W}$ be the proportional importance of time step t.

$$C^*_{f,W} = \frac{\sum_{t \in W} p^2_{f,t,W} - 1/|W|}{1 - 1/|W|} \tag{2}$$

C^* ranges from 0 (uniform influence) to 1 (single-day spike), providing a scale-invariant measure of temporal sparsity.

These statistics are concatenated to form the final feature vector $v \in \mathbb{R}^D$ used for clustering.

Hierarchical Concept Clustering

We define the concept vocabulary $\mathcal{C}$ via unsupervised clustering of the feature vectors v (which comprise the concatenated statistics from all windows). To determine the optimal number of concepts, we analyzed the manifold structure using Principal Component Analysis (PCA). The analysis revealed that the first principal component captures a disproportionately large share of the variance, creating a strict bipartite separation between memory-driven (baseflow) and event-driven (flash flood) dynamics. A flat clustering approach fails to capture subtler variations within these regimes. Therefore, we employ a hierarchical K-Means strategy: we first partition the space into $K_{super} = 2$ global regimes, and subsequently cluster each regime into $K_{sub} = 3$ sub-patterns based on Silhouette optimization. This yields a total of $K = 6$ distinct semantic concepts (e.g., "Delayed Snowmelt," "Convective Flash Flood") that serve as the ground truth labels for our distillation process (see Figure 3).

3.2 Context-Aware Concept Distillation

Problem Setting: Let f_θ be a pre-trained hydrological model (specifically an LSTM) that maps dynamic forcings x_d and static basin attributes x_s to streamflow predictions y_{LSTM}. We consider this model as a fixed "Teacher" and aim to train an interpretable "Student" surrogate.

We construct a distillation dataset of extreme events $\mathcal{D} = \{(h_i, y_{LSTM,i}, x_{s,i})\}_{i=1}^N$, where each sample i represents a specific target date for a specific basin. We focus exclusively on high-flow events (highest 10% streamflow) because they are the most societally impactful events, ensuring the interpretation targets critical hydrological hazards. Here, h_i is the LSTM's final hidden state, $x_{s,i}$ denotes the static basin attributes, and $y_{LSTM,i}$ is the teacher's prediction.

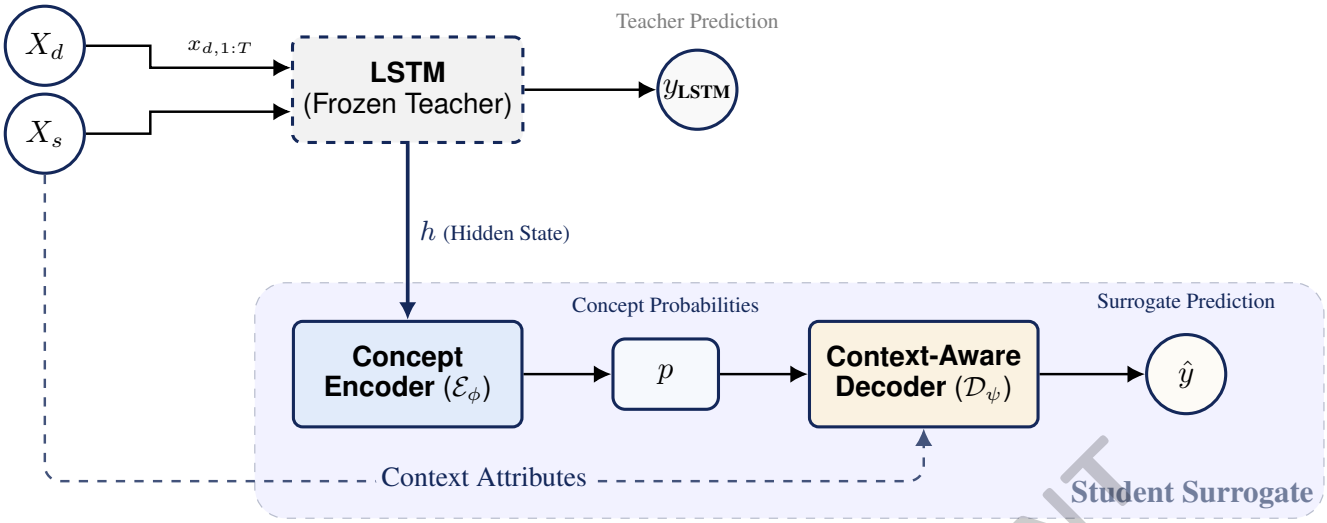

Figure 1: **Context-Aware Concept Distillation Architecture.** The inputs are split into dynamic forcings (X_d) and static attributes (X_s). The Stage 1 Encoder distills the LSTM hidden state (h) into interpretable concept probabilities (p). The Stage 2 **Context-Aware Decoder** ($\mathcal{D}_\psi$) uses the static context (X_s) to modulate the mapping from concepts (p) to streamflow ($\hat{y}$), allowing for basin-specific adaptations.

Stage 1: Distillation via Concept Encoder

The Concept Encoder $\mathcal{E}_\phi$ acts as a translator, mapping the high-dimensional LSTM's hidden state h_i to the probability distribution over the $K = 6$ concepts (see Figure 3). We parameterize $\mathcal{E}_\phi$ as an MLP with two hidden layers ($256 \to 64 \to 32 \to K$) using ReLU activations and dropout ($p = 0.2$), trained to minimize Cross-Entropy loss against SHAP cluster labels.

$$p_i = \text{Softmax}(\text{MLP}_{enc}(h_i)) \in \mathbb{R}^K \tag{3}$$

Here, $p_{i,k}$ represents the probability that extreme event i is driven by hydrological concept k. By freezing this encoder after Stage 1, we ensure that the intermediate representation remains grounded in the discovered SHAP semantics.

Stage 2: Reconstruction via Context-Aware Decoder

In Stage 2, we map the concept probabilities p_i to streamflow $\hat{y}_i$. To ensure interpretability, this mapping must remain linear ($y = W \cdot p + b$). However, a standard linear decoder is insufficient because the hydrological impact of a concept varies by geography (e.g., a "Snowmelt" concept may yield massive flow in steep rocky basins but low flow in flat permeable ones). A single global weight vector fails to capture these spatially dependent variations. We introduce a Context-Aware Decoder governed by a Residual Hypernetwork (Figure 2), which dynamically generates basin-specific weights $W(x_s)$. This preserves linear interpretability for dynamic concepts while enabling non-linear spatial modulation. This design imposes a structural inductive bias mirroring the physical principle:

$$\text{Streamflow} \propto \underbrace{\text{Basin Sensitivity } (W)}_{\text{Spatial } (x_s)} \times \underbrace{\text{Dynamic Forcing } (p_i)}_{\text{Temporal } (p_i)}$$

First, a Static Processor compresses the high-dimensional basin attributes $x_{s,i}$ into a compact latent context embedding

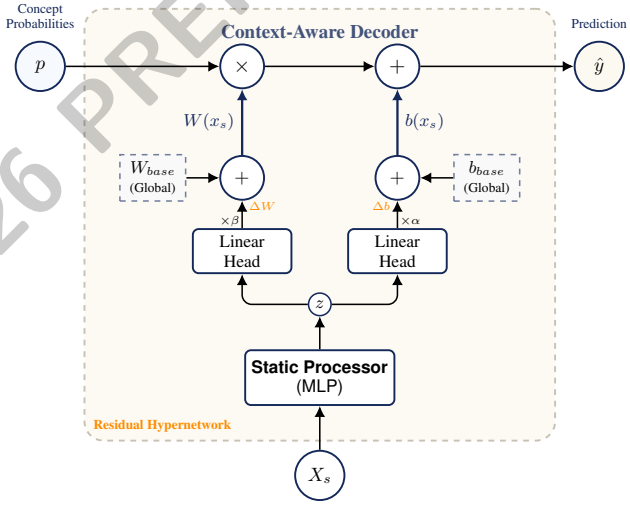

Figure 2: **Zoom-in of the Stage 2 Context-Aware Decoder.** The Residual Hypernetwork processes static attributes (X_s) to generate a compact latent context embedding (z). This embedding drives two linear heads to produce residual corrections ($\Delta W, \Delta b$), which shift the global hydrological laws (W_{base}, b_{base}) to create basin-specific weights ($W(x_s), b(x_s)$).

z_i. We parameterize this as a 2-layer MLP ($51 \to 64 \to 64$) with ReLU activations and Dropout ($p = 0.1$):

$$z_i = \text{MLP}_{stat}(x_{s,i}) \in \mathbb{R}^{64} \tag{4}$$

This embedding z_i serves as a low-dimensional signature of the basin's physical identity (e.g., "Steep & Impermeable"). Next, the hypernetwork generates basin-specific parameters via two parallel linear heads. We define the effective Bias $b(x_{s,i})$ and Weight Matrix $W(x_{s,i})$ as residual corrections to

the global average:

$$b(x_{s,i}) = b_{base} + \alpha \cdot \text{Linear}_{bias}(z_i) \quad (5)$$

$$W(x_{s,i}) = \underbrace{W_{base}}_{\text{Global Law}} + \underbrace{\beta \cdot \text{Linear}_{weight}(z_i)}_{\text{Local Adjustment}} \quad (6)$$

Here, W_{base} represents the mean hydrological response across all basins (e.g., precipitation generally increases flow), while the second term captures local deviations driven by specific traits z_i (e.g., steep slopes amplify the response). α and β are learnable scalars that scale these context-aware adjustments. The final prediction for event i is the dot product of the dynamic concepts and the context-aware weights:

$$\hat{y}_i = W(x_{s,i}) \cdot p_i + b(x_{s,i}) \quad (7)$$

Training Objective: The decoder minimizes the MSE between the surrogate prediction $\hat{y}_i$ and the teacher LSTM's prediction $y_{LSTM,i}$, rather than ground truth observations. This ensures strict fidelity to the LSTM's internal reasoning.

Crucially, while the hypernetwork is non-linear in x_s, the final prediction remains linear in concepts p_i. This enables direct interpretation of concept importance via the effective weights $W(x_{s,i})$, enforcing a strict separation of concerns between temporal dynamics and spatial heterogeneity.

4 Experimental Setup

4.1 Dataset: Caravan

We evaluate our framework on the Caravan dataset [Kratzert *et al.*, 2023], utilizing a curated set of 5,203 basins spanning diverse hydro-climatic zones globally. Caravan was selected as it is the standard state-of-the-art benchmark for global hydrological machine learning, providing the only multi-continental dataset in a uniform format. To ensure robust generalization to unseen future climates, we employ a temporal split: Training (2001–2014), Validation (2015–2017), and Testing (2018–2020).

Input Features: The model utilizes two types of inputs from the Caravan dataset. (1) **Dynamic Forcings (x_d):** 14 time-varying meteorological and state variables, including precipitation, temperature (mean/min/max), shortwave/longwave radiation, dewpoint, soil moisture (4 layers), snow water equivalent, surface pressure, wind components (u/v), and potential evaporation. (2) **Static Attributes (x_s):** 51 time-invariant basin properties describing topography (e.g., area, elevation, slope), climate indices (e.g., aridity, seasonality, precipitation frequency/duration), and soil/land characteristics (e.g., groundwater depth, 15 vegetation classes, and 21 land cover classes).

4.2 The Teacher Model (LSTM)

Our "Teacher" model is a standard Long Short-Term Memory (LSTM) network. It processes the sequence of dynamic inputs $x_{d,1:T}$ ($T = 365$) to update a hidden state $h_t \in \mathbb{R}^{256}$. To incorporate spatial context, the static attributes x_s are passed through a fully-connected embedding network (two layers: $64 \to 32$ neurons, Tanh activation) and concatenated with the dynamic inputs at each time step.

Training Objective: The LSTM is trained to minimize the Basin-Averaged Nash-Sutcliffe Efficiency (NSE*) [Kratzert *et al.*, 2019]. Unlike standard MSE, NSE* normalizes errors by the specific basin's variance (σ_b^2), preventing high-discharge basins from dominating the gradient:

$$\mathcal{L}_{NSE^*} = \frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{N}\frac{(y_{LSTM,b,t} - y_{obs,b,t})^2}{(\sigma_b + \epsilon)^2} \quad (8)$$

where y_{LSTM} is the model prediction, y_{obs} is the observed streamflow, and ϵ is a smoothing term. The model achieves robust accuracy against ground truth observations (Median NSE ≈ 0.60), providing a high-quality distillation target.

4.3 Defining Extreme Events

Since our goal is to interpret critical hydrological hazards, we focus on **extreme events**. We construct the distillation dataset $\mathcal{D}$ by selecting the top 10% of days by streamflow volume for each basin from the training period. This yields a large-scale dataset of high-impact events, ensuring the concepts discovered (e.g., Flash Flood) represent significant hydrological mechanisms rather than background noise.

4.4 Evaluation Metrics

We evaluate the fidelity of the surrogate model using three complementary metrics, each targeting a specific dimension of model quality: (1) **Global NSE** measures the ability to capture large-scale event magnitudes (volume-weighted); (2) **Median Local NSE** measures the spatial consistency and reliability across diverse geographies (unweighted); and (3) **Success Rate** (percentage of basins with NSE > 0.7) quantifies utility, corresponding to "Good" to "Very Good" performance in hydrological standards [Moriasi *et al.*, 2007].

Target Definition: To measure distillation fidelity, all metrics use the LSTM's predictions (y_{LSTM}) as the target rather than observed streamflow. Thus, a high fidelity score (e.g., Median NSE 0.70) demonstrates the surrogate successfully mimics the LSTM, independent of the LSTM's baseline accuracy against Nature (0.60).

4.5 Baselines

To quantify the value of our architectural choices, we compare the Context-Aware Decoder against three ablative baselines: **Naive Linear (Global Weights):** A global linear regression decoder that learns a single, universal weight vector $W \in \mathbb{R}^K$ and bias b for all basins ($\hat{y} = W \cdot p + b$). This tests whether basin-specific modulation is actually required or if a global "average" law suffices. **Deep MLP (Black Box):** A non-linear decoder that concatenates concept probabilities and static features into a standard multi-layer perceptron ($\hat{y} = \text{MLP}(p, x_s)$). This tests whether our interpretability constraint (enforcing linearity in concepts) comes at the cost of performance. **Simple Summation (Feature Ablation):** To validate the necessity of our high-fidelity feature engineering, we generate alternative concepts using a standard coarse-grained approach: simple summation of SHAP values over three broad windows (Recent, Short-term, Long-term). We then retrain the full surrogate model (Encoder and Decoder) using these coarse concept labels. This isolates the

Model Architecture	Global NSE (Volume)	Median NSE (Spatial)	Success Rate (> 0.7)
Naive Linear (Global)	0.67	0.25	18.8%
Deep MLP (Black Box)	0.83	0.60	39.9%
Context-Aware (Ours)	**0.88**	**0.70**	**50.3%**

Table 1: Fidelity comparison on the full global dataset. The Context-Aware model significantly outperforms the Deep MLP in spatial generalization (Median NSE 0.70 vs 0.60). Success Rate denotes the percentage of basins achieving high fidelity (NSE > 0.7), corresponding to the "Good" to "Very Good" range in hydrological standards [Moriasi *et al.*, 2007].

contribution of the log-temporal statistics (e.g., Peak Lag) to the model's success.

5 Results and Analysis

We evaluate the framework along three dimensions: (1) **Fidelity**, ensuring the surrogate model accurately mimics the teacher; (2) **Necessity**, justifying the advanced feature engineering; and (3) **Interpretability**, confirming that the discovered concepts represent distinct hydrological processes.

5.1 Fidelity Analysis

We compare the Context-Aware Decoder against the Naive Linear and Deep MLP baselines on the full global dataset (5,203 basins). Table 1 summarizes the results.

Generalization beyond the Black Box: Remarkably, our interpretable surrogate model outperforms the black-box Deep MLP (Median NSE 0.70 vs 0.60). As shown in Figure 4, the Context-Aware model exhibits stochastic dominance over the baselines, maintaining superior performance across the basin population. This suggests that the Residual Hypernetwork design acts as a regularizer. While the Deep MLP overfits to training correlations, our architecture enforces a physical separation between global laws (W_{base}) and local adaptations (ΔW), enabling superior generalization to the unseen test period (2018–2020).

Error Analysis (The Variance Paradox): We investigated the tail of basins with low fidelity scores (NSE < 0) and found they are heavily concentrated in low-variance catchments ($\log_{10}(\sigma^2) < -1$). Paradoxically, these "failed" basins exhibit significantly lower absolute errors (Median RMSE = 0.14) compared to the high-fidelity basins (Median RMSE = 0.51). This empirical evidence confirms that the low NSE scores are not due to predictive model failure, but rather due to the inherent sensitivity of the NSE metric to vanishing denominators in low-flow regimes.

5.2 Ablation Study: Feature Engineering

To validate the necessity of our upstream feature engineering (log-temporal windows, soft peak lag), we compare it against the "Simple Summation" baseline (Table 2).

The results are stark: the Success Rate ($NSE > 0.7$) drops from 50.3% to 10.9% when using simple summation. As visualized in Figure 4, the Legacy model (Orange) degrades significantly across the entire distribution compared to the Advanced model (Blue). While simple sums capture the total moisture volume (preserving decent Global NSE), they destroy the temporal morphology of the signal. Including Soft Peak Lag and Concentration enables the clustering algorithm to distinguish between short, intense flash floods and prolonged saturation events, a structural distinction necessary for the model to generalize spatially across diverse basins.

Feature Engineering	Median NSE	Success Rate
Legacy (Simple Sum)	0.24	10.9%
Advanced (Log-Window)	**0.70**	**50.3%**

Table 2: Ablation of the Feature Engineering module. Using simple window sums causes a collapse in spatial generalization.

ID	Interpretative Label	Dominant Drivers (Evidence)
C_1	Subsurface Flow	Negative Precip Impact, Slope, Shallow Groundwater
C_2	Delayed Snowmelt	Snow Water Equivalent, Temperature
C_3	Steep Saturated Runoff	Steep Slope, Lagged Precip (2–4 wks)
C_4	Rainfall-Runoff	Precip (0–7 days), Slope
C_5	Wet-Soil Storms	Precip + Soil Moisture (Layers 1–3)
C_6	Convective Flash Flood	Immediate Precip (< 2d), Steep Slope

Table 3: The Hydrological Language. Cluster IDs are mapped to *candidate* labels based on their dominant feature drivers (visualized in Figure 3). Future work will refine these semantics with our domain expert.

5.3 Qualitative Analysis: The Hydrological Language

Figure 3 visualizes the $K = 6$ concepts discovered via clustering. Each plot aggregates SHAP attribution patterns for that cluster. Crucially, these are not just post-hoc visualizations; they define the semantic vocabulary the Concept Encoder learns in Stage 1. The hierarchical clustering reveals two distinct reasoning regimes:

Memory-Driven Concepts (C_1–C_3): These regimes are dominated by state variables. For example, C_2 (labeled as Delayed Snowmelt) is driven by temperature and snow depth with long temporal lags. Notably, C_1 (Subsurface Flow) exhibits a pattern where recent precipitation is negatively correlated with the concept's activation. This reflects a dry regime where flow is sustained by slow release rather than immediate runoff. **Event-Driven Concepts (C_4–C_6):** These regimes are dominated by meteorological forcing. C_6 (Convective Flash Flood) is the purest example, driven almost entirely by immediate precipitation (0–2 days) with high concentration.

It is important to emphasize that while these concepts align with hydrological intuition, they are fundamentally data-driven abstractions. As shown in Table 3, we assign *candidate* semantic labels to facilitate communication, but these definitions serve as a starting point for domain investigation rather than final physical ground truths. Crucially, the domain expert verified these candidate labels based on samples of the actual data points within each cluster, rather than solely relying on the SHAP plots. The primary validity of this language is functional: the high fidelity achieved by the surrogate (Median NSE 0.70) confirms that this compact vocabulary is sufficient to reconstruct the teacher's reasoning, validating the concepts' utility for flood prediction.

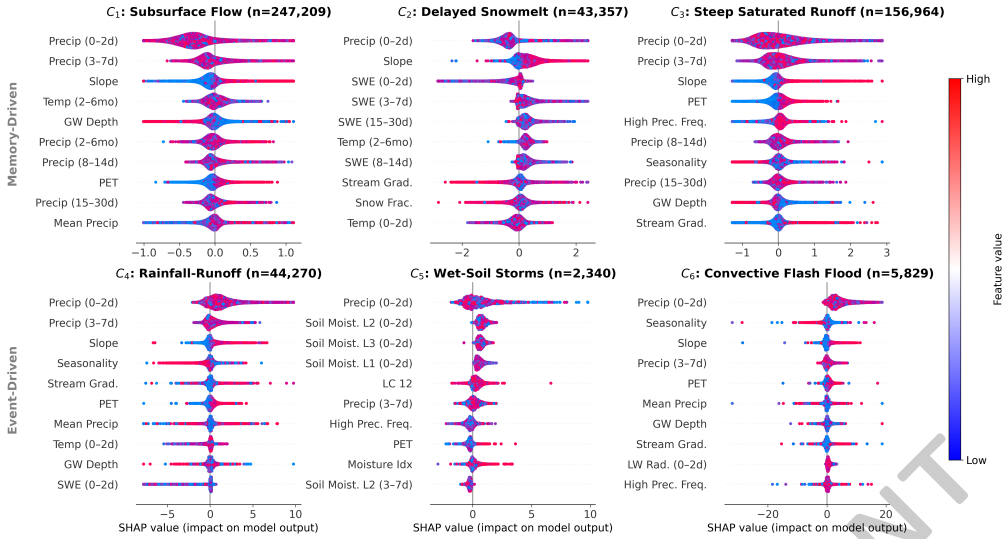

Figure 3: **The Discovered Hydrological Language.** SHAP summary plots for the $K = 6$ concept clusters. These distinct patterns serve as the ground-truth targets for the Concept Encoder. The model successfully disentangles distinct physical mechanisms. Top Row (Memory-Driven): Concepts like C_2 (Delayed Snowmelt) are driven by temperature and snow depth with long temporal lags. Bottom Row (Event-Driven): Concepts like C_6 (Convective Flash Flood) are dominated by immediate precipitation (0-2 days) with high concentration. Blue/Red points indicate low/high feature values.

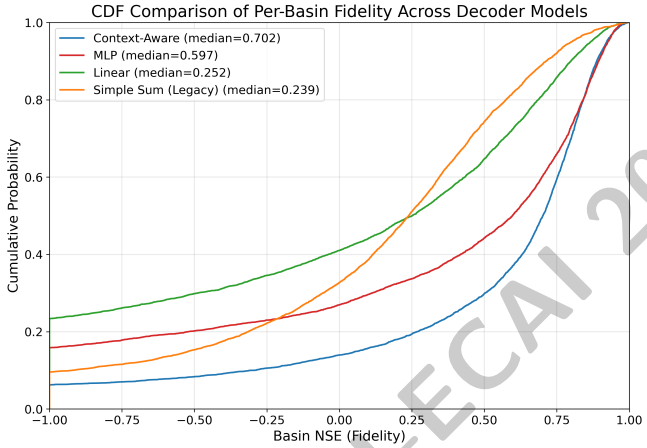

Figure 4: **Cumulative Distribution of NSE Fidelity Scores.** Comparison of per-basin performance across 5,203 basins. The Context-Aware model (Blue) consistently dominates the black-box MLP (Red). Furthermore, the gap between Blue and Orange (Simple Sum) visually demonstrates the necessity of high-fidelity feature engineering; without it, the model fails to capture the temporal dynamics required for generalization.

6 Conclusion

We presented Context-Aware Concept Distillation, a framework that transforms opaque deep sequence models into transparent, hydrology-aware surrogate models. By combining high-fidelity unsupervised concept discovery with a Residual Hypernetwork, we successfully disentangled the temporal dynamics of hydrological data from the spatial heterogeneity of diverse basins.

Our extensive evaluation on the Caravan dataset yields a compelling finding: the interpretable surrogate model generalizes to future test data better than a standard black-box baseline (Median NSE 0.70 vs 0.60). This suggests that the structural constraints of our architecture—forcing the model to reason via global laws modulated by local context—prevent the overfitting common in pure black-box approaches. Furthermore, our analysis of low-fidelity basins revealed that many apparent failures are artifacts of the NSE metric in low-variance regimes, rather than predictive shortcomings.

Real-World Impact: CACD enables trustworthy flood prediction models whose behavior can be expressed through a compact set of automatically discovered, human-interpretable concepts. Importantly, XAI should expose model flaws, not mask them. To avoid "interpretable misinformation", where a systematic LSTM error receives a plausible explanation, these narratives must be evaluated alongside historical error bounds. By demonstrating that accurate streamflow predictions can be reconstructed using only these concepts, the framework provides empirical evidence that model decisions rely on stable and inspectable abstractions rather than opaque internal dynamics. This reduces the risk of "right-for-the-wrong-reasons" behavior, directly increasing trust, auditability, and accountability for AI-assisted flood forecasting in public safety contexts. This capability allows emergency responders and authorities not just to use AI predictive outputs but to rely on them, understanding and trusting the insights and reasoning behind the automated alerts. Future work includes qualitative stakeholder evaluations to further validate the operational utility of these concepts.

Acknowledgements

Claudia V. Goldman was supported by the David Goldman Data-driven Innovation Research Center at the Hebrew Uni-

versity Business School at the Hebrew University. This study was partly supported by the ISF grant no. 1999/22.

References

[Bramm *et al.*, 2025] Andrei M. Bramm, Pavel V. Matrenin, and Alexandra I. Khalyasmaa. A review of xai methods applications in forecasting runoff and water level hydrological tasks. *Mathematics*, 13(17):2830, 2025.

[Dikshit *et al.*, 2024] Abhirup Dikshit, Biswajeet Pradhan, Sahar S. Matin, Ghassan Beydoun, M. Santosh, Hyuck-Jin Park, and Khairul Nizam Abdul Maulud. Artificial intelligence: A new era for spatial modelling and interpreting climate-induced hazard assessment. *Geoscience Frontiers*, 15(4):101815, 2024.

[Frame *et al.*, 2022] Jonathan M. Frame, Frederik Kratzert, Daniel Klotz, Martin Gauch, Guy Shalev, Oren Gilon, Lorne M. Qualls, Hoshin V. Gupta, and Grey S. Nearing. Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13):3377–3392, 2022.

[Jiang *et al.*, 2022] Shijie Jiang, Emanuele Bevacqua, and Jakob Zscheischler. River flooding mechanisms and their changes in europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, 26(24):6339–6359, 2022.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677. PMLR, 2018.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5338–5348. PMLR, 2020.

[Kratzert *et al.*, 2019] Frederik Kratzert, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing. 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, 2019.

[Kratzert *et al.*, 2023] Frederik Kratzert, Grey Nearing, Nans Addor, Tyler Erickson, Martin Gauch, Oren Gilon, Lukas Gudmundsson, Avinatan Hassidim, Daniel Klotz, Sella Nevo, Guy Shalev, and Yossi Matias. Caravan: A global community dataset for large-sample hydrology. *Scientific Data*, 10(1):61, 2023.

[Liu *et al.*, 2024] Jiangtao Liu, Yuchen Bian, Kathryn Lawson, and Chaopeng Shen. Probing the limit of hydrologic predictability with the transformer network. *Journal of Hydrology*, 637:131389, 2024.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017.

[Mamalakis *et al.*, 2022] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A. Barnes. Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In Andreas Holzinger, Randy Goebel, Ruth Fong, Tero Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI*, volume 13200 of *Lecture Notes in Computer Science*, pages 315–339. Springer, Cham, 2022.

[Moriasi *et al.*, 2007] Daniel N Moriasi, Jeffrey G Arnold, Michael W Van Liew, Ronald L Bingner, Robert Daren Harmel, and Tamie L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900, 2007.

[Morin *et al.*, 2009] Efrat Morin, Yael Jacoby, Shilo Navon, and Erez Bet-Halachmi. Towards flash-flood prediction in the dry dead sea region utilizing radar rainfall information. *Advances in Water Resources*, 32(7):1066–1076, 2009.

[Nearing *et al.*, 2024] Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, Sella Nevo, Florian Pappenberger, Christel Prudhomme, Guy Shalev, Shlomo Shenzis, Tadele Yednkachw Tekalign, Dana Weitzner, and Yossi Matias. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627:559–563, 2024.

[Nevo *et al.*, 2022] Sella Nevo, Efrat Morin, Adi Gerzi Rosenthal, Asher Metzger, Chen Barshai, Dana Weitzner, Dafi Voloshin, Frederik Kratzert, Gal Elidan, Gideon Dror, Gregory Begelman, Grey Nearing, Guy Shalev, Hila Noga, Ira Shavitt, Liora Yuklea, Moriah Royz, Niv Giladi, Nofar Peled Levi, Ofir Reich, Oren Gilon, Ronnie Maor, Shahar Timnat, Tal Shechter, Vladimir Anisimov, Yotam Gigi, Yuval Levin, Zach Moshe, Zvika Ben-Haim, Avinatan Hassidim, and Yossi Matias. Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, 26(15):4013–4032, 2022.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016.

[Rinat *et al.*, 2018] Yair Rinat, Francesco Marra, Davide Zoccatelli, and Efrat Morin. Controls of flash flood peak discharge in mediterranean basins and the special role of runoff-contributing areas. *Journal of Hydrology*, 565:846–860, 2018.

[Rojat *et al.*, 2021] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (xai) on timeseries data: A survey. arXiv preprint arXiv:2104.00950, 2021.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[Simonyan *et al.*, 2014] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, 2014.

[Visave, 2025] Jaideep Visave. Transparency in ai for emergency management: building trust and accountability. *AI and Ethics*, 5:3967–3980, Mar 2025.

[Vu *et al.*, 2025] Kiana Vu, İsmet Selçuk Özer, Phung Lai, Zheng Wu, Thilanka Munasinghe, and Jennifer Wei. From black box to insight: Explainable ai for extreme event preparedness. arXiv preprint arXiv:2511.13712, 2025.