GAUGECAST++: A Physics-Informed Latent Forecasting System for Localized Flood Prediction

Adrian Kucia
School of Computer Science
Northumbria University, UK
adrian.kucia@northumbria.ac.uk

Edward Rollason
School of Geography and Natural Sciences
Northumbria University, UK
ed.rollason@northumbria.ac.uk

Wai Lok Woo*
School of Computer Science
Northumbria University, UK
wailok.woo@northumbria.ac.uk

Abstract

Sparse-gauge flood forecasting is challenging because river-level dynamics combine transport, storage, rainfall forcing, and sharp event-scale phase shifts. We introduce GAUGECAST++, a physics-informed latent forecasting framework that learns a coordinate system in which flood dynamics become easier to evolve. Its key component is a gauge-conditioned pullback operator, which reparameterizes the forecast field and suppresses residual transport, yielding a more stable latent evolution than direct prediction in observation space. To address delayed or shifted flood peaks, GAUGECAST++ adds semigroup-consistent phase-transport learning, enforcing agreement between long-horizon forecasts and composed short-horizon rollouts in both amplitude and event phase. Finally, a decoupled inverse rainfall ambiguity module estimates uncertainty caused by imperfect future rainfall forcing. Experiments across six UK river-gauge catchments show consistent gains in event-level MSE, NSE, peak timing, and peak magnitude against strong neural forecasting baselines.

FloodAI Datasets Code

1 Introduction

Accurate flood-stage forecasting is central to early warning, emergency response, reservoir operation, and infrastructure protection. In many operational catchments, however, forecasts must be produced from sparse river-gauge observations rather than dense measurements of the full hydrodynamic state. This creates a difficult prediction problem: the model must infer future water-level trajectories from limited historical context, uncertain rainfall forcing, delayed catchment storage, and rapidly evolving event-scale dynamics. Unlike generic time-series forecasting, small phase errors can be operationally severe in flood prediction, since a forecast that captures the correct peak magnitude but predicts it several hours too early or too late may still fail as a warning system.

Physics-based hydrological and hydraulic models address this problem by encoding conservation laws, routing dynamics, rainfall–runoff processes, and calibrated catchment parameters. These models provide physical interpretability, but they often require detailed spatial inputs, boundary conditions, channel geometry, soil properties, and substantial calibration effort. Such information may be

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

unavailable, noisy, or incomplete in sparsely monitored basins. Recent machine learning approaches have therefore become increasingly attractive for hydrological prediction, especially recurrent models, process-guided learning, mass-conserving architectures, graph-based hydrological models, and differentiable process-based systems [4, 9, 14, 20, 24, 30]. More broadly, physics-informed learning and neural operators have shown that neural networks can benefit from differential constraints, weak-form residuals, operator-level supervision, and scientific inductive bias [1, 10, 12, 13, 15, 17, 18, 23].

Despite this progress, sparse-gauge flood forecasting remains challenging for modern neural forecasting systems. Strong sequence models such as N-BEATS, N-HiTS, PatchTST, Crossformer, TSMixer, PatchMixer, iTransformer and Mamba have achieved impressive results in long-horizon time-series forecasting [2, 3, 5–7, 19, 21, 22, 36]. However, most of these models operate directly in observation space and are optimized mainly through pointwise prediction losses. This makes them poorly matched to three core failure modes of flood events. First, flood-stage dynamics are transport-dominated: hydrological signals move, deform, and diffuse through the catchment, so forecasting requires learning not only amplitude but also event propagation. Second, long-horizon predictions suffer from phase drift, where short-term forecasts remain accurate but later rollouts become temporally misaligned with the observed rising limb or crest. Third, future rainfall forcing is uncertain, and deterministic models often conflate forcing error with model error, making it difficult to diagnose whether a failed hydrograph forecast is due to imperfect rainfall information or a deficient hydrological representation. Complementary work has investigated counterfactual explanations for hydrological forecasting, providing an XAI perspective on model behaviour in flood-related applications [8].

In this work, we introduce **GAUGECAST++**, a physics-informed sparse-gauge flood forecasting system designed around these three failure modes. The central idea is to learn a latent coordinate system in which flood dynamics become easier to evolve. Instead of enforcing physics directly in the original observation coordinates, GAUGECAST++ learns a gauge-conditioned pullback representation that reparameterizes the forecast field through a diffeomorphic coordinate map. Our key contributions are as follows:

- We propose GAUGECAST++, a physics-informed sparse-gauge flood forecasting system that learns latent coordinate systems for stable event-scale prediction.
- We introduce a gauge-conditioned pullback operator that suppresses residual transport in latent coordinates and provides a stability-controlled formulation for transport-dominated flood dynamics.
- We develop semigroup-consistent phase-transport learning and a decoupled inverse rainfall ambiguity module to address long-horizon peak drift and forcing-side uncertainty.
- We demonstrate consistent improvements across six UK catchments against modern time-series baselines, with ablations showing that the gains arise from the proposed structural components.

2 Related Work

Physics-informed learning. Physics-informed neural networks (PINNs) introduced the integration of data loss with PDE residual constraints and remain a core framework in scientific machine learning Raissi et al. [23]. Later work expanded this idea into physics-informed machine learning more broadly, emphasizing scientific priors, automatic differentiation, and constrained learning under sparse observations Karniadakis et al. [10]. Variational and weak-form extensions, such as hp-VPINNs, further showed that physics can be enforced more flexibly and often more robustly than by strict strong-form residual matching Kharazmi et al. [12]. Beyond pointwise prediction, neural operators learn mappings between function spaces and have become an important direction for PDE-driven learning. The Fourier Neural Operator (FNO) showed that such architectures can approximate parametric PDE solution maps efficiently Li et al. [17], while the broader neural operator framework formalized function-space learning more generally Kovachki et al. [13]. Physics-Informed Neural Operators (PINO) combined operator learning with equation-based supervision Li et al. [18], and graph-based solvers such as Message Passing Neural PDE Solvers demonstrated the value of relational inductive biases for physical systems Brandstetter et al. [1].

Hydrology and process-guided modeling. In Earth system science, deep learning has increasingly been used to complement rather than replace process knowledge [25, 28, 32, 33]. This view was made explicit in process-guided deep learning, where physical reasoning is embedded directly into the predictive framework Read et al. [24]. In hydrology, LSTM-based rainfall–runoff models have shown strong predictive skill, including in ungauged and regional settings Kratzert et al. [14]. At the same time, hydrological machine learning has been argued to require not only accuracy, but also physical consistency and interpretability Nearing et al. [20]. Recent work on spatiotemporal graph learning and differentiable process-based hydrological models further supports the importance of hybrid physically guided approaches Feng et al. [4], Sun et al. [30].

Structure-constrained and conservation-aware architectures. A complementary direction incorporates domain structure directly into model design. Mass-conserving architectures such as MC-LSTM embed conservation laws into recurrent state updates Hoedt et al. [9], while the Mass-Conserving Perceptron extends this idea beyond recurrent models Wang and Gupta [34]. In parallel, monotonic neural networks enforce known order relationships between inputs and outputs, from early monotonic architectures Sill [29] to more expressive modern formulations Runje and Shankaranarayana [26], Wehenkel and Louppe [35]. These constraints are particularly relevant in hydrology, where conservation and monotonic discharge–stage relations are often physically required.

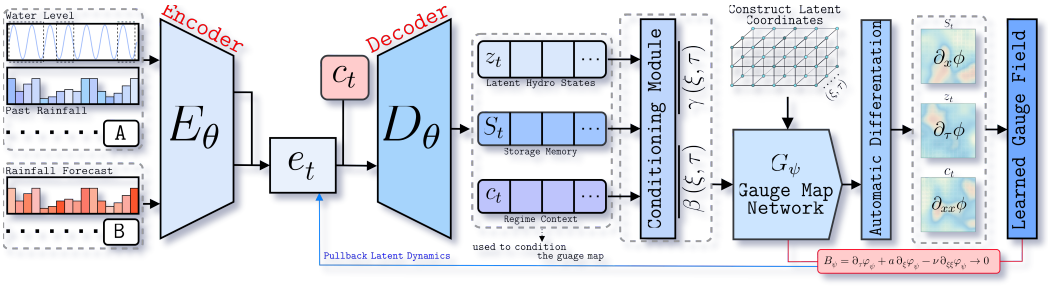


Figure 1: **Contribution 1: gauge-conditioned pullback field learning.** Given the encoded latent hydrological state, the gauge field network predicts a sample-conditioned coordinate potential $\phi_\psi(\tau, x)$ and its derivatives. Automatic differentiation produces the spatial derivative $\partial_x\phi_\psi$, curvature term $\partial_{xx}\phi_\psi$, temporal derivative $\partial_\tau\phi_\psi$, and residual transport coefficient B_ψ. These quantities define the learned coordinate field and latent transport coefficients used by the pullback physics layer to transform the forecasting dynamics into a better-conditioned latent coordinate system.

3 Proposed Method

3.1 Preliminaries

We introduce only the notation needed for the main development and defer the complete notation table to Appendix B. Bold uppercase letters such as $\mathbf{X}$, $\mathbf{U}$, and $\mathbf{Y}$ denote discrete input and output sequences, while lowercase letters such as u, v, and ϕ denote latent continuous fields or coordinate maps. We use ∂_ξ and ∂_τ for partial derivatives with respect to the latent spatial and temporal coordinates, respectively, and $\|\cdot\|_{L^2}$ to denote the standard L^2 norm on the latent domain.

Our formulation is built on a canonical latent spacetime domain $\Omega = (0,1) \times (0,1)$ with coordinates (ξ, τ), where ξ denotes latent space and τ denotes normalized forecast time. A learned gauge map ϕ_ψ reparameterizes the latent spatial coordinate, and v_θ denotes the transformed latent solution after pullback. The most important symbols used in the main text are summarized in Table 9.

3.2 Problem Formulation

We study a supervised sequence-to-trajectory forecasting problem. Let

$$\mathcal{D} = \{(\mathbf{X}^{(i)}, \mathbf{U}^{(i)}, \mathbf{Y}^{(i)})\}_{i=1}^{N}, \qquad (\mathbf{X}^{(i)}, \mathbf{U}^{(i)}, \mathbf{Y}^{(i)}) \in \mathbb{R}^{L\times d_x} \times \mathbb{R}^{H\times d_u} \times \mathbb{R}^{H\times d_y}. \tag{1}$$

denote a dataset of training examples, where each sample consists of a historical context sequence $\mathbf{X}^{(i)} \in \mathbb{R}^{L\times d_x}$, an optional future conditioning sequence $\mathbf{U}^{(i)} \in \mathbb{R}^{H\times d_u}$ over a prediction horizon of length H, and a target future trajectory $\mathbf{Y}^{(i)} \in \mathbb{R}^{H\times d_y}$. The objective is to learn a parametric

forecaster F_Θ that maps $(\mathbf{X}, \mathbf{U})$ to a prediction $\widehat{\mathbf{Y}} = F_\Theta(\mathbf{X}, \mathbf{U}) \in \mathbb{R}^{H \times d_y}$. A purely discrete sequence-to-sequence view is often insufficient for capturing sharp transitions, phase shifts, and operator-level structure. We therefore posit that each forecast is generated by an underlying latent continuous field evolving on a canonical spacetime domain $\Omega := (0,1) \times (0,1)$, where $\xi \in (0,1)$ denotes a latent spatial coordinate and $\tau \in (0,1)$ denotes normalized forecast time. Specifically, we assume that there exists a latent field $u_\Theta : \Omega \to \mathbb{R}^{d_z}$ together with a readout operator $\mathcal{R}$ such that the forecast can be written as

$$\widehat{\mathbf{Y}} = \mathcal{R}[u_\Theta] = \big(\mathcal{R}_1[u_\Theta], \dots, \mathcal{R}_H[u_\Theta]\big) \in \mathbb{R}^{H \times d_y}, \qquad \mathcal{R} : \mathcal{U} \to \mathbb{R}^{H \times d_y}, \tag{2}$$

Here, u_Θ should be understood as a functional representation of the predicted trajectory, while $\mathcal{R}$ extracts its discrete horizon-wise output. To encode inductive bias at the operator level, we assume that u_Θ approximately obeys a sample-conditioned transport–diffusion relation of the form

$$\partial_\tau u(\xi,\tau) + a(\xi,\tau;\beta)\, \partial_\xi u(\xi,\tau) - \nu(\xi,\tau;\beta)\, \partial_{\xi\xi} u(\xi,\tau) = g(\xi,\tau;\beta), \tag{3}$$

where the coefficient tuple $\beta = \beta(\mathbf{X}, \mathbf{U})$ depends on the input sample. Equation (3) is not introduced as a domain-specific governing law, but as a generic operator prior for sequential prediction problems in which transport, smoothing, and external driving coexist. In this view, the forecasting problem becomes one of learning a representation whose latent dynamics are simultaneously data-consistent and operator-consistent. The main difficulty is that the first-order transport component in (3) can induce stiffness, phase delay, and unstable optimization when enforced directly in physical coordinates. Our first contribution addresses this by introducing a learned diffeomorphic gauge map that pulls back (3) into a latent coordinate system where the residual transport can be explicitly controlled and, in the ideal case, annihilated.

3.3 Gauge-Conditioned Pullback Representation

A central idea of our method is to avoid forecasting directly in the original transport-dominated coordinates. Instead, we learn a coordinate transformation that rewrites the dynamics in a latent domain where the induced evolution is better conditioned. Figure 2 illustrates this mechanism. We begin from the reduced transport–diffusion surrogate

$$\partial_t h(x,t) + a(x,t)\, \partial_x h(x,t) - \nu(x,t)\, \partial_{xx} h(x,t) = f(x,t), \qquad (x,t) \in \Omega := (0,L) \times (0,T), \tag{4}$$

where h denotes the target field, a is an effective transport coefficient, $\nu > 0$ is diffusivity, and f is an external forcing term. Rather than approximating h directly in the original coordinates, we introduce a learned gauge map $y = \phi_\psi(x,t)$, and represent the solution through the pullback ansatz as follows,

$$h(x,t) = v_\theta(\phi_\psi(x,t), t). \tag{5}$$

This decomposition separates the problem into two coupled tasks: learning the coordinate system ϕ_ψ and learning the latent evolution v_θ in that coordinate system. Applying the chain rule to (5) yields a transformed latent operator of the form

$$\begin{aligned} &\partial_t v_\theta - \nu(x,t)\big(\partial_x \phi_\psi(x,t)\big)^2\, \partial_{yy} v_\theta + B_\psi(x,t)\, \partial_y v_\theta = f(x,t), \\ &B_\psi(x,t) := \partial_t \phi_\psi(x,t) + a(x,t)\, \partial_x \phi_\psi(x,t) - \nu(x,t)\, \partial_{xx} \phi_\psi(x,t). \end{aligned} \tag{6}$$

Equation (6) is the key object in our formulation. It measures the residual transport that remains after pullback by the learned gauge map. When B_ψ is large, the transformed operator remains transport dominated; when B_ψ is small, the latent evolution becomes increasingly diffusion dominated.

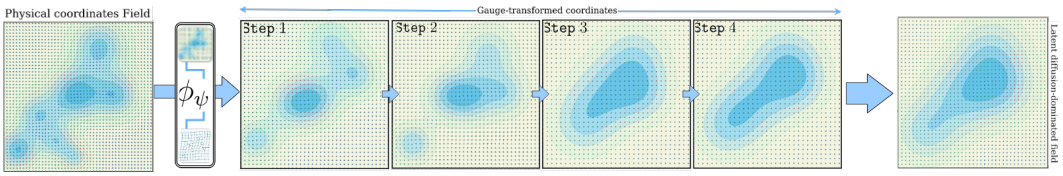


Figure 2: A learned field reparameterizes dynamics into a smooth, diffusion-dominated latent field.

Our first contribution is the proposal to learn ϕ_ψ so that the residual transport term is explicitly suppressed. This is encoded by the gauge equation, or equivalently $B_\psi(x,t) = 0$. Under exact gauge satisfaction, (6) reduces to

$$\partial_t v_\theta(y,t) - \nu(x,t)\big(\partial_x \phi_\psi(x,t)\big)^2\, \partial_{yy} v_\theta(y,t) = f(x,t), \tag{7}$$

which is a geometry-modulated parabolic evolution with no explicit first-order transport term. In this sense, the learned gauge acts as an operator preconditioner: rather than forecasting directly in the original coordinates, the model learns a coordinate system in which the induced dynamics are easier to evolve. The pullback construction above has two immediate theoretical consequences: it reduces transport stiffness in latent coordinates and yields a stability estimate controlled by the gauge residual. Full derivations and proofs are deferred to Appendix D. We first quantify residual transport in the pulled-back coordinates by defining the transformed Peclet number

$$\mathrm{Pe}_\psi(x,t) := |B_\psi(x,t)|\,\ell\,\nu(x,t)^{-1}\big(\partial_x\phi_\psi(x,t)\big)^{-2} \tag{8}$$

where $\ell > 0$ is a latent reference length scale. This ratio measures the competition between residual transport and effective latent diffusion.

Proposition 3.1 (Gauge-induced stiffness reduction)**.** *Assume that $\phi_\psi \in C^{2,1}(\Omega)$, $\partial_x\phi_\psi(x,t) \geq s_0 > 0$, and $\nu(x,t) \geq \nu_0 > 0$ on Ω. If $\|B_\psi\|_{L^\infty(\Omega)} \leq \varepsilon$, then*

$$\mathrm{Pe}_\psi(x,t) \leq \varepsilon\,\ell(\nu_0 s_0^2)^{-1}, \qquad \forall (x,t) \in \Omega \tag{9}$$

In particular, exact gauge satisfaction implies $\mathrm{Pe}_\psi \equiv 0$, so the pulled-back dynamics are purely parabolic in latent coordinates.

Proposition 3.1 shows that minimizing the gauge residual is not merely a regularization heuristic: it directly controls the transport dominance of the transformed operator. Hence, the learned coordinate system improves optimization by reducing the stiffness of the latent evolution. We next state the latent stability consequence of approximate gauge satisfaction.

Theorem 3.2 (Latent stability under approximate gauge satisfaction)**.** *Assume that $\nu(x,t) \geq \nu_0 > 0$ and $\partial_x\phi_\psi(x,t) \geq s_0 > 0$ on Ω, and let v_θ satisfy (6) with sufficient regularity and homogeneous boundary terms. Then there exists a constant $C(\nu_0, s_0) > 0$ such that*

$$\frac{1}{2}\frac{d}{dt}\|v_\theta\|_{L^2}^2 + C(\nu_0,s_0)\|\partial_y v_\theta\|_{L^2}^2 \;\leq\; \frac{1}{2}\|f\|_{L^2}^2 + \left(\frac{1}{2} + \frac{\|B_\psi\|_{L^\infty}^2}{C(\nu_0,s_0)}\right)\|v_\theta\|_{L^2}^2. \tag{10}$$

Consequently, the latent system inherits a coercive parabolic estimate whenever the gauge residual is sufficiently small, and this estimate becomes strictly sharper as $\|B_\psi\|_{L^\infty} \to 0$.

Theorem 3.2 formalizes the main effect of the gauge-conditioned pullback: learning ϕ_ψ to suppress B_ψ converts a transport-dominated forecasting problem into a better-conditioned latent evolution with stronger parabolic stability. This is the theoretical basis for our use of gauge conditioning as a forecasting mechanism rather than as a purely geometric regularizer.

3.4 Semigroup-Consistent Phase-Transport Learning

A key failure mode in long-horizon physical forecasting is *horizon-dependent phase drift*: forecasts remain accurate at short lead times, yet progressively misalign in event timing as the horizon increases. In flood-stage prediction, this manifests as delayed rising limbs and shifted crests at long lead times even when short-horizon amplitudes are well captured. This phenomenon indicates not merely pointwise prediction error, but a deeper inconsistency in the learned temporal dynamics: forecasts over long horizons do not compose coherently from shorter-horizon evolutions. To address this, we introduce a *semigroup-consistent phase-transport learning principle* that regularizes the learned forecasting dynamics to remain compositionally and temporally consistent across horizons.

Let $\mathcal{T}_\Delta$ denote the learned forecasting operator over lead time Δ, mapping the historical input x_t to a future stage trajectory. In an ideal dynamical system, the horizon-indexed operators satisfy the semigroup relation $\mathcal{T}_{\Delta_1+\Delta_2} \approx \mathcal{T}_{\Delta_2} \circ \mathcal{T}_{\Delta_1}$, meaning that a direct long-horizon forecast should agree with the composition of two shorter transitions. We therefore define the semigroup defect

$$\mathcal{E}_{m,n}(x_t) := \mathcal{T}_{m+n}(x_t) - \mathcal{C}_n\big(\mathcal{T}_m(x_t), x_t\big), \tag{11}$$

where $\mathcal{C}_n$ is a continuation operator that rolls the model forward for another n steps from the m-step forecast state. Pointwise semigroup consistency alone is insufficient for event forecasting because hydrographs may agree in amplitude while remaining misaligned in time. To explicitly regularize

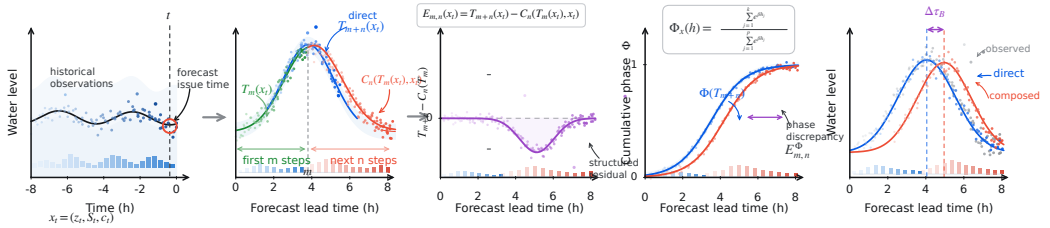


Figure 3: **Semigroup-consistent phase transport.** Direct and composed forecasts are compared through the trajectory defect $E_{m,n}$ and cumulative phase discrepancy $E^{\Phi}m, n$. Their disagreement quantifies compositional and event-phase inconsistency across forecast lead times, linking later-lead-time errors within the evaluated 8-hour horizon to smooth crest-time drift $\Delta\tau\beta$.

event timing, we define a differentiable cumulative phase map

$$\Phi_k(h) = \Big(\sum_{j=1}^{k} e^{\beta h_j}\Big)\Big(\sum_{j=1}^{P} e^{\beta h_j}\Big)^{-1}, \qquad k = 1, \ldots, P, \tag{12}$$

where $h = (h_1, \ldots, h_P)$ is a forecast trajectory over P steps and $\beta > 0$ controls the sharpness of the peak-sensitive weighting. The map $\Phi(h)$ provides a smooth event-phase representation of the forecast, emphasizing the transport of high-stage mass across the horizon. Using (12), we define the *phase semigroup defect*

$$\mathcal{E}^{\Phi}_{m,n}(x_t) := \Phi\big(\mathcal{T}_{m+n}(x_t)\big) - \Phi\Big(\mathcal{C}_n\big(\mathcal{T}_m(x_t), x_t\big)\Big). \tag{13}$$

Whereas (11) enforces trajectory-level compositional consistency, (13) explicitly penalizes horizon-wise event-time misalignment.

Theoretical implication. The central role of (11)–(13) is that long-horizon timing error becomes explicitly linked to the compositional inconsistency of the learned dynamics. This is summarized by the following proposition.

Proposition 3.3 (Phase drift bound via semigroup inconsistency)**.** *Let $\mathcal{T}_\Delta$ be the learned horizon-indexed forecasting operator, let $\Pi[h]$ denote the crest-time functional, and let $\mathcal{R}_{\text{hydro}}$ denote the latent hydrodynamic residual. Assume that: (i) $\mathcal{T}_\Delta$ is locally Lipschitz on the admissible trajectory class; (ii) the continuation operator $\mathcal{C}_n$ is locally Lipschitz; (iii) Π is locally Lipschitz in a neighborhood of an isolated nondegenerate peak. Then there exist constants $C_1, C_2, C_3 > 0$ such that*

$$\big|\hat{\tau}^{\star}_{m+n} - \tau^{\star}_{m+n}\big| \le C_1\big\|\mathcal{E}_{m,n}(x_t)\big\| + C_2\big\|\mathcal{E}^{\Phi}_{m,n}(x_t)\big\| + C_3\big\|\widehat{\mathcal{R}}_{\text{hydro}}\big\|_2, \tag{14}$$

where $\tau^{\star}_{m+n}$ and $\hat{\tau}^{\star}_{m+n}$ denote the true and predicted crest times at horizon $m+n$.

Proposition 3.3 shows that long-horizon peak drift is controlled by three quantities: trajectory-level semigroup inconsistency, phase-level semigroup inconsistency, and deviation from the latent hydrodynamic manifold. Thus, the phase-transport loss directly targets the mechanism that produces shifted flood crests rather than treating timing error as ordinary pointwise noise. The proof and extended derivation are provided in Appendix E.

3.5 Decoupled Inverse Rainfall Ambiguity Modeling

Flood-stage forecast errors are often confounded: an inaccurate hydrograph may result from an imperfect hydrological operator, but it may also arise from uncertainty in the future rainfall forcing available at issue time. We therefore introduce a *decoupled inverse rainfall ambiguity module* that diagnoses forcing-side uncertainty without feeding corrections back into the base hydrograph predictor. The base GAUGECAST++ forecast remains fixed; the ambiguity module only estimates how unresolved rainfall discrepancy may explain the frozen pullback residual and how that uncertainty propagates to the predicted water level.

Let $\delta\mathbf{r}_t := \mathbf{r}^{\text{obs}}_{t+1:t+P} - \mathbf{r}^{\text{fc}}_{t+1:t+P}$ denote the unknown discrepancy between realized rainfall and an external forecast-side rainfall proxy over the prediction horizon. Conditioned on an information set

$\mathcal{Z}_t$ containing the historical hydrological context, forecast rainfall proxy, base hydrograph prediction, and pullback diagnostics, we model the discrepancy using a horizon-wise heteroscedastic posterior:

$$q_\psi(\delta \mathbf{r}_t \mid \mathcal{Z}_t) = \prod_{k=1}^{P} \mathcal{N}\big(\delta r_{t,k}; \mu_{\psi,k}(\mathcal{Z}_t), \sigma^2_{\psi,k}(\mathcal{Z}_t)\big)\,. \tag{15}$$

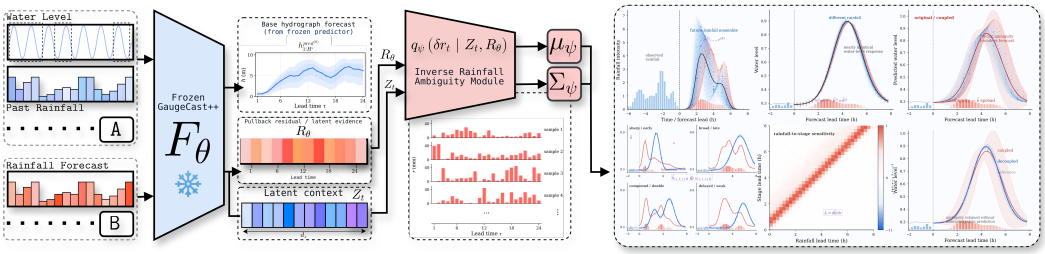


Figure 4: **Decoupled inverse rainfall ambiguity.** Different plausible rainfall forcings can induce similar water-level responses; the proposed decoupling isolates forcing ambiguity from the hydrological forecast across diverse event regimes.

Here, $\boldsymbol{\mu}_\psi$ estimates the latent rainfall-discrepancy field, while $\boldsymbol{\sigma}^2_\psi$ represents forcing-side uncertainty. Unlike a standard uncertainty head, the latent variable is not hydrograph error itself; it is an inverse forcing discrepancy constrained by the frozen physics residual.

Let $\mathcal{R}_\Theta$ denote the pullback residual computed by the frozen GAUGECAST++ predictor. If unresolved forecast error is dominated by source misspecification, then the missing rainfall-induced source should approximately cancel this residual. We therefore project the posterior mean discrepancy through the same rainfall-to-source family used by the predictor:

$$\Pi_\kappa(\boldsymbol{\mu}_\psi)_k = \mu_{\psi,k} e^{-\kappa x_k} \;\approx\; \mathcal{R}_{\Theta,k}, \qquad k = 1, \ldots, P. \tag{16}$$

This relation is the central physical constraint of the ambiguity module: the inferred rainfall discrepancy is not free to explain arbitrary forecast errors, but must account for unresolved source mismatch in the frozen pullback operator.

To quantify how rainfall uncertainty affects the hydrograph, we propagate the posterior variance through the rainfall-to-stage sensitivity of the frozen predictor. Let

$$J_t := D_{\mathbf{r}^{\text{pred}}_{t+1:t+P}} \widehat{\mathbf{h}}^{(0)}_t \in \mathbb{R}^{P\times P}, \qquad \Sigma_{\psi,t} := \operatorname{diag}(\sigma^2_{\psi,1}, \ldots, \sigma^2_{\psi,P}).$$

We define the sensitivity-weighted ambiguity score as

$$\mathcal{A}_t = P^{-1} \operatorname{tr}\left(J_t \Sigma_{\psi,t} J_t^\top\right). \tag{17}$$

Thus, ambiguity is not posterior variance alone; it is rainfall-discrepancy uncertainty filtered through the local hydrological sensitivity of the frozen forecasting system.

Theoretical implication. Let $D_\kappa := \operatorname{diag}(e^{-\kappa x_1}, \ldots, e^{-\kappa x_P})$ be the projected rainfall-to-source operator. If the dominant unresolved component of the frozen pullback residual enters through the source channel, then $\mathcal{R}_\Theta = D_\kappa \delta \mathbf{r}^\star_t + \boldsymbol{\varepsilon}_t$, where $\delta \mathbf{r}^\star_t$ is the latent forcing discrepancy and $\boldsymbol{\varepsilon}_t$ is the non-source residual remainder. Under this local source-dominance assumption, any candidate posterior mean satisfies

$$\left\|\boldsymbol{\mu}_\psi - \delta \mathbf{r}^\star_t\right\|_2 \le \|D_\kappa^{-1}\|_2 \left\|D_\kappa \boldsymbol{\mu}_\psi - \mathcal{R}_\Theta\right\|_2 + \|D_\kappa^{-1}\|_2 \|\boldsymbol{\varepsilon}_t\|_2. \tag{18}$$

Hence minimizing the projected mismatch $\|D_\kappa \boldsymbol{\mu}_\psi - \mathcal{R}_\Theta\|_2^2$ recovers the latent rainfall discrepancy up to the non-source residual remainder. Furthermore, for $\delta \mathbf{r}_t \sim q_\psi(\delta \mathbf{r}_t \mid \mathcal{Z}_t)$ and the first-order perturbation $\delta \widehat{\mathbf{h}}_t \approx J_t \delta \mathbf{r}_t$, uncertainty propagation gives

$$P^{-1} \mathbb{E}\left[\|\delta \widehat{\mathbf{h}}_t\|_2^2 \,\middle|\, \mathcal{Z}_t\right] \approx P^{-1} \operatorname{tr}\left(J_t \Sigma_{\psi,t} J_t^\top\right) = \mathcal{A}_t. \tag{19}$$

Therefore, $\mathcal{A}_t$ estimates the hydrograph variability induced specifically by forcing uncertainty. The full theorem, assumptions, posterior training losses, and finite-difference sensitivity estimator are provided in Appendix F.

3.6 End-to-End Architecture and Training Objective

We first summarize how the components of GAUGECAST++ instantiate the generic forecaster F_Θ introduced in Section 3.2. Let X_t denote the historical water-level and meteorological context and let U_t denote the available future-rainfall forcing over the prediction horizon. A sequence encoder first produces a shared hydrological context, $c_t = E_{\theta_e}(X_t)$. A storage module then combines this context with the future forcing to construct horizon-wise hydrological memory states, $S_{1:H} = S_{\theta_s}(c_t, U_t)$, and a multi-horizon decoder produces a latent representation for each forecast lead time, $z_{1:H} = D_{\theta_d}(c_t, S_{1:H}, U_t)$. For a normalized forecast coordinate (ξ_k, τ_k), the context-conditioned gauge network defines the transformed spatial coordinate $y_k = \phi_\psi(\xi_k, \tau_k; c_t)$. The corresponding pullback solution network evaluates the base water-level prediction in the learned coordinate system, $\widehat{h}_k^{(0)} = v_{\theta_v}(y_k, \tau_k, z_k)$, while the event-residual branch predicts an additive correction for rapid event-scale variations, $r_k^e = R_{\theta_r}(z_k, c_t)$.

Training objective. The deterministic predictor is trained using a weighted combination of supervised, geometric, physics-informed, storage, and temporal-consistency objectives:

$$\mathcal{L}_{\text{det}} = \lambda_{\text{data}}\mathcal{L}_{\text{data}} + \lambda_g\mathcal{L}_g + \lambda_{\text{PDE}}\mathcal{L}_{\text{PDE}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{SM}}\mathcal{L}_{\text{SM}} + \lambda_{\text{SG}}\mathcal{L}_{\text{SG}} + \lambda_{\text{PT}}\mathcal{L}_{\text{PT}}.$$

Here, $\mathcal{L}_{\text{data}}$ supervises the final H-step water-level trajectory; $\mathcal{L}_g$ suppresses the gauge residual B_ψ; $\mathcal{L}_{\text{PDE}}$ penalizes the pulled-back transport–diffusion residual; $\mathcal{L}_{\text{diff}}$ regularizes the learned coordinate geometry to remain smooth, positive, and near-identity; $\mathcal{L}_{\text{SM}}$ supervises the storage-derived soil-moisture state; and $\mathcal{L}_{\text{SG}}$ and $\mathcal{L}_{\text{PT}}$ impose trajectory-level semigroup and phase-transport consistency, respectively. The event-residual module R_{θ_r} is a trainable network rather than an additional loss term; its output contributes to $\widehat{\mathbf{Y}}$ and is supervised through $\mathcal{L}_{\text{data}}$.

4 Experiments

We conduct extensive experiments to evaluate the performance and efficiency of GAUGECAST++, covering short-term forecasting, including real-world benchmark. The detailed model and experiment configurations are summarized in Appendix B.

Benchmarks. We evaluate GAUGECAST++ on sparse-gauge flood forecasting benchmarks derived from UK river monitoring sites at 15-minute resolution obtained from the Environment Agency Hydrology Data Explorer[2]. The primary benchmark contains six Environment Agency catchments—Acomb Millersfield, Acomb Garden House, Stocksfield, Nunnykirk, Knitsley, and Kielder. Each model receives 32 historical steps and predicts the next 32 steps, corresponding to an 8-hour input window and an 8-hour forecast horizon. We additionally benchmark newly installed sensors with shorter training records. In addition to full test-set evaluation, we assess event-scale forecasting on three flood windows and report MSE, NSE, peak-timing error, and peak-height error. Additional experiments on an expanded set of catchment-specific datasets, including sites with rainfall, soil moisture, soil temperature, and forecast-rainfall forcing, are provided in Appendix C

Baselines. We compare GAUGECAST++ with strong neural time-series forecasting baselines spanning Transformer-based, MLP-based, patch-based, state-space, and hierarchical forecasting architectures. The Transformer-family baselines include Crossformer [36], PatchTST [21], and iTransformer [19]. We also include lightweight MLP and patch-mixing models, namely TSMixer [3] and PatchMixer [5]. To evaluate against recent long-sequence state-space models, we compare with Mamba [6]. Finally, we include competitive hierarchical and basis-expansion forecasting models, N-HiTS [2] and N-BEATS [22]. All baselines are trained and evaluated under the same chronological data splits, input length, prediction horizon, and preprocessing protocol as GAUGECAST++.

Implementation Details. All experiments are implemented in PyTorch and conducted on a single workstation equipped with an NVIDIA RTX 4090 GPU with 24 GB memory, an Intel Core i9-14900KS CPU, and 256 GB RAM. No multi-GPU or distributed training is used. Unless otherwise stated, all models are trained with the Adam optimizer using an initial learning rate of 1×10^{-4}, a mini-batch size of 64, and mean-squared prediction loss as the primary data-fitting objective.

[2] `https://environment.data.gov.uk/hydrology/explore`

Table 1: Primary catchment-level benchmark. Test-set and event-level MSE (↓) and NSE (↑) across six UK river-gauge catchments, with three flood-event windows per catchment. Values for GAUGECAST++ report mean ± seed uncertainty and relative improvement over the strongest baseline. ✗ indicates that no observed event exceeded the station-specific flood threshold during test period.

Data	Model	GAUGECAST++ (OURS)		CROSSFORMER		PATCHTST		TSMIXER		PATCHMIXER		MAMBA		ITRANSFORMER		N-HITS		N-BEATS	
	Metrics	MSE↓ ±unc., Δ%	NSE↑ ±unc., Δ	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑
ACOMB GRN	Event 1	**0.032**$_{\pm 0.004}^{-72.6\%}$	**0.933**$_{\pm 0.014}^{+0.179}$	**0.117**	**0.754**	0.471	0.009	0.127	0.733	0.117	0.753	0.317	0.333	0.122	0.744	0.362	0.238	0.337	0.290
	Event 2	**0.075**$_{\pm 0.002}^{+2.7\%}$	**0.895**$_{\pm 0.037}^{-0.002}$	0.093	0.869	0.385	0.460	**0.073**	**0.897**	0.082	0.884	0.222	0.689	0.106	0.851	0.341	0.522	0.311	0.564
	Event 3	**0.577**$_{\pm 0.009}^{-37.7\%}$	**0.586**$_{\pm 0.020}^{+0.251}$	0.951	0.318	2.485	-0.783	**0.926**	**0.335**	1.514	-0.087	1.357	0.026	0.968	0.305	1.682	-0.207	1.712	-0.229
	Test	**0.008**$_{\pm 0.000}^{-20.0\%}$	**0.922**$_{\pm 0.001}^{+0.018}$	0.011	0.901	0.026	0.762	**0.010**	**0.904**	0.013	0.876	0.016	0.852	0.011	0.897	0.020	0.815	0.019	0.821
ACOMB MFS	Event 1	**0.034**$_{\pm 0.014}^{-34.6\%}$	**0.927**$_{\pm 0.065}^{+0.037}$	0.077	0.836	0.443	0.061	**0.052**	**0.890**	0.064	0.863	0.324	0.314	0.103	0.781	0.382	0.191	0.428	0.092
	Event 2	**0.017**$_{\pm 0.008}^{-32.0\%}$	**0.972**$_{\pm 0.031}^{+0.015}$	0.058	0.902	0.326	0.450	**0.025**	**0.957**	0.109	0.817	0.208	0.649	0.076	0.871	0.328	0.446	0.339	0.427
	Event 3	**0.402**$_{\pm 0.018}^{-24.6\%}$	**0.606**$_{\pm 0.053}^{+0.128}$	0.706	0.309	1.131	-0.107	**0.533**	**0.478**	0.553	0.458	0.872	0.146	0.752	0.264	1.192	-0.167	1.323	-0.295
	Test	**0.006**$_{\pm 0.000}^{-0.0\%}$	**0.930**$_{\pm 0.005}^{+0.003}$	0.008	0.904	0.015	0.811	**0.006**	**0.927**	0.006	0.925	0.011	0.855	0.008	0.898	0.015	0.811	0.016	0.795
STOCKSFIELD	Event 1	**0.044**$_{\pm 0.000}^{-82.4\%}$	**0.943**$_{\pm 0.008}^{+0.265}$	0.475	0.389	0.879	-0.130	0.279	0.642	**0.250**	**0.678**	0.568	0.270	0.443	0.430	-1.01	-0.299	1.097	-0.410
	Event 2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 3	**0.264**$_{\pm 0.003}^{-26.3\%}$	**0.542**$_{\pm 0.013}^{+0.161}$	0.361	0.370	0.607	-0.051	**0.358**	**0.381**	0.698	-0.209	0.442	0.234	0.486	0.158	0.673	-0.167	0.757	-0.311
	Test	**0.013**$_{\pm 0.001}^{-7.1\%}$	**0.881**$_{\pm 0.008}^{+0.008}$	0.016	0.851	4.059	-2.665	**0.014**	**0.873**	0.016	0.859	0.019	0.830	0.020	0.817	0.025	0.773	0.026	0.762
NUNNYKIRK	Event 1	**0.027**$_{\pm 0.019}^{-82.9\%}$	**0.968**$_{\pm 0.018}^{+0.155}$	0.257	0.695	0.325	0.614	0.212	0.748	**0.158**	**0.813**	0.273	0.675	0.184	0.781	0.343	0.593	0.382	0.546
	Event 2	**0.050**$_{\pm 0.003}^{-7.4\%}$	**0.913**$_{\pm 0.064}^{+0.006}$	0.056	0.902	0.249	0.566	**0.054**	**0.907**	0.282	0.509	0.133	0.768	0.093	0.839	0.180	0.686	0.216	0.624
	Event 3	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Test	**0.003**$_{\pm 0.000}^{-25.0\%}$	**0.972**$_{\pm 0.005}^{+0.010}$	0.004	0.958	0.009	0.925	**0.004**	**0.962**	0.005	0.951	0.006	0.944	0.005	0.954	0.009	0.923	0.009	0.922
KNITSLEY	Event 1	**0.011**$_{\pm 0.000}^{-35.3\%}$	**0.940**$_{\pm 0.002}^{+0.030}$	**0.017**	**0.910**	0.160	0.164	0.029	0.845	0.037	0.808	0.122	0.362	0.027	0.856	0.148	0.224	0.143	0.251
	Event 2	**0.035**$_{\pm 0.024}^{-40.7\%}$	**0.942**$_{\pm 0.032}^{+0.041}$	0.089	0.854	0.473	0.219	**0.059**	**0.901**	0.135	0.777	0.323	0.466	0.178	0.707	0.421	0.306	0.405	0.332
	Event 3	**0.007**$_{\pm 0.002}^{-41.7\%}$	**0.936**$_{\pm 0.038}^{+0.046}$	0.017	0.839	0.091	0.168	0.021	0.803	**0.012**	**0.890**	0.072	0.299	0.033	0.697	0.092	0.152	0.093	0.147
	Test	**0.003**$_{\pm 0.000}^{-0.0\%}$	**0.945**$_{\pm 0.013}^{+0.003}$	0.004	0.928	0.012	0.810	**0.003**	**0.942**	0.004	0.930	0.008	0.862	0.005	0.911	0.011	0.821	0.011	0.824
KIELDER	Event 1	**0.009**$_{\pm 0.000}^{-30.8\%}$	**0.955**$_{\pm 0.005}^{+0.022}$	0.015	0.920	0.140	0.269	0.016	0.918	**0.013**	**0.933**	0.091	0.527	0.031	0.837	0.137	0.286	0.123	0.361
	Event 2	**0.004**$_{\pm 0.001}^{-73.3\%}$	**0.983**$_{\pm 0.003}^{+0.049}$	0.029	0.869	0.087	0.610	0.017	0.700	**0.015**	**0.934**	0.081	0.637	0.047	0.788	0.068	0.692	0.059	0.735
	Event 3	**0.007**$_{\pm 0.000}^{-53.3\%}$	**0.874**$_{\pm 0.006}^{+0.206}$	0.015	0.634	0.040	0.280	**0.019**	**0.668**	0.021	0.629	0.021	0.621	0.023	0.618	0.040	0.284	0.042	0.260
	Test	**0.003**$_{\pm 0.000}^{-25.0\%}$	**0.968**$_{\pm 0.006}^{+0.017}$	0.004	0.942	0.014	0.826	**0.004**	**0.951**	0.004	0.946	0.009	0.894	0.005	0.940	0.013	0.844	0.013	0.845
Best (↑)	Count	21	21	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0

4.1 Main Results

Catchment-level forecasting. Table 1 reports test-set and event-level results across six UK catchments. GAUGECAST++ achieves the strongest overall performance, obtaining the best MSE and NSE in 21 out of 22 evaluated comparisons. On the full test splits, it reduces MSE by 20.0%, 0.0%, 7.1%, 25.0%, 0.0%, and 25.0% on Acomb Garden House, Acomb Millersfield, Stocksfield, Nunnykirk, Knitsley, and Kielder, respectively. These consistent improvements over Crossformer, PatchTST, TSMixer, PatchMixer, Mamba, iTransformer, N-HiTS, and N-BEATS show that the proposed gauge-conditioned latent dynamics improve sparse-gauge flood forecasting beyond strong sequence baselines.

Post-hoc uncertainty diagnosis. Table 2 evaluates no-retraining uncertainty scores. Contribution 3 obtains the best AUPRC and top-k precision for identifying high-error windows, indicating that forcing-attributable ambiguity provides a useful real-time risk signal beyond deterministic forecast accuracy.

Table 2: Post-hoc uncertainty benchmark. Validation is done on Stocksfield dataset and the results are averaged.

Method	r_{MSE} ↑	ROC↑	PR↑	P@10↑	P@20↑
C3 ambiguity (**ours**)	0.596$_{\pm 0.091}$	**0.757**$_{\pm 0.090}$	**0.574**$_{\pm 0.074}$	**0.544**$_{\pm 0.049}$	**0.330**$_{\pm 0.042}$
Deep ensemble [16]	0.285$_{\pm 0.009}$	0.575$_{\pm 0.022}$	0.352$_{\pm 0.061}$	0.337$_{\pm 0.030}$	0.199$_{\pm 0.013}$
Variance-only [11]	**0.650**$_{\pm 0.079}$	0.748$_{\pm 0.081}$	0.544$_{\pm 0.059}$	0.503$_{\pm 0.018}$	**0.321**$_{\pm 0.037}$
Residual indicator [31]	**0.672**$_{\pm 0.004}$	**0.763**$_{\pm 0.072}$	**0.555**$_{\pm 0.071}$	**0.528**$_{\pm 0.079}$	0.319$_{\pm 0.038}$
Sensitivity-only [27]	0.280$_{\pm 0.097}$	0.727$_{\pm 0.078}$	0.446$_{\pm 0.072}$	0.444$_{\pm 0.045}$	0.270$_{\pm 0.043}$

r_{MSE}: Pearson correlation with window MSE. ROC/PR/P@10/P@20 use top-10% high-error labels.

Event-scale forecasting. The gains are especially pronounced during flood-event windows, where accurate timing and peak magnitude are critical. For example, GAUGECAST++ reduces event MSE by up to 72.6% on Acomb Garden House and up to 73.3% on Kielder, while also improving NSE across difficult high-amplitude events. This suggests that the pullback physics and phase-consistency terms are most beneficial under transient, transport-dominated regimes.

Extended benchmark. Table 3 further evaluates GAUGECAST++ on additional UK catchments. The model remains the strongest overall method, achieving the best MSE and NSE in 15 out of 19 evaluated comparisons. These results indicate that the proposed structural components generalize beyond the primary six-catchment benchmark. Extended Table appears in Appendix C

Table 3: Extended catchment-level benchmark. Test-set and event-level MSE (↓) and NSE (↑) across additional UK river-gauge catchments and flood-event windows.

Data	Model / Metrics	GaugeCast++ (Ours) MSE↓ ±unc., Δ%	GaugeCast++ (Ours) NSE↑ ±unc., Δ	CrossFormer MSE↓	CrossFormer NSE↑	PatchTST MSE↓	PatchTST NSE↑	TSMixer MSE↓	TSMixer NSE↑	PatchMixer MSE↓	PatchMixer NSE↑	Mamba MSE↓	Mamba NSE↑	iTransformer MSE↓	iTransformer NSE↑	N-HiTS MSE↓	N-HiTS NSE↑	N-Beats MSE↓	N-Beats NSE↑
Acomb Orchard Avenue	Event 1	**0.001**$_{\pm 0.000}^{-0.0\%}$	**0.793**$_{\pm 0.055}^{+0.017}$	0.001	0.732	0.001	0.745	0.001	0.723	**0.001**	**0.776**	0.087	-19.47	0.001	0.746	0.001	0.708	0.001	0.766
	Event 2	**0.009**$_{\pm 0.000}^{-18.2\%}$	**0.797**$_{\pm 0.008}^{+0.020}$	**0.011**	**0.777**	0.014	0.693	0.016	0.665	0.020	0.575	0.308	-5.600	0.011	0.756	0.015	0.687	0.011	0.774
	Event 3	**0.028**$_{\pm 0.002}^{-12.5\%}$	**0.321**$_{\pm 0.048}^{+0.107}$	**0.032**	**0.214**	0.034	0.164	0.044	-0.084	0.035	0.141	0.110	-1.682	0.033	0.192	0.039	0.049	0.039	0.056
	Test	**0.002**$_{\pm 0.000}^{-33.3\%}$	**0.896**$_{\pm 0.015}^{+0.012}$	0.003	0.848	0.003	0.865	0.004	0.822	0.003	0.859	0.057	3.881	**0.003**	**0.884**	0.004	0.847	0.003	0.851
Haltwhistle Town Foot	Event 1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 2	**0.017**$_{\pm 0.001}^{-0.0\%}$	**0.542**$_{\pm 0.015}^{+0.001}$	0.018	0.523	0.028	0.362	0.028	0.257	**0.017**	**0.541**	0.080	-1.120	0.021	0.428	0.021	0.445	0.026	0.284
	Event 3	**0.027**$_{\pm 0.000}^{-20.6\%}$	**0.611**$_{\pm 0.014}^{+0.106}$	0.054	0.221	**0.034**	**0.505**	0.035	0.489	0.052	0.240	0.079	-0.137	0.055	0.214	0.054	0.218	0.041	0.407
	Test	**0.029**$_{\pm 0.001}^{+7.4\%}$	**0.704**$_{\pm 0.017}^{-0.018}$	**0.027**	**0.722**	0.039	0.575	0.036	0.631	0.029	0.696	0.127	-0.293	0.029	0.699	0.030	0.694	0.034	0.655
Stocksfield Lead Road	Event 1	**0.001**$_{\pm 0.000}^{-0.0\%}$	**0.910**$_{\pm 0.022}^{+0.004}$	0.001	0.877	0.002	0.829	0.001	0.859	0.001	0.901	0.064	-4.685	0.001	0.903	**0.001**	**0.906**	0.001	0.887
	Event 2	**0.000**$_{\pm 0.000}^{0.0\%}$	**0.877**$_{\pm 0.020}^{-0.059}$	0.001	0.865	**0.000**	**0.936**	0.001	0.745	0.000	0.868	0.015	-3.769	0.000	0.859	0.000	0.864	0.000	0.863
	Event 3	**0.001**$_{\pm 0.000}^{-0.0\%}$	**0.790**$_{\pm 0.086}^{+0.028}$	**0.001**	**0.762**	0.001	0.609	0.003	0.238	0.002	0.490	0.017	-3.860	0.001	0.636	0.002	0.515	0.002	0.503
	Test	**0.004**$_{\pm 0.000}^{-0.0\%}$	**0.952**$_{\pm 0.007}^{+0.006}$	0.008	0.902	0.005	0.936	0.009	0.904	**0.004**	**0.946**	0.027	-2.389	0.005	0.943	0.005	0.941	0.007	0.917
Hepscott A192	Event 1	**0.041**$_{\pm 0.000}^{+241.7\%}$	**0.247**$_{\pm 0.125}^{-0.536}$	0.045	0.218	0.043	0.243	**0.012**	**0.783**	0.059	-0.023	0.570	-8.880	0.046	0.224	0.070	-0.275	0.055	0.035
	Event 2	**0.019**$_{\pm 0.000}^{-20.8\%}$	**0.790**$_{\pm 0.026}^{+0.057}$	0.030	0.677	0.024	0.730	0.031	0.663	**0.025**	**0.733**	0.454	-3.337	0.033	0.633	0.028	0.687	0.027	0.713
	Event 3	**0.008**$_{\pm 0.001}^{-0.0\%}$	**0.337**$_{\pm 0.083}^{+0.067}$	0.009	0.135	0.009	0.244	0.011	-0.003	0.009	0.199	0.036	-2.182	**0.008**	**0.270**	0.011	0.017	0.010	0.124
	Test	**0.003**$_{\pm 0.000}^{+50.0\%}$	**0.829**$_{\pm 0.017}^{-0.032}$	0.004	0.806	**0.002**	**0.861**	0.004	0.817	0.004	0.815	0.038	-1.138	0.003	0.824	0.003	0.825	0.003	0.827
Riding Mill Millfield Road	Event 1	**0.002**$_{\pm 0.000}^{-0.0\%}$	**0.935**$_{\pm 0.004}^{+0.020}$	0.004	0.855	0.004	0.856	0.006	0.767	0.002	0.904	0.239	-7.750	0.005	0.809	**0.002**	**0.915**	0.004	0.871
	Event 2	**0.007**$_{\pm 0.002}^{-0.0\%}$	**0.953**$_{\pm 0.021}^{+0.002}$	0.010	0.936	0.015	0.900	0.013	0.913	**0.007**	**0.951**	0.070	-3.766	0.008	0.938	0.011	0.928	0.008	0.946
	Event 3	**0.023**$_{\pm 0.002}^{-8.0\%}$	**0.360**$_{\pm 0.044}^{+0.034}$	0.026	0.286	0.030	0.172	0.030	0.177	0.027	0.265	0.122	-2.328	**0.025**	**0.326**	0.026	0.300	0.029	0.205
	Test	**0.005**$_{\pm 0.000}^{-16.7\%}$	**0.942**$_{\pm 0.002}^{+0.004}$	0.009	0.909	0.007	0.921	0.007	0.921	0.007	0.924	0.033	-2.673	0.006	0.936	**0.006**	**0.938**	0.007	0.931
Best (↑)	Count	15	15	1	1	2	2	1	1	0	0	0	0	0	0	0	0	0	0

4.2 Ablation Study

To quantify the contribution of each component in GaugeCast++, we conduct a component-wise ablation study over five random seeds. Starting from the full model, we remove one modelling component at a time: the learned gauge transform ϕ_ψ, the pullback PDE residual $\mathcal{L}_{\text{PDE}}$, the soil-moisture dynamics module $\mathcal{S}_\theta$, the event-residual refinement branch $\mathcal{R}_e$, and the phase-transport consistency regularisation. We also include a data-only variant trained using only supervised prediction loss, with all physics-informed, storage-aware, and temporal-consistency components removed. This design allows us to separate the effects of learned geometry, physics regularisation, hydrological memory, event refinement, and temporal consistency.

Table 4: Ablation study of GaugeCast++. Values are reported as mean with seed-wise uncertainty. Light-blue and pale-blue cells indicate best and second-best results, respectively.

Model	ϕ_ψ	$\mathcal{L}_g$	$\mathcal{L}_{\text{PDE}}$	$\mathcal{L}_{\text{diff}}$	$\mathcal{S}_\theta$	$\mathcal{L}_{\text{SM}}$	$\mathcal{R}_e$	SG/PT	MSE_{Event} ↓±unc., Δ%	NSE_{Test} ↑±unc., Δ	NSE_{Event} ↑±unc., Δ	Δt_p (h)↓±unc., Δ%	Δh_p ↓±unc., Δ%
① GaugeCast++	✓	✓	✓	✓	✓	✓	✓	✓	**0.026**$_{\pm 0.003}^{-0.0\%}$	**0.952**$_{\pm 0.020}^{+0.000}$	**0.940**$_{\pm 0.006}^{+0.000}$	**-3.25**$_{\pm 2.29}^{0.0\%}$	**-0.096**$_{\pm 0.065}^{0.0\%}$
② w/o gauge transform, $\phi_\psi(x,t)=x$	✗	✗	✓	✗	✓	✓	✓	✓	**0.033**$_{\pm 0.004}^{+26.9\%}$	**0.941**$_{\pm 0.007}^{-0.011}$	**0.924**$_{\pm 0.008}^{-0.016}$	**-5.25**$_{\pm 0.00}^{+61.5\%}$	**-0.063**$_{\pm 0.051}^{-34.4\%}$
③ w/o pullback PDE residual, $\mathcal{L}_{\text{PDE}}$	✓	✓	✗	✓	✓	✓	✓	✓	**0.106**$_{\pm 0.057}^{+307.7\%}$	**0.883**$_{\pm 0.060}^{-0.069}$	**0.761**$_{\pm 0.128}^{-0.179}$	**+3.00**$_{\pm 0.25}^{-7.7\%}$	**+0.128**$_{\pm 0.054}^{+33.3\%}$
④ w/o storage dynamics, $\mathcal{S}_\theta$	✓	✓	✓	✓	✗	✗	✓	✓	**0.108**$_{\pm 0.029}^{+315.4\%}$	**0.811**$_{\pm 0.062}^{-0.141}$	**0.756**$_{\pm 0.065}^{-0.184}$	**+6.25**$_{\pm 2.60}^{+92.3\%}$	**+0.074**$_{\pm 0.064}^{-22.9\%}$
⑤ w/o event-residual branch, $\mathcal{R}_e$	✓	✓	✓	✓	✓	✓	✗	✓	**0.038**$_{\pm 0.004}^{+46.2\%}$	**0.945**$_{\pm 0.018}^{-0.007}$	**0.915**$_{\pm 0.010}^{-0.025}$	**+2.50**$_{\pm 0.00}^{-23.1\%}$	**-0.147**$_{\pm 0.023}^{+53.1\%}$
⑥ w/o semigroup/phase consistency	✓	✓	✓	✓	✓	✓	✓	✗	**0.030**$_{\pm 0.003}^{+15.4\%}$	**0.951**$_{\pm 0.025}^{-0.001}$	**0.935**$_{\pm 0.005}^{-0.005}$	**-3.25**$_{\pm 2.45}^{0.0\%}$	**-0.105**$_{\pm 0.062}^{+09.4\%}$
⑦ data loss only, $\mathcal{L}_{\text{data}}$	✗	✗	✗	✗	✗	✗	✗	✗	**0.121**$_{\pm 0.047}^{+365.4\%}$	**0.820**$_{\pm 0.065}^{-0.132}$	**0.726**$_{\pm 0.107}^{-0.214}$	**+8.08**$_{\pm 0.58}^{+148.6\%}$	**-0.153**$_{\pm 0.099}^{+59.4\%}$

Ablation analysis. The ablation results in Table 4 show that the performance of GaugeCast++ is structural rather than incidental. The full model achieves the best overall event MSE and event NSE, while the data-only variant exhibits the weakest performance, demonstrating that supervised sequence fitting alone is insufficient for robust flood-event forecasting. Removing the pullback PDE residual or the storage dynamics module causes the largest degradation, indicating that physics-constrained latent transport and catchment-state memory are the dominant contributors to event-scale predictive reliability. The gauge transform also improves performance, suggesting that the learned coordinate map ϕ_ψ helps align sparse single-gauge observations with a smoother latent flow representation.

5 Conclusion

We introduced GaugeCast++, a physics-informed sparse-gauge flood forecasting framework combining gauge-conditioned pullback dynamics, storage-aware forcing, phase-consistent training, and post-hoc rainfall ambiguity diagnosis. Experiments across UK catchments show consistent gains over strong neural forecasting baselines, especially during flood events, while C3 identifies high-risk forecast windows without changing deterministic predictions. Future work will extend the model to multi-gauge spatial forecasting and validate C3 in real-time warning systems.

Acknowledgments and Disclosure of Funding

This project is funded by DEFRA (NUM015) as part of the £200 million Flood and Coastal Innovation Programmes which is managed by the Environment Agency. The programmes will drive innovation in flood and coastal resilience and adaptation to a changing climate.

References

[1] Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. In *International Conference on Learning Representations*, 2022.

[2] Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-HiTS: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854.

[3] Si-An Chen, Chun-Liang Li, Sercan O. Arik, Nathanael Christian Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023.

[4] Dapeng Feng, Jiangtao Liu, Kathryn Lawson, and Chaopeng Shen. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, 58(10):e2022WR032404, 2022. doi: 10.1029/2022WR032404.

[5] Zeying Gong, Yujin Tang, and Junwei Liang. Patchmixer: A patch-mixing architecture for long-term time series forecasting, 2023.

[6] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[7] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[8] Lucas Hayes, Jacob Sanderson, Alvin Kho, Hua Mao, and Wai Lok Woo. Temporally coherent counterfactual explanations for time series forecasting. In Michelangelo Ceci, Dragi Kocev, Anna Monreale, Celine Vens, Peter Flach, Giuseppe Manco, Myra Spiliopoulou, and Elio Masciari, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 521–537, Cham, 2027. Springer Nature Switzerland. ISBN 978-3-032-37664-0.

[9] Pieter-Jan Hoedt, Frederik Kratzert, Daniel Klotz, Christina Halmich, Markus Holzleitner, Grey S. Nearing, Sepp Hochreiter, and Guenter Klambauer. Mc-lstm: Mass-conserving LSTM. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4275–4286, 2021.

[10] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[12] Ehsan Kharazmi, Zhongqiang Zhang, and George E. M. Karniadakis. hp-vpinns: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, 2021. doi: 10.1016/j.cma.2020.113547.

[13] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[14] Frederik Kratzert, Daniel Klotz, Mathew Herrnegger, Alden K. Sampson, Sepp Hochreiter, and Grey S. Nearing. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55(12):11344–11354, 2019. doi: 10.1029/2019WR026065.

[15] Adrian Kucia and Wai Lok Woo. Apilanet: Adaptive physics-informed latent network for single-sensor forecasting. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management*, CIKM '26, page 2358–2368, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725395. doi: 10.1145/3799682.3840886. URL `https://doi.org/10.1145/3799682.3840886`.

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[17] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

[18] Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. In *International Conference on Learning Representations*, 2022.

[19] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

[20] Grey S. Nearing, Frederik Kratzert, Alden Keefe Sampson, Craig S. Pelissier, Daniel Klotz, Jonathan M. Frame, Cristina Prieto, and Hoshin V. Gupta. What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3):e2020WR028091, 2021. doi: 10.1029/2020WR028091.

[21] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[22] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[23] Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

[24] Jordan S. Read, Xiaowei Jia, Jared Willard, Alison P. Appling, Jacob A. Zwart, Samantha K. Oliver, Anuj Karpatne, Gretchen J. A. Hansen, Paul C. Hanson, William Watkins, Michael Steinbach, and Vipin Kumar. Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, 55(11): 9173–9190, 2019. doi: 10.1029/2019WR024922.

[25] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566 (7743):195–204, 2019. doi: 10.1038/s41586-019-0912-1.

[26] Davor Runje and Sharath M. Shankaranarayana. Constrained monotonic neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29338–29353, 2023.

[27] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, 2008.

[28] Jacob Sanderson, Naruephorn Tengtrairat, Wai Lok Woo, Hua Mao, and Raid Rafi Al-Nima. Xfimnet: an explainable deep learning architecture for versatile flood inundation mapping with synthetic aperture radar and multi-spectral optical images. *International Journal of Remote Sensing*, 44(24):7755–7789, 2023. doi: 10.1080/01431161.2023.2288945. URL `https://doi.org/10.1080/01431161.2023.2288945`.

[29] Joseph Sill. Monotonic networks. In *Advances in Neural Information Processing Systems 10*, 1997.

[30] Alexander Y. Sun, Peishi Jiang, Maruti K. Mudunuru, and Xingyuan Chen. Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research*, 57(12): e2021WR030394, 2021. doi: 10.1029/2021WR030394.

[31] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. *Journal of Computational and Applied Mathematics*, 50(1–3):67–83, 1994.

[32] Jialou Wang, Jacob Sanderson, and Wai Lok Woo. Flash flood forecasting in north east england through weak label-guided mixture of experts with multi-scale explainability. *Journal of Hydrology: Regional Studies*, 65:103402, 2026. ISSN 2214-5818. doi: https://doi.org/10.1016/j.ejrh.2026.103402. URL `https://www.sciencedirect.com/science/article/pii/S2214581826003009`.

[33] Jialou Wang, Jacob Sanderson, and Wai Lok Woo. Hyphynet: Hydrology gnn with structure augmentation and physical knowledge integration for flood forecasting. In Mitra Baratchi, Francesco Bonchi, Riccardo Guidotti, Francesco Gullo, Paolo Mignone, Mouzhi Ge, Vincenzo Moscato, Chiara Celsi, Ruggero G. Pensa, Fabrizio Silvestri, Peter Flach, Giuseppe Manco, Myra Spiliopoulou, and Elio Masciari, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, Demo Track and Industrial Track*, pages 3–18, Cham, 2027. Springer Nature Switzerland. ISBN 978-3-032-37685-5.

[34] Yuan-Heng Wang and Hoshin V. Gupta. A mass-conserving-perceptron for machine-learning-based modeling of geoscientific systems. *Water Resources Research*, 60(4):e2023WR036461, 2024. doi: 10.1029/2023WR036461.

[35] Antoine Wehenkel and Gilles Louppe. Unconstrained monotonic neural networks. In *Advances in Neural Information Processing Systems 32*, 2019.

[36] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

A Limitations.

The present study focuses on sparse-gauge flood-stage forecasting using single-site historical observations and forecast-side rainfall inputs. Although the proposed gauge-conditioned pullback and storage-aware dynamics improve performance across the evaluated UK catchments, the model is still constrained by the quality, resolution, and representativeness of the rainfall forcing. Severe distribution shifts, unseen land-use changes, sensor failures, or rainfall forecasts with systematic bias may reduce reliability.

B Implementation Details

We summarized details of datasets, evaluation metrics, experiments and visualizations in this section.

Datasets Details We evaluate all models on an expanded set of UK hydrological forecasting datasets derived from river-gauge and catchment monitoring sites. Water-level and rainfall observations for the original benchmark sites are obtained from the Environment Agency Hydrology Data Explorer[3], using the following station references: Acomb Millersfield (**023063**), Acomb Garden House (**023062**), Stocksfield (**023066**), Nunnykirk (**022016**), Knitsley (**023839**), and Kielder (**023055**). To extend the experimental coverage, we further construct ten additional catchment-specific datasets from the processed monitoring exports: Hepscott B1337, Hepscott A192, Hepscott Crofts Park, Acomb Garden House, Haltwhistle Town Foot, Haltwhistle Woodhead, Haltwhistle Gibbs Hill, Riding Millfield Road, Riding Dipton Ford, and Stocks A695 Bridge. Each additional dataset combines water-level observations with rainfall, soil volumetric water content, soil temperature, and forecast rainfall forcing at a 15-minute temporal resolution. Future rainfall forcing is obtained from Met Office rainfall forecasts and used as an exogenous input over the prediction horizon.

Metric Details We evaluate forecasting performance using both global sequence-level metrics and event-specific hydrological metrics. Let $\mathbf{y} = \{y_i\}_{i=1}^{N}$ denote the observed water-level sequence and $\hat{\mathbf{y}} = \{\hat{y}_i\}_{i=1}^{N}$ denote the corresponding model prediction, where N is the number of evaluated forecast points. We report mean squared error (MSE) and Nash–Sutcliffe efficiency (NSE).

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \qquad \mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2 + \epsilon}, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{20}$$

where ϵ is a small constant used for numerical stability. Lower MSE indicates smaller prediction error, while higher NSE indicates better hydrological forecasting skill. An NSE value close to one corresponds to high agreement with observations, whereas values near or below zero indicate weak predictive skill relative to the observed mean.

For event-level evaluation, the same MSE and NSE definitions are computed only over the selected flood-event window $\mathcal{E}$. We denote these as $\mathrm{MSE}_{\mathrm{Event}}$ and $\mathrm{NSE}_{\mathrm{Event}}$. In addition, we evaluate peak behaviour using the predicted and observed peak indices

$$i_p = \arg\max_{i \in \mathcal{E}} y_i, \qquad \hat{i}_p = \arg\max_{i \in \mathcal{E}} \hat{y}_i. \tag{21}$$

The signed peak-timing and peak-height errors are defined as

$$\Delta t_p = t_{\hat{i}_p} - t_{i_p}, \qquad \Delta h_p = \hat{y}_{\hat{i}_p} - y_{i_p}. \tag{22}$$

Here, the sign of Δt_p distinguishes early from delayed peak predictions, while the sign of Δh_p distinguishes peak underprediction from overprediction. For quantitative accuracy comparisons, we report the absolute errors

$$|\Delta t_p| = \left| t_{\hat{i}_p} - t_{i_p} \right|, \qquad |\Delta h_p| = \left| \hat{y}_{\hat{i}_p} - y_{i_p} \right|. \tag{23}$$

The timing error $|\Delta t_p|$ is measured in hours, while $|\Delta h_p|$ is measured in physical water-level units. Lower values indicate better peak-time alignment and more accurate flood-peak magnitude estimation.

All reported results are averaged over five random seeds. We report each metric as mean $\pm$ standard deviation to quantify seed-wise uncertainty:

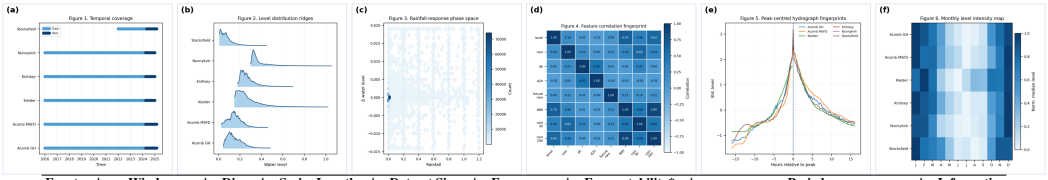
$$\mu_m = \frac{1}{S}\sum_{s=1}^{S} m_s, \qquad \sigma_m = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(m_s - \mu_m)^2}, \tag{24}$$

[3] https://environment.data.gov.uk/hydrology/explore

Table 5: Dataset **A** detailed descriptions. The dataset size is reported as (Train, Validation, Test).

Task	Dataset	Dim	Series Length	Dataset Size	Frequency	Train Period	Test Period	Forecastability* / Info
Water-level Forecasting	Acomb MSFD	10	(32, 32)	(231377, 57845, 32038)	15 min	2016-01-01–2024-03-31	2024-04-01–2025-02-28	0.62 / Weather
	Acomb GH	10	(32, 32)	(231377, 57845, 31368)	15 min	2016-01-01–2024-03-31	2024-04-01–2025-02-28	0.62 / Weather
	Kielder	10	(32, 32)	(231398, 57850, 26316)	15 min	2016-01-01–2024-03-31	2024-04-01–2024-12-31	0.43 / Weather
	Knitsley	10	(32, 32)	(231398, 57850, 26287)	15 min	2016-01-01–2024-03-31	2024-04-01–2024-12-30	0.52 / Weather
	Nunnykirk	10	(32, 32)	(231377, 57845, 26283)	15 min	2016-01-01–2024-03-31	2024-04-01–2024-12-30	0.52 / Weather
	Stocksfield	10	(32, 32)	(65356, 16340, 29161)	15 min	2022-01-01–2024-04-30	2024-05-01–2025-02-28	0.52 / Weather

* The forecastability is calculated as one minus the normalized spectral entropy of the Fourier power spectrum of the water-level series. A larger value indicates better predictability.

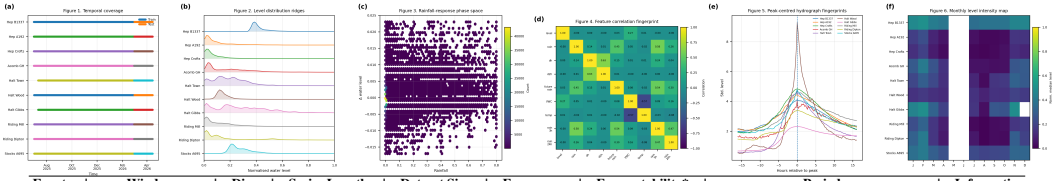


Event	Window	Dim	Series Length	Dataset Size	Frequency	Forecastability*	Period	Information
Event 1	April event	10	(32, 32)	192	15 min	Site-dependent	2024-04-08 00:00–2024-04-09 23:45	Weather
Event 2	May event	10	(32, 32)	288	15 min	Site-dependent	2024-05-28 00:00–2024-05-30 23:45	Weather
Event 3	October event	10	(32, 32)	192	15 min	Site-dependent	2024-10-16 00:00–2024-10-17 23:45	Weather

Table 6: Dataset **B** detailed descriptions. The dataset size is reported as (Train, Validation, Test).

Tasks	Dataset	Dim	Series Length	Dataset Size	Frequency	Train Period	Test Period	Forecastability*	Information
Water-level Forecasting	Hepscott B1337	10	(32, 32)	(18588, 4648, 4197)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.62	Weather
	Hepscott A192	10	(32, 32)	(18605, 4652, 4198)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.66	Weather
	Hepscott Crofts Park	10	(32, 32)	(18605, 4652, 4198)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.61	Weather
	Acomb Garden House	10	(32, 32)	(18632, 4658, 4274)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.76	Weather
	Haltwhistle Town Foot	10	(32, 32)	(17809, 4453, 4279)	15 min	2025-07-11–2026-02-28	2026-03-01–2026-04-14	0.64	Weather
	Haltwhistle Woodhead	10	(32, 32)	(18551, 4638, 4217)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.46	Weather
	Haltwhistle Gibbs Hill	10	(32, 32)	(13281, 3321, 4276)	15 min	2025-07-11–2026-02-28	2026-03-01–2026-04-14	0.74	Weather
	Riding Millfield Road	10	(32, 32)	(18548, 4638, 4265)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.72	Weather
	Riding Dipton Ford	10	(32, 32)	(18558, 4640, 4276)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.72	Weather
	Stocks A695 Bridge	10	(32, 32)	(18543, 4636, 4198)	15 min	2025-07-01–2026-02-28	2026-03-01–2026-04-14	0.67	Weather

* The forecastability is calculated as one minus the normalized spectral entropy of the Fourier power spectrum of the water-level series. A larger value indicates better predictability.



Event	Window	Dim	Series Length	Dataset Size	Frequency	Forecastability*	Period	Information
Event 1	Early March peak	10	(32, 32)	192	15 min	Site-dependent	2026-03-06 00:00–2026-03-07 23:45	Weather
Event 2	Mid-March peak	10	(32, 32)	192	15 min	Site-dependent	2026-03-12 00:00–2026-03-13 23:45	Weather
Event 3	Late March peak	10	(32, 32)	192	15 min	Site-dependent	2026-03-24 00:00–2026-03-25 23:45	Weather

where m_s is the metric value for seed s and $S = 5$ in our experiments.

Experimental Details All experiments are implemented in PyTorch and conducted on a single workstation equipped with an NVIDIA RTX 4090 GPU with 24 GB memory, an Intel Core i9-14900KS CPU, and 256 GB RAM. No multi-GPU or distributed training is used. Unless otherwise stated, all models are trained with the Adam optimizer using an initial learning rate of 1×10^{-4}, a mini-batch size of 64, and mean-squared prediction loss as the primary data-fitting objective. Training loaders are shuffled, while validation and test loaders are kept deterministic and unshuffled. Each experiment is repeated over five random seeds. For each seed, the model parameters are reinitialised and the training data loader is reconstructed with the same seed to ensure reproducible shuffling. Final results are reported as mean $\pm$ standard deviation across seeds. For multi-seed event visualisation, predictions from the independently trained seed models are averaged to form the consensus forecast, while the seed-wise standard deviation is used as an estimate of epistemic uncertainty. The default forecasting configuration uses 32 historical input steps and predicts the following 32 steps, corresponding to an 8-hour lookback window and an 8-hour forecast horizon. The same preprocessing, feature engineering, chronological train/validation/test split, and min–max scaling protocol are applied across all datasets and ablation settings.

C Additional Experiments and Visualizations

Table 7: Catchment-level forecasting. Test-set MSE (↓) and NSE (↑) across twelve UK catchments and three events per catchment. Best results are **bold**; second-best are **blue**.

Data	Model	GaugeCast++ (Ours)		CrossFormer		PatchTST		TSMixer		PatchMixer		Mamba		iTransformer		N-HiTS		N-Beats	
	Metrics	MSE↓ ±unc., Δ%	NSE↑ ±unc., Δ	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑	MSE↓	NSE↑
Acomb Garden House	Event 1	**0.000**$_{\pm 0.000}^{-0.0\%}$	**0.890**$_{\pm 0.034}^{+0.034}$	0.001	0.668	**0.000**	**0.856**	0.001	0.761	0.001	0.801	0.051	-12.23	0.001	0.775	0.001	0.726	0.001	0.743
	Event 2	**0.002**$_{\pm 0.001}^{-50.0\%}$	**0.941**$_{\pm 0.021}^{+0.064}$	**0.004**	**0.877**	0.006	0.818	0.007	0.790	0.010	0.687	0.179	-4.697	0.007	0.787	0.006	0.812	0.009	0.725
	Event 3	**0.004**$_{\pm 0.001}^{-0.0\%}$	**0.749**$_{\pm 0.031}^{+0.005}$	0.007	0.605	**0.004**	**0.744**	0.006	0.650	0.007	0.610	0.051	-2.010	0.009	0.476	0.005	0.721	0.008	0.552
	Test	**0.006**$_{\pm 0.000}^{-25.0\%}$	**0.926**$_{\pm 0.019}^{+0.027}$	0.012	0.865	**0.008**	**0.899**	0.013	0.851	0.009	0.880	0.337	-3.241	0.009	0.877	0.010	0.875	0.011	0.862
Acomb Orchard Avenue	Event 1	**0.001**$_{\pm 0.000}^{-0.0\%}$	**0.793**$_{\pm 0.055}^{+0.017}$	0.001	0.732	0.001	0.745	0.001	0.723	**0.001**	**0.776**	0.087	-19.47	0.001	0.746	0.001	0.708	0.001	0.766
	Event 2	**0.009**$_{\pm 0.000}^{-18.2\%}$	**0.797**$_{\pm 0.008}^{+0.020}$	**0.011**	**0.777**	0.014	0.693	0.016	0.665	0.020	0.575	0.308	-5.600	0.011	0.756	0.015	0.687	0.011	0.774
	Event 3	**0.028**$_{\pm 0.002}^{-12.5\%}$	**0.321**$_{\pm 0.048}^{+0.107}$	**0.032**	**0.214**	0.034	0.164	0.044	-0.084	0.035	0.141	0.110	-1.682	0.033	0.192	0.039	0.049	0.039	0.056
	Test	**0.002**$_{\pm 0.000}^{-33.3\%}$	**0.896**$_{\pm 0.015}^{+0.012}$	0.003	0.848	0.003	0.865	0.004	0.822	0.003	0.859	0.057	3.881	**0.003**	**0.884**	0.004	0.847	0.003	0.851
Haltwhistle Gibbs Hill	Event 1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 2	**0.185**$_{\pm 0.024}^{-2.1\%}$	**0.275**$_{\pm 0.096}^{+0.018}$	0.206	0.177	0.201	0.185	0.229	0.097	**0.189**	**0.257**	0.668	-1.631	0.215	0.154	0.190	0.250	0.194	0.243
	Event 3	**0.049**$_{\pm 0.001}^{-3.9\%}$	**0.708**$_{\pm 0.011}^{+0.012}$	**0.051**	**0.696**	0.062	0.633	0.078	0.535	0.113	0.332	0.207	-0.227	0.056	0.674	0.061	0.639	0.053	0.686
	Test	**0.016**$_{\pm 0.001}^{+33.3\%}$	**0.833**$_{\pm 0.017}^{-0.055}$	0.014	0.865	**0.012**	**0.888**	0.016	0.851	0.015	0.861	0.216	-0.941	**0.014**	**0.874**	0.015	0.837	0.015	0.834
Haltwhistle Town Foot	Event 1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 2	**0.017**$_{\pm 0.001}^{-0.0\%}$	**0.542**$_{\pm 0.015}^{+0.001}$	0.018	0.523	0.028	0.362	0.028	0.257	**0.017**	**0.541**	0.080	-1.120	0.021	0.428	0.021	0.445	0.026	0.284
	Event 3	**0.027**$_{\pm 0.000}^{-20.6\%}$	**0.611**$_{\pm 0.014}^{+0.106}$	0.054	0.221	**0.034**	**0.505**	0.035	0.489	0.052	0.240	0.079	-0.137	0.055	0.214	0.054	0.218	0.041	0.407
	Test	**0.029**$_{\pm 0.001}^{+7.4\%}$	**0.704**$_{\pm 0.017}^{+0.018}$	**0.027**	**0.722**	0.039	0.575	0.036	0.631	0.029	0.696	0.127	-0.293	0.029	0.699	0.030	0.694	0.034	0.655
Haltwhistle Wood Head	Event 1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 2	**0.028**$_{\pm 0.000}^{-6.7\%}$	**0.368**$_{\pm 0.007}^{+0.054}$	0.033	0.257	**0.030**	**0.314**	0.051	-0.185	0.037	0.135	0.055	-0.263	0.046	-0.059	0.042	0.040	0.043	0.037
	Event 3	**0.078**$_{\pm 0.002}^{-1.3\%}$	**0.361**$_{\pm 0.018}^{+0.012}$	0.089	0.270	**0.079**	**0.349**	0.148	-0.214	0.093	0.237	0.124	-0.012	0.130	-0.069	0.113	0.071	0.087	0.289
	Test	**0.029**$_{\pm 0.000}^{+163.6\%}$	**0.401**$_{\pm 0.012}^{-0.373}$	**0.025**	**0.483**	**0.011**	**0.774**	0.036	0.265	0.026	0.464	0.047	0.051	0.030	0.383	0.031	0.368	0.031	0.362
Stocksfield A695	Event 1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 3	**0.002**$_{\pm 0.000}^{-33.3\%}$	**0.692**$_{\pm 0.008}^{+0.038}$	0.007	0.155	0.007	0.357	**0.003**	**0.654**	0.006	0.433	0.014	-0.655	0.004	0.423	0.005	0.396	0.005	0.410
	Test	**0.008**$_{\pm 0.000}^{+33.3\%}$	**0.343**$_{\pm 0.008}^{-0.145}$	0.011	0.273	0.008	0.372	**0.006**	**0.488**	0.007	0.450	0.026	-0.687	0.007	0.388	0.008	0.379	**0.007**	**0.452**
Stocksfield Lead Road	Event 1	**0.001**$_{\pm 0.000}^{-0.0\%}$	**0.910**$_{\pm 0.022}^{+0.004}$	0.001	0.877	0.002	0.829	0.001	0.859	0.001	0.901	0.064	-4.685	0.001	0.903	**0.001**	**0.906**	0.001	0.887
	Event 2	**0.000**$_{\pm 0.000}^{0.0\%}$	**0.877**$_{\pm 0.020}^{+0.059}$	0.001	0.865	**0.000**	**0.936**	0.001	0.745	0.000	0.868	0.015	-3.769	0.000	0.859	0.000	0.864	0.000	0.863
	Event 3	**0.001**$_{\pm 0.000}^{-0.0\%}$	**0.790**$_{\pm 0.086}^{+0.028}$	**0.001**	**0.762**	0.001	0.609	0.003	0.238	0.002	0.490	0.017	-3.860	0.001	0.636	0.002	0.515	0.002	0.503
	Test	**0.004**$_{\pm 0.000}^{-0.0\%}$	**0.952**$_{\pm 0.007}^{+0.006}$	0.008	0.902	0.005	0.936	0.009	0.904	**0.004**	**0.946**	0.027	-2.389	0.005	0.943	0.005	0.941	0.007	0.917
Hepscott Crofts Park	Event 1	**0.003**$_{\pm 0.001}^{+200.0\%}$	**0.172**$_{\pm 0.303}^{-0.625}$	0.003	0.133	0.001	0.528	**0.001**	**0.797**	**0.001**	**0.633**	0.023	-6.010	0.003	0.282	0.003	0.174	0.001	0.598
	Event 2	**0.010**$_{\pm 0.000}^{-23.1\%}$	**0.806**$_{\pm 0.018}^{+0.050}$	0.015	0.728	**0.014**	**0.733**	0.015	0.726	0.013	0.756	0.200	-2.721	0.020	0.633	0.015	0.712	0.019	0.690
	Event 3	**0.004**$_{\pm 0.001}^{-0.0\%}$	**0.393**$_{\pm 0.133}^{+0.085}$	0.005	0.151	**0.004**	**0.308**	0.005	0.118	0.004	0.241	0.012	-1.157	0.005	0.191	0.005	0.129	0.004	0.283
	Test	**0.002**$_{\pm 0.000}^{+100.0\%}$	**0.826**$_{\pm 0.015}^{-0.055}$	0.001	0.856	0.001	0.865	0.002	0.817	**0.001**	**0.881**	0.015	-0.701	**0.001**	**0.877**	0.002	0.831	0.002	0.821
Hepscott A192	Event 1	**0.041**$_{\pm 0.000}^{+241.7\%}$	**0.247**$_{\pm 0.125}^{+0.536}$	0.045	0.218	0.043	0.243	**0.012**	**0.783**	0.059	-0.023	0.570	-8.880	0.046	0.224	0.070	-0.275	0.055	0.035
	Event 2	**0.019**$_{\pm 0.000}^{-20.8\%}$	**0.790**$_{\pm 0.026}^{+0.057}$	0.030	0.677	0.024	0.730	0.031	0.663	**0.025**	**0.733**	0.454	-3.337	0.033	0.633	0.028	0.687	0.027	0.713
	Event 3	**0.008**$_{\pm 0.001}^{-0.0\%}$	**0.337**$_{\pm 0.083}^{+0.067}$	0.009	0.135	0.009	0.244	0.011	-0.003	0.009	0.199	0.036	-2.182	**0.008**	**0.270**	0.011	0.017	0.010	0.124
	Test	**0.003**$_{\pm 0.000}^{+50.0\%}$	**0.829**$_{\pm 0.017}^{+0.032}$	0.004	0.806	**0.002**	**0.861**	0.004	0.817	0.004	0.815	0.038	-1.138	0.003	0.824	0.003	0.825	0.003	0.827
Hepscott B1337	Event 1	**0.025**$_{\pm 0.000}^{+78.6\%}$	**0.059**$_{\pm 0.022}^{-0.425}$	**0.020**	**0.245**	0.032	-0.174	0.028	-0.012	**0.014**	**0.484**	0.150	-4.554	0.025	0.072	0.032	-0.209	0.040	-0.415
	Event 2	**0.008**$_{\pm 0.000}^{-38.5\%}$	**0.847**$_{\pm 0.015}^{+0.073}$	0.016	0.706	0.014	0.751	0.019	0.662	0.014	0.747	0.203	-2.673	0.020	0.636	0.014	0.746	**0.013**	**0.774**
	Event 3	**0.002**$_{\pm 0.000}^{-0.0\%}$	**0.597**$_{\pm 0.061}^{+0.080}$	**0.002**	**0.517**	0.003	0.266	0.005	-0.046	0.002	0.399	0.010	-1.227	0.003	0.434	0.004	0.157	0.003	0.304
	Test	**0.004**$_{\pm 0.000}^{+33.3\%}$	**0.818**$_{\pm 0.008}^{-0.048}$	**0.003**	**0.848**	**0.003**	**0.866**	0.006	0.722	0.003	0.849	0.036	-0.642	0.004	0.829	0.004	0.798	0.005	0.789
Riding Mill Millfield Road	Event 1	**0.002**$_{\pm 0.000}^{-0.0\%}$	**0.935**$_{\pm 0.004}^{+0.020}$	0.004	0.855	0.004	0.856	0.006	0.767	0.002	0.904	0.239	-7.750	0.005	0.809	**0.002**	**0.915**	0.004	0.871
	Event 2	**0.007**$_{\pm 0.002}^{-0.0\%}$	**0.953**$_{\pm 0.021}^{+0.002}$	0.010	0.936	0.015	0.900	0.013	0.913	**0.007**	**0.951**	0.070	-3.766	0.008	0.938	0.011	0.928	0.008	0.946
	Event 3	**0.023**$_{\pm 0.002}^{-8.0\%}$	**0.360**$_{\pm 0.044}^{+0.034}$	0.026	0.286	0.030	0.172	0.030	0.177	0.027	0.265	0.122	-2.328	**0.025**	**0.326**	0.026	0.300	0.029	0.205
	Test	**0.005**$_{\pm 0.000}^{-16.7\%}$	**0.942**$_{\pm 0.002}^{+0.004}$	0.009	0.909	0.007	0.921	0.007	0.921	0.007	0.924	0.033	-2.673	0.006	0.936	**0.006**	**0.938**	0.007	0.931
Riding Mill Dipton Ford	Event 1	**0.027**$_{\pm 0.002}^{-3.6\%}$	**0.858**$_{\pm 0.007}^{+0.012}$	0.048	0.739	**0.028**	**0.846**	0.032	0.828	0.029	0.841	0.569	-2.077	0.046	0.752	0.035	0.813	0.033	0.822
	Event 2	**0.044**$_{\pm 0.004}^{+25.7\%}$	**0.918**$_{\pm 0.010}^{+0.016}$	**0.035**	**0.934**	0.048	0.911	0.058	0.893	0.053	0.901	0.212	-2.936	0.047	0.912	0.058	0.892	0.049	0.908
	Event 3	**0.028**$_{\pm 0.002}^{-9.7\%}$	**0.862**$_{\pm 0.008}^{+0.013}$	0.039	0.809	0.040	0.807	0.040	0.803	0.036	0.826	0.337	-0.645	**0.031**	**0.849**	0.050	0.757	0.033	0.840
	Test	**0.015**$_{\pm 0.001}^{+36.4\%}$	**0.902**$_{\pm 0.008}^{-0.026}$	0.018	0.883	**0.011**	**0.928**	0.014	0.904	**0.014**	**0.906**	0.295	-0.921	0.019	0.876	0.017	0.891	0.016	0.894
Best (↑)	Count	29	29	2	2	5	5	3	3	2	2	0	0	1	1	0	0	0	0

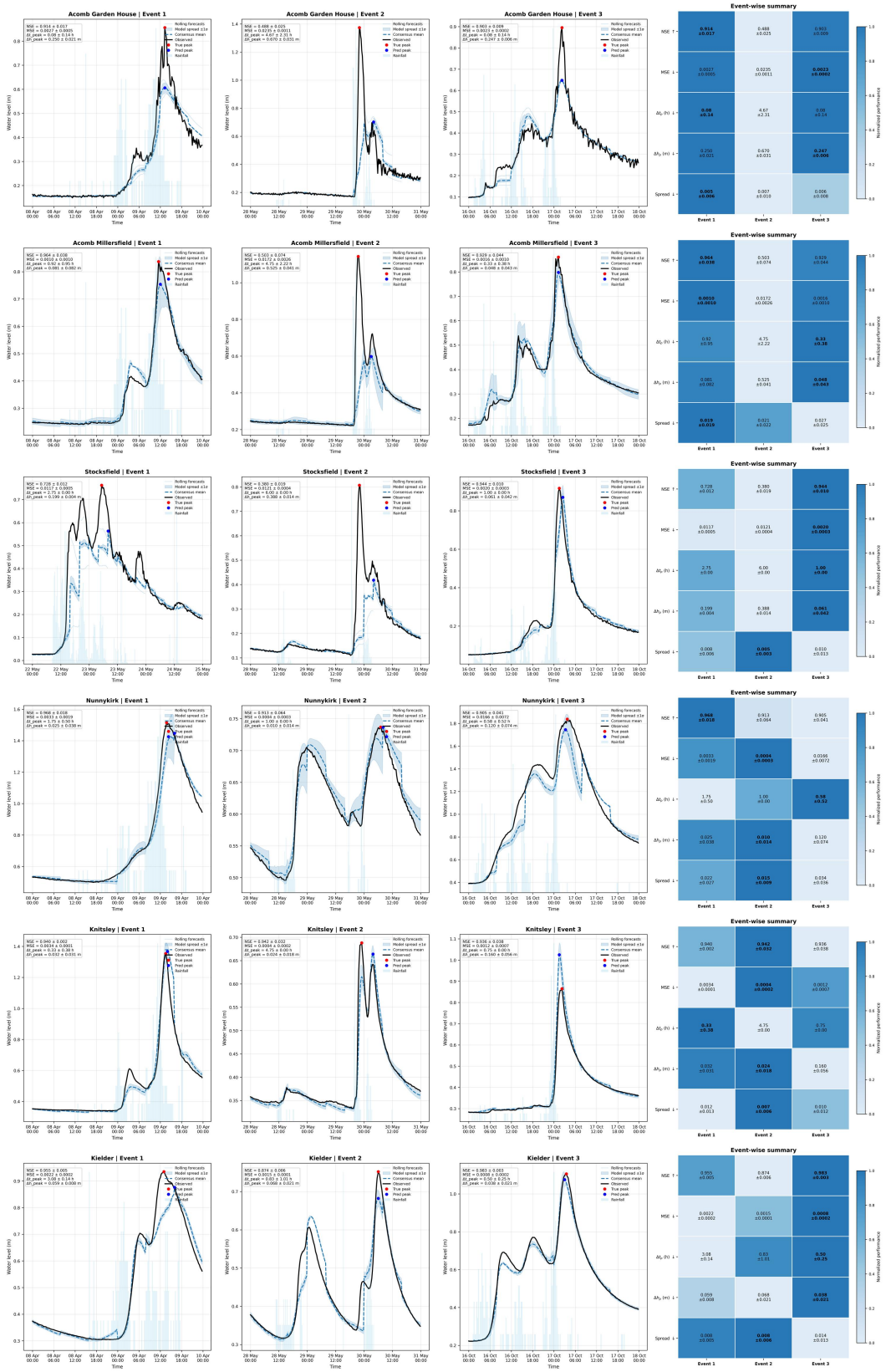


Figure 5: Event-level forecasting visualizations across the six hydrological sites.

Table 8: Recurrent-backbone benchmark under the same joint direct multi-output forecasting protocol. All models receive identical historical inputs and the same 32-step future-rainfall forcing and predict the complete 32-step water-level trajectory. We report event-level and full-test MSE (↓) and NSE (↑). ✗ indicates that no observed event exceeded the station-specific flood threshold during the test period.

Data	Metric	GaugeCast++ (Ours) MSE↓±unc., Δ%	GaugeCast++ (Ours) NSE↑±unc., Δ	BiLSTM MSE↓	BiLSTM NSE↑	xLSTM MSE↓	xLSTM NSE↑	LSTM MSE↓	LSTM NSE↑	GRU MSE↓	GRU NSE↑
Acomb Grn	Event 1	$\mathbf{0.032}_{\pm 0.004}{}^{-30.4\%}$	$\mathbf{0.933}_{\pm 0.014}{}^{+0.030}$	**0.046**	**0.903**	0.088	0.815	0.053	0.889	0.074	0.845
	Event 2	$\mathbf{0.075}_{\pm 0.002}{}^{-27.9\%}$	$\mathbf{0.895}_{\pm 0.037}{}^{+0.041}$	0.129	0.819	0.164	0.770	**0.104**	**0.854**	0.105	0.853
	Event 3	$\mathbf{0.577}_{\pm 0.009}{}^{-3.5\%}$	$\mathbf{0.586}_{\pm 0.020}{}^{+0.015}$	**0.598**	**0.571**	0.865	0.379	0.635	0.544	0.706	0.493
	Test	$\mathbf{0.008}_{\pm 0.000}{}^{-27.3\%}$	$\mathbf{0.922}_{\pm 0.001}{}^{+0.029}$	**0.011**	**0.893**	0.014	0.869	0.012	0.890	0.012	0.892
Acomb Mfs	Event 1	$\mathbf{0.034}_{\pm 0.014}{}^{-17.1\%}$	$\mathbf{0.927}_{\pm 0.065}{}^{+0.014}$	0.053	0.888	0.044	0.906	0.055	0.883	**0.041**	**0.913**
	Event 2	$\mathbf{0.017}_{\pm 0.008}{}^{-75.0\%}$	$\mathbf{0.972}_{\pm 0.031}{}^{+0.087}$	0.095	0.840	**0.068**	**0.885**	0.094	0.842	0.085	0.857
	Event 3	$\mathbf{0.402}_{\pm 0.018}{}^{-39.4\%}$	$\mathbf{0.606}_{\pm 0.053}{}^{+0.255}$	0.765	0.251	0.731	0.284	**0.663**	**0.351**	0.758	0.258
	Test	$\mathbf{0.006}_{\pm 0.000}{}^{-0.0\%}$	$\mathbf{0.930}_{\pm 0.005}{}^{+0.011}$	0.008	0.904	**0.006**	**0.919**	0.007	0.907	0.007	0.916
Stocksfield	Event 1	$\mathbf{0.044}_{\pm 0.000}{}^{-69.0\%}$	$\mathbf{0.943}_{\pm 0.008}{}^{+0.125}$	0.161	0.793	0.226	0.710	**0.142**	**0.818**	0.194	0.750
	Event 2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Event 3	$\mathbf{0.264}_{\pm 0.003}{}^{-39.2\%}$	$\mathbf{0.542}_{\pm 0.013}{}^{+0.293}$	0.495	0.142	0.492	0.147	**0.434**	**0.249**	0.471	0.184
	Test	$\mathbf{0.013}_{\pm 0.001}{}^{-31.6\%}$	$\mathbf{0.881}_{\pm 0.008}{}^{+0.056}$	**0.019**	**0.825**	0.023	0.789	0.023	0.794	0.021	0.809
Nunnykirk	Event 1	$\mathbf{0.027}_{\pm 0.019}{}^{-84.8\%}$	$\mathbf{0.968}_{\pm 0.018}{}^{+0.179}$	0.245	0.709	0.263	0.687	**0.178**	**0.789**	0.220	0.739
	Event 2	$\mathbf{0.050}_{\pm 0.003}{}^{-49.0\%}$	$\mathbf{0.913}_{\pm 0.064}{}^{+0.084}$	0.175	0.695	0.112	0.805	**0.098**	**0.829**	0.118	0.795
	Event 3	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	Test	$\mathbf{0.003}_{\pm 0.000}{}^{-50.0\%}$	$\mathbf{0.972}_{\pm 0.005}{}^{+0.029}$	0.007	0.936	**0.006**	**0.943**	**0.006**	0.942	0.007	0.933
Knitsley	Event 1	$\mathbf{0.011}_{\pm 0.000}{}^{-52.2\%}$	$\mathbf{0.940}_{\pm 0.002}{}^{+0.058}$	**0.023**	0.881	0.024	0.874	**0.023**	**0.882**	**0.023**	0.878
	Event 2	$\mathbf{0.035}_{\pm 0.024}{}^{-39.7\%}$	$\mathbf{0.942}_{\pm 0.032}{}^{+0.038}$	**0.058**	**0.904**	0.145	0.761	**0.058**	**0.904**	0.099	0.837
	Event 3	$\mathbf{0.007}_{\pm 0.002}{}^{-50.0\%}$	$\mathbf{0.936}_{\pm 0.038}{}^{+0.068}$	**0.014**	**0.868**	**0.014**	0.867	0.016	0.851	0.020	0.820
	Test	$\mathbf{0.003}_{\pm 0.000}{}^{-40.0\%}$	$\mathbf{0.945}_{\pm 0.013}{}^{+0.036}$	**0.005**	0.906	**0.005**	0.906	0.006	0.898	**0.005**	**0.909**
Kielder	Event 1	$\mathbf{0.009}_{\pm 0.000}{}^{-35.7\%}$	$\mathbf{0.955}_{\pm 0.005}{}^{+0.026}$	0.018	0.907	**0.014**	**0.929**	0.022	0.886	0.033	0.827
	Event 2	$\mathbf{0.004}_{\pm 0.001}{}^{-76.5\%}$	$\mathbf{0.983}_{\pm 0.003}{}^{+0.061}$	0.022	0.903	0.030	0.864	0.026	0.883	**0.017**	**0.922**
	Event 3	$\mathbf{0.007}_{\pm 0.000}{}^{-73.1\%}$	$\mathbf{0.874}_{\pm 0.006}{}^{+0.337}$	0.032	0.426	0.027	0.524	0.032	0.439	**0.026**	**0.537**
	Test	$\mathbf{0.003}_{\pm 0.000}{}^{-25.0\%}$	$\mathbf{0.968}_{\pm 0.006}{}^{+0.021}$	0.006	0.929	0.006	0.922	0.006	0.931	**0.004**	**0.947**
Best (↑)	**Count**	22	22	0	0	1	0	0	0	0	0

Table 9: Main notation used throughout the paper.

Symbol	Meaning
$\mathcal{D}$	training dataset of input–output examples
$\mathbf{X} \in \mathbb{R}^{L \times d_x}$	historical context sequence
$\mathbf{U} \in \mathbb{R}^{H \times d_u}$	future conditioning sequence
$\mathbf{Y} \in \mathbb{R}^{H \times d_y}$	target future trajectory
L	input/context length
H	prediction horizon length
F_Θ	sequence-to-trajectory predictor
$\widehat{\mathbf{Y}}$	predicted future trajectory
$\Omega = (0,1) \times (0,1)$	canonical latent spacetime domain
$\xi \in (0,1)$	latent spatial coordinate
$\tau \in (0,1)$	normalized forecast-time coordinate
$u_\Theta : \Omega \to \mathbb{R}^{d_z}$	latent field induced by the predictor
$\mathcal{R}$	readout operator from latent field to discrete forecast
$a(\xi, \tau; \beta)$	latent transport coefficient
$\nu(\xi, \tau; \beta)$	latent diffusion coefficient
$g(\xi, \tau; \beta)$	latent forcing/source term
β	sample-conditioned operator coefficient tuple
$\phi_\psi(\xi, \tau)$	learned gauge map / latent coordinate transform
$y = \phi_\psi(\xi, \tau)$	transformed latent coordinate
$v_\theta(y, \tau)$	latent solution in transformed coordinates
$s_\psi = \partial_\xi \phi_\psi$	spatial derivative of the gauge map
$m_\psi = \partial_\tau \phi_\psi$	temporal derivative of the gauge map
$B_\psi(\xi, \tau; \beta)$	gauge residual / residual transport field
$\mathrm{Pe}_\psi(\xi, \tau; \beta)$	transformed Peclet number
ℓ	characteristic latent length scale
Θ	parameters of the forecasting architecture
θ	parameters of the latent solution network
ψ	parameters of the gauge map

D Gauge-Conditioned Pullback of the Transport–Diffusion Operator

This appendix presents the full mathematical development of the gauge-conditioned pullback formulation introduced in the main text. The key idea is to learn a diffeomorphic coordinate map that acts as an operator preconditioner: after pullback, the first-order transport component can be explicitly controlled and, in the ideal case, annihilated. This converts the original transport–diffusion relation into a geometry-modulated parabolic equation in latent coordinates.

D.1 Setup and standing assumptions

Let $I_x = (0, L)$, $I_y = (0, L_y)$, and $t \in (0, T)$. We consider

$$\Omega_x := I_x \times (0, T).$$

For each fixed t, the learned gauge map

$$\phi_{\psi,t} : I_x \to I_y, \qquad \phi_{\psi,t}(x) := \phi_\psi(x, t),$$

is assumed to be a C^2 diffeomorphism onto the fixed latent interval I_y. We consider the transport–diffusion model

$$\partial_t h(x,t) + a(x,t)\,\partial_x h(x,t) - \nu(x,t)\,\partial_{xx} h(x,t) = f(x,t), \qquad (x,t) \in I_x \times (0,T). \tag{25}$$

Throughout the appendix, we assume

$$a, \nu, f \in C^1(\overline{\Omega_x}), \qquad \nu(x,t) \geq \nu_0 > 0, \tag{26}$$

$$\phi_\psi \in C^{2,1}(\overline{\Omega_x}), \qquad 0 < s_0 \leq \partial_x \phi_\psi(x,t) \leq s_1 < \infty, \tag{27}$$

$$\phi_\psi(0,t) = 0, \qquad \phi_\psi(L,t) = L_y, \tag{28}$$

$$v_\theta \in C^{2,1}(I_y \times (0,T)). \tag{29}$$

The lower bound on $\partial_x \phi_\psi$ ensures that $x \mapsto \phi_\psi(x,t)$ is invertible for each fixed t. We denote the inverse map by

$$\chi_\psi(y,t) := \phi_{\psi,t}^{-1}(y).$$

D.2 Pullback representation and exact derivative identities

We introduce the latent coordinate $y = \phi_\psi(x,t)$, and represent the state as

$$h(x,t) = v_\theta(\phi_\psi(x,t), t) = v_\theta(y,t) \tag{30}$$

Define the gauge Jacobian and temporal drift by

$$s_\psi(x,t) := \partial_x \phi_\psi(x,t), \qquad m_\psi(x,t) := \partial_t \phi_\psi(x,t) \tag{31}$$

Under Assumption (27), $s_\psi(x,t) > 0$, so the map $x \mapsto \phi_\psi(x,t)$ is locally invertible for each fixed t.

Lemma D.1 (Exact pullback derivatives)**.** *If $h(x,t) = v_\theta(\phi_\psi(x,t), t)$, then*

$$\partial_x h(x,t) = s_\psi(x,t)\,\partial_y v_\theta(y,t), \tag{32}$$

$$\partial_t h(x,t) = \partial_t v_\theta(y,t) + m_\psi(x,t)\,\partial_y v_\theta(y,t), \tag{33}$$

$$\partial_{xx} h(x,t) = \partial_x s_\psi(x,t)\,\partial_y v_\theta(y,t) + s_\psi(x,t)^2\,\partial_{yy} v_\theta(y,t) \tag{34}$$

Proof. From (30),

$$\partial_x h(x,t) = \partial_y v_\theta(y,t)\,\partial_x \phi_\psi(x,t) = s_\psi(x,t)\,\partial_y v_\theta(y,t),$$

which proves (32). Similarly,

$$\partial_t h(x,t) = \partial_t v_\theta(y,t) + \partial_y v_\theta(y,t)\,\partial_t \phi_\psi(x,t) = \partial_t v_\theta(y,t) + m_\psi(x,t)\,\partial_y v_\theta(y,t),$$

which proves (33). Differentiating (32) once more in x yields

$$\partial_{xx} h = \partial_x(s_\psi\,\partial_y v_\theta) = (\partial_x s_\psi)\,\partial_y v_\theta + s_\psi\,\partial_x(\partial_y v_\theta)$$

Applying the chain rule again,

$$\partial_x(\partial_y v_\theta(y,t)) = \partial_{yy} v_\theta(y,t)\,\partial_x \phi_\psi(x,t) = s_\psi(x,t)\,\partial_{yy} v_\theta(y,t),$$

$$\Rightarrow \partial_{xx} h(x,t) = \partial_x s_\psi(x,t)\,\partial_y v_\theta(y,t) + s_\psi(x,t)^2\,\partial_{yy} v_\theta(y,t). \qquad \square$$

The identities above are exact and require no approximation. They show that the pullback separates the transformed dynamics into a Jacobian-weighted first derivative and a geometry-weighted second derivative.

For later use, define the inverse gauge map and pulled-back log-Jacobian

$$\chi_\psi(y,t) := \phi_{\psi,t}^{-1}(y), \qquad \widetilde{\rho}_\psi(y,t) := \log s_\psi(\chi_\psi(y,t),t).$$

Equivalently,

$$\widetilde{\rho}_\psi(\phi_\psi(x,t),t) = \log s_\psi(x,t).$$

Lemma D.2 (Log-Jacobian form of the pulled-back curvature)**.** *Along* $y = \phi_\psi(x,t)$*,*

$$\partial_x s_\psi = s_\psi^2\, \partial_y \widetilde{\rho}_\psi, \tag{35}$$

and therefore

$$\partial_{xx} h = s_\psi^2 \left(\partial_{yy} v_\theta + (\partial_y \widetilde{\rho}_\psi)\, \partial_y v_\theta\right). \tag{36}$$

Proof. Since

$$\widetilde{\rho}_\psi(\phi_\psi(x,t),t) = \log s_\psi(x,t),$$

differentiation in x gives

$$(\partial_y \widetilde{\rho}_\psi)(\partial_x \phi_\psi) = (\partial_x s_\psi) s_\psi^{-1}.$$

Using $\partial_x \phi_\psi = s_\psi$,

$$(\partial_y \widetilde{\rho}_\psi) s_\psi = (\partial_x s_\psi) s_\psi^{-1} \quad \Rightarrow \quad \partial_y \widetilde{\rho}_\psi = (\partial_x s_\psi) s_\psi^{-2} \quad \Rightarrow \quad \partial_x s_\psi = s_\psi^2\, \partial_y \widetilde{\rho}_\psi.$$

Substituting into Lemma D.1,

$$\partial_{xx} h = (\partial_x s_\psi)\partial_y v_\theta + s_\psi^2 \partial_{yy} v_\theta = s_\psi^2 \left(\partial_{yy} v_\theta + (\partial_y \widetilde{\rho}_\psi)\partial_y v_\theta\right) \qquad \square$$

Equation (36) makes explicit that the second-order pullback contains not only the canonical curvature term $\partial_{yy} v_\theta$, but also an additional first-order correction induced by spatial variation of the log-Jacobian. This correction is the geometric source of the residual transport term that appears in the pulled-back operator.

For later reference, we may summarize the pullback at the operator level by writing

$$\partial_x = s_\psi \partial_y, \qquad \partial_{xx} = s_\psi^2 \partial_{yy} + (\partial_x s_\psi)\partial_y = s_\psi^2 \left(\partial_{yy} + (\partial_y \widetilde{\rho}_\psi)\partial_y\right) \tag{37}$$

D.3 Pulled-back operator and the gauge residual

Substituting Lemma D.1 into (25) yields the exact latent-coordinate form of the operator.

Proposition D.3 (Exact pulled-back operator)**.** *Define*

$$B_\psi(x,t) := \partial_t \phi_\psi(x,t) + a(x,t)\, \partial_x \phi_\psi(x,t) - \nu(x,t)\, \partial_{xx} \phi_\psi(x,t) \tag{38}$$

Then the pullback representation (30) *transforms* (25) *into*

$$\begin{aligned} &\partial_t v_\theta(y,t) - \nu(x,t) s_\psi(x,t)^2\, \partial_{yy} v_\theta(y,t) + B_\psi(x,t)\, \partial_y v_\theta(y,t) = f(x,t) \\ \iff \quad &\partial_t v_\theta(y,t) - \widetilde{\nu}_\psi(y,t) \widetilde{s}_\psi(y,t)^2\, \partial_{yy} v_\theta(y,t) + \widetilde{B}_\psi(y,t)\, \partial_y v_\theta(y,t) = \widetilde{f}_\psi(y,t). \end{aligned} \tag{39}$$

Proof. Using (32), (33), and (34) in (25), we obtain

$$\partial_t v_\theta + m_\psi\, \partial_y v_\theta + a\, s_\psi\, \partial_y v_\theta - \nu \left((\partial_x s_\psi)\, \partial_y v_\theta + s_\psi^2\, \partial_{yy} v_\theta\right) = f$$

Collecting the $\partial_y v_\theta$ terms gives

$$\partial_t v_\theta - \nu s_\psi^2\, \partial_{yy} v_\theta + (m_\psi + a s_\psi - \nu\, \partial_x s_\psi)\, \partial_y v_\theta = f \qquad \square$$

Using $m_\psi = \partial_t \phi_\psi$ and $\partial_x s_\psi = \partial_{xx} \phi_\psi$ yields (39).

Equation (39) shows that the learned coordinate map enters the transformed dynamics through two distinct mechanisms: 1. The geometry-modulated diffusion coefficient νs_ψ^2; 2. The residual transport coefficient B_ψ.

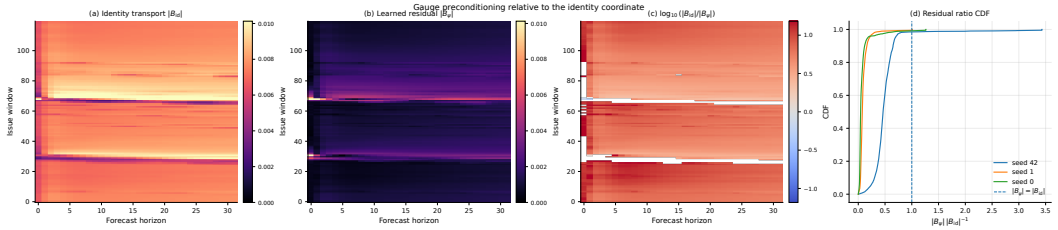


Figure 6: Gauge preconditioning relative to the identity coordinate. For the identity map $\phi(x,t) = x$, the residual transport coefficient is $B_{\mathrm{id}} = a$. For the learned gauge, the residual coefficient is $B_\psi = \partial_t \phi_\psi + a\partial_x \phi_\psi - \nu \partial_{xx} \phi_\psi$. Panels compare the identity residual, learned gauge residual, log-scale transport reduction, and the empirical CDF of $|B_\psi|\,|B_{\mathrm{id}}|^{-1}$ across seeds and diagnostic windows. Values below one indicate that the learned gauge reduces explicit transport relative to the identity coordinate. In this diagnostic set, the median residual ratio is 0.090, and 99.03% of issue–horizon entries satisfy $|B_\psi| < |B_{\mathrm{id}}|$.

D.4 Transport-flattening gauge principle

The central principle of our method is to learn ϕ_ψ so that the residual transport term in (39) vanishes.

Definition D.4 (Transport-flattening gauge condition)**.** *The gauge map ϕ_ψ is said to satisfy the transport-flattening gauge condition if*

$$\partial_t \phi_\psi(x,t) + a(x,t)\,\partial_x \phi_\psi(x,t) - \nu(x,t)\,\partial_{xx}\phi_\psi(x,t) = 0 \qquad \text{on } \Omega. \tag{40}$$

Equivalently, $B_\psi(x,t) \equiv 0$.

Corollary D.5 (Canonical latent parabolic form)**.** *If (40) holds exactly, then (39) reduces to*

$$\partial_t v_\theta(y,t) - \nu(x,t)\, s_\psi(x,t)^2\, \partial_{yy} v_\theta(y,t) = f(x,t) \tag{41}$$

Proof. Immediate from Proposition D.3 by setting $B_\psi \equiv 0$. □

Thus, the learned gauge map acts as an operator preconditioner: it removes the explicit transport term from the non-divergence latent equation, while the remaining geometry enters through the modulated diffusion coefficient $D_\psi = \nu s_\psi^2$.

D.5 Transformed Peclet number and stiffness reduction

To quantify the remaining transport dominance after pullback, we introduce a transformed Peclet number.

Definition D.6 (Transformed Peclet number)**.** *Let $\ell > 0$ be a characteristic latent length scale. The transformed Peclet number is*

$$\mathrm{Pe}_\psi(x,t) := |B_\psi(x,t)|\, \ell\, \nu(x,t)^{-1} s_\psi(x,t)^{-2} \tag{42}$$

Proposition D.7 (Residual transport controls latent stiffness)**.** *Assume*

$$\nu(x,t) \geq \nu_0 > 0, \qquad s_\psi(x,t) \geq s_0 > 0 \qquad \forall (x,t) \in \Omega$$

If $\|B_\psi\|_{L^\infty(\Omega)} \leq \varepsilon$, then

$$\mathrm{Pe}_\psi(x,t) \leq \varepsilon\, \ell\, \nu_0^{-1} s_0^{-2}, \qquad \forall (x,t) \in \Omega. \tag{43}$$

In particular, if $B_\psi \equiv 0$, then $\mathrm{Pe}_\psi \equiv 0$.

Proof. From (42),

$$\mathrm{Pe}_\psi(x,t) = |B_\psi(x,t)|\, \ell\, \nu(x,t)^{-1} s_\psi(x,t)^{-2} \leq \|B_\psi\|_{L^\infty(\Omega)}\, \ell\, \nu_0^{-1} s_0^{-2} \leq \varepsilon\, \ell\, \nu_0^{-1} s_0^{-2}$$ □

Proposition D.7 shows that minimizing the gauge residual directly reduces the transport dominance of the transformed operator.

D.6 Energy stability under approximate gauge satisfaction

We now derive an energy estimate for the pulled-back latent equation under approximate gauge satisfaction. Since the diffusion coefficient in latent coordinates is spatially varying, the correct coercive estimate must account not only for the residual transport field B_ψ, but also for the spatial variation of the effective latent diffusivity.

Define

$$D_\psi(x,t) := \nu(x,t)\, s_\psi(x,t)^2, \tag{44}$$

and introduce the effective lower-order coefficient

$$\widetilde{B}_\psi(x,t) := B_\psi(x,t) + \partial_y D_\psi(x,t) \tag{45}$$

Then the pulled-back equation $\partial_t v - \nu s_\psi^2\, \partial_{yy} v + B_\psi\, \partial_y v = f$ can be equivalently written in divergence form as

$$\partial_t v - \partial_y\big(D_\psi\, \partial_y v\big) + \widetilde{B}_\psi\, \partial_y v = f \tag{46}$$

Let v satisfy (46) on the fixed latent interval I_y.

Theorem D.8 (Energy stability under approximate gauge satisfaction)**.** *Assume $0 < s_0 \le s_\psi(x,t) \le s_1 < \infty$, $\nu(x,t) \ge \nu_0 > 0$, and let $D_0 = \nu_0 s_0^2$. Let v satisfy*

$$\partial_t v - \partial_y(D_\psi \partial_y v) + \widetilde{B}_\psi \partial_y v = f$$

on the fixed latent interval I_y, with boundary conditions such that integration-by-parts boundary terms vanish. Then

$$\frac{1}{2}\frac{d}{dt}\|v(\cdot,t)\|^2_{L^2(I_y)} + \frac{D_0}{2}\|\partial_y v(\cdot,t)\|^2_{L^2(I_y)} \le \frac{1}{2}\|f(\cdot,t)\|^2_{L^2(I_y)} + \left(\frac{1}{2} + \frac{\|\widetilde{B}_\psi(\cdot,t)\|^2_{L^\infty(I_y)}}{2D_0}\right)\|v(\cdot,t)\|^2_{L^2(I_y)}.$$

The main text states the simplified consequence in terms of B_ψ. The sharper appendix estimate shows that, for variable geometry, the lower-order perturbation is controlled by $\widetilde{B}_\psi = B_\psi + \partial_y D_\psi$. In our implementation, smoothness and diffeomorphism regularization control $\partial_y D_\psi$, while the gauge loss directly suppresses the explicit residual transport B_ψ.

Proof. Starting from (46), multiply by v and integrate over the latent domain $\phi_\psi(\Omega)$:

$$\int v\, \partial_t v\, dy - \int v\, \partial_y(D_\psi\, \partial_y v)\, dy + \int \widetilde{B}_\psi\, v\, \partial_y v\, dy = \int f v\, dy. \Rightarrow \int v\, \partial_t v\, dy = \frac{1}{2}\frac{d}{dt}\|v\|^2_{L^2}.$$

Integrating the diffusion term by parts and using the vanishing boundary contribution,

$$-\int v\, \partial_y(D_\psi\, \partial_y v)\, dy = \int D_\psi\, |\partial_y v|^2\, dy.$$

$$D_\psi(x,t) = \nu(x,t) s_\psi(x,t)^2 \ge \nu_0 s_0^2 = D_0, \Rightarrow \int D_\psi\, |\partial_y v|^2\, dy \ge D_0 \|\partial_y v\|^2_{L^2}$$

For the lower-order perturbation term,

$$\left|\int \widetilde{B}_\psi\, v\, \partial_y v\, dy\right| \le \|\widetilde{B}_\psi\|_{L^\infty}\, \|v\|_{L^2}\, \|\partial_y v\|_{L^2}$$

Applying Young's inequality with parameter D_0 yields

$$\|\widetilde{B}_\psi\|_{L^\infty}\, \|v\|_{L^2}\, \|\partial_y v\|_{L^2} \le \frac{D_0}{2}\|\partial_y v\|^2_{L^2} + \frac{\|\widetilde{B}_\psi\|^2_{L^\infty}}{2D_0}\|v\|^2_{L^2}. \simeq \int f v\, dy \le \frac{1}{2}\|f\|^2_{L^2} + \frac{1}{2}\|v\|^2_{L^2}$$

Combining the above estimates gives

$$\frac{1}{2}\frac{d}{dt}\|v\|^2_{L^2} + D_0\|\partial_y v\|^2_{L^2} \le \frac{1}{2}\|f\|^2_{L^2} + \frac{1}{2}\|v\|^2_{L^2} + \frac{D_0}{2}\|\partial_y v\|^2_{L^2} + \frac{\|\widetilde{B}_\psi\|^2_{L^\infty}}{2D_0}\|v\|^2_{L^2} \qquad \square$$

which immediately yields (45).

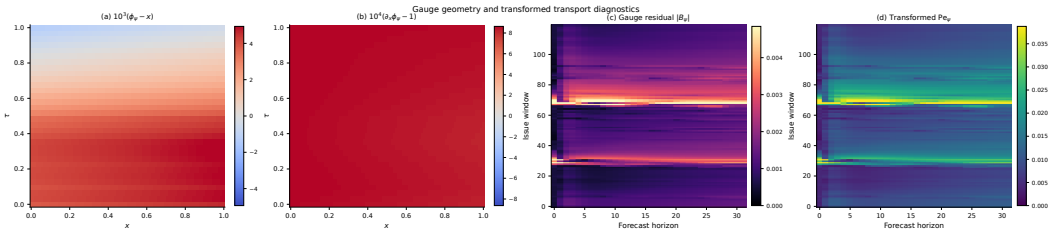


Figure 7: Gauge-geometry and transformed-transport diagnostics for held-out forecast windows. Panels show the scaled coordinate displacement $10^3(\phi_\psi - x)$, the scaled Jacobian deviation $10^4(\partial_x\phi_\psi - 1)$, the gauge residual magnitude $|B_\psi|$, and the transformed Peclet number Pe_ψ. The learned coordinate correction remains smooth and near-identity, while the Jacobian remains positive and close to one, indicating no folding of the learned gauge map. Residual and Peclet heatmaps show that transport-dominated corrections are localized in issue time and forecast horizon.

The estimate (45) shows that approximate gauge satisfaction leads to a stable latent evolution, with the perturbation controlled by the effective lower-order coefficient $\widetilde{B}_\psi = B_\psi + \partial_y D_\psi$. In particular, exact cancellation of the explicit transport term B_ψ still improves conditioning, but the sharp energy bound depends on both the residual transport and the spatial variation of the geometry-modulated diffusivity.

D.7 Diffeomorphism preservation

The usefulness of the gauge map depends on preserving invertibility throughout training.

Proposition D.9 (Sufficient condition for diffeomorphism preservation)**.** *Assume that $\phi_\psi \in C^{3,1}(\overline{\Omega_x})$ satisfies the exact gauge equation*

$$\partial_t\phi_\psi + a\partial_x\phi_\psi - \nu\partial_{xx}\phi_\psi = 0,$$

and define $s_\psi = \partial_x\phi_\psi$. Then s_ψ satisfies

$$\partial_t s_\psi + \partial_x(a s_\psi) - \partial_x(\nu\partial_x s_\psi) = 0.$$

If a, ν are bounded with bounded derivatives, $\nu \geq \nu_0 > 0$, and the initial and boundary data satisfy $s_\psi(x,0) > 0$ and $s_\psi > 0$ on the parabolic boundary, then the parabolic maximum principle implies $s_\psi(x,t) > 0$ on Ω_x. Hence $x \mapsto \phi_\psi(x,t)$ remains strictly monotone for each t.

If the coefficients are regular, $\nu \geq \nu_0 > 0$, and the initial and boundary data satisfy $s_\psi(x,0) > 0$, with compatible positive boundary conditions, then $s_\psi(x,t) > 0$ for all $(x,t) \in \Omega$. Hence $x \mapsto \phi_\psi(x,t)$ remains strictly monotone for each t.

Proof. Differentiate (40) with respect to x:

$$\partial_x\partial_t\phi_\psi + \partial_x(a\,\partial_x\phi_\psi) - \partial_x(\nu\,\partial_{xx}\phi_\psi) = 0 \qquad \square$$

Since $s_\psi = \partial_x\phi_\psi$, this stands. Positivity then follows from the parabolic maximum principle under the stated assumptions.

D.8 Variational weak form

Let $D_\psi(y,t) := \widetilde{\nu}_\psi(y,t)\widetilde{s}_\psi(y,t)^2$, and let

$$\widetilde{B}_\psi(y,t) := B_\psi(\chi_\psi(y,t),t) + \partial_y D_\psi(y,t).$$

The pulled-back equation can be written in divergence form as

$$\partial_t v - \partial_y(D_\psi\partial_y v) + \widetilde{B}_\psi\partial_y v = \widetilde{f}_\psi.$$

For any test function $\varphi \in H_0^1(I_y)$, the weak form is

$$\int_0^T \int_{I_y} \Big(\partial_t v\,\varphi + D_\psi\,\partial_y v\,\partial_y\varphi + \widetilde{B}_\psi\,\partial_y v\,\varphi - \widetilde{f}_\psi\,\varphi\Big)\,dy\,dt = 0. \tag{47}$$

Under exact gauge satisfaction $B_\psi = 0$, the explicit transport contribution vanishes, but the weak form still contains the geometric lower-order term $\partial_y D_\psi \, \partial_y v \, \varphi$ whenever the effective latent diffusivity is spatially variable. If D_ψ is constant or slowly varying, this reduces to the standard weighted parabolic weak form.

E Technical Appendix: Semigroup-Consistent Phase-Transport Learning

This appendix provides the mathematical derivation and implementation details underlying the semigroup-consistent phase-transport regularization introduced in Section 3.4. The central goal is to control *horizon-dependent phase drift* by requiring that horizon-indexed forecasting operators compose consistently in both trajectory space and event-phase space.

E.1 Operator-theoretic formulation

Let x_t denote the historical input window ending at time t, and let $\mathcal{T}_\Delta$ denote the learned forecasting operator over lead time Δ, so that $\hat{h}^{(\Delta)}(x_t) := \mathcal{T}_\Delta(x_t) \in \mathbb{R}^\Delta$, where $\hat{h}^{(\Delta)}(x_t)$ is the predicted stage trajectory of length Δ. In an exact time-homogeneous dynamical system, one expects the semigroup property $\mathcal{T}_{\Delta_1+\Delta_2} = \mathcal{T}_{\Delta_2} \circ \mathcal{T}_{\Delta_1}$, meaning that a direct long-horizon forecast equals the composition of two shorter transitions. In learned forecasting systems, this relation typically fails, especially at long horizons, which leads to accumulated temporal misalignment.

To quantify this inconsistency, define a continuation operator $\mathcal{C}_n$ that rolls the model forward for another n steps from the intermediate forecast state produced at horizon m. The *semigroup defect* is then

$$\mathcal{E}_{m,n}(x_t) := \mathcal{T}_{m+n}(x_t) - \mathcal{C}_n\big(\mathcal{T}_m(x_t), x_t\big) \tag{48}$$

If exact semigroup held exactly, then $\mathcal{E}_{m,n}(x_t) = 0$ for all admissible x_t.

Implemented prefix-consistency surrogate. The continuation operator C_n in Eq. (48) describes the ideal operator-level semigroup relation. In the implemented regularizer, however, we do not construct a continuation trajectory using future ground-truth observations. Instead, for a full model rollout $T_P(x_t) \in \mathbb{R}^P$ and a shorter rollout $T_m(x_t) \in \mathbb{R}^m$, with $m < P$, we compare the short prediction against the corresponding prefix of the full prediction.

Let $\Pi_m : \mathbb{R}^P \to \mathbb{R}^m$ denote the prefix operator retaining the first m forecast steps. The implemented trajectory-consistency defect is

$$\widetilde{E}_m(x_t) = T_m(x_t) - \mathrm{sg}[\Pi_m T_P(x_t)] \,, \tag{49}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Thus, both trajectories are generated solely from the same historical input x_t; no future observed water-level values are inserted into either rollout.

The corresponding phase-consistency defect is

$$\widetilde{E}^{\Phi}_m(x_t) = \Phi(T_m(x_t)) - \mathrm{sg}[\Pi_m \Phi(T_P(x_t))] \,. \tag{50}$$

These prefix-consistency terms provide the practical training surrogate for the ideal semigroup relation while avoiding teacher-forced future-stage continuation.

E.2 Phase-transport map

Trajectory-level semigroup consistency alone is insufficient for long-horizon event prediction, because two hydrographs may be close in pointwise norm while still exhibiting shifted peaks. To make event timing explicit, we define a differentiable phase representation.

For a forecast trajectory $h = (h_1, \ldots, h_P) \in \mathbb{R}^P$, define the soft peak weights

$$s_j(h) = \exp(\beta h_j)\Big(\sum_{\ell=1}^{P} \exp(\beta h_\ell)\Big)^{-1}, \qquad j = 1, \ldots, P \tag{51}$$

where $\beta > 0$ controls concentration around high-stage steps. Since $s_j(h) \geq 0$ and $\sum_{j=1}^{P} s_j(h) = 1$, the vector $s(h)$ defines a probability distribution over forecast time.

We then define the cumulative phase map

$$\Phi_k(h) = \sum_{j=1}^{k} s_j(h) = \left(\sum_{j=1}^{k} \exp(\beta h_j)\right)\left(\sum_{j=1}^{P} \exp(\beta h_j)\right)^{-1}, \qquad k = 1, \dots, P \tag{52}$$

The map $\Phi(h) = (\Phi_1(h), \dots, \Phi_P(h))$ is monotone, bounded in $[0,1]^P$, and satisfies $\Phi_P(h) = 1$. It may be interpreted as a smooth cumulative transport map of peak-sensitive mass across the horizon.

The corresponding phase-consistency defect is

$$\mathcal{E}_m^{\Phi}(x_t) := \Phi_m(\mathcal{T}_m(x_t)) - \Phi_m(\Pi_m \mathcal{T}_P(x_t)) \rightharpoonup \mathcal{L}_\Phi = \|\Phi_m(\mathcal{T}_m(x_t)) - \mathrm{sg}[\Phi_m(\Pi_m \mathcal{T}_P(x_t))]\|_2^2. \tag{53}$$

where Φ_m denotes the phase map applied to an m-step trajectory.

Soft crest-time functional. A smooth crest-time surrogate can be obtained from the same weights:

$$\begin{aligned} \widetilde{\tau}_\beta(h) &:= \sum_{j=1}^{P} j\, s_j(h), s_j(h) \qquad\qquad = \Phi_j(h) - \Phi_{j-1}(h), \qquad \Phi_0(h) := 0, \\ \Rightarrow \quad \widetilde{\tau}_\beta(h) &= \sum_{j=1}^{P} j\big(\Phi_j(h) - \Phi_{j-1}(h)\big) \end{aligned} \tag{54}$$

This identity is useful because it makes the soft crest time an explicit functional of the cumulative phase map.

Lemma E.1 (Phase control of the soft crest time)**.** *For any two trajectories $h, g \in \mathbb{R}^P$,*

$$\big|\widetilde{\tau}_\beta(h) - \widetilde{\tau}_\beta(g)\big| \le (P-1)\, \|\Phi(h) - \Phi(g)\|_\infty. \tag{55}$$

Proof. Let

$$\Delta_k := \Phi_k(h) - \Phi_k(g), \qquad \Delta_0 = 0, \qquad \Delta_P = 0,$$

because both phase maps end at one. Since

$$s_j(h) - s_j(g) = \Delta_j - \Delta_{j-1},$$

we have

$$\widetilde{\tau}_\beta(h) - \widetilde{\tau}_\beta(g) = \sum_{j=1}^{P} j(\Delta_j - \Delta_{j-1}).$$

Applying summation by parts,

$$\sum_{j=1}^{P} j(\Delta_j - \Delta_{j-1}) = -\sum_{j=1}^{P-1} \Delta_j \to \big|\widetilde{\tau}_\beta(h) - \widetilde{\tau}_\beta(g)\big| \le \sum_{j=1}^{P-1} |\Delta_j| \le (P-1)\|\Phi(h) - \Phi(g)\|_\infty \quad \square$$

Lemma E.1 shows that phase-map consistency directly controls a smooth surrogate of crest-time error.

E.3 Phase drift bound

We now formalize the relationship between semigroup inconsistency and long-horizon timing error.

Assumption E.2. *For admissible trajectories, the forecasting operators $\mathcal{T}_\Delta$ and continuation operator $\mathcal{C}_n$ are locally Lipschitz in the relevant neighborhood of the data manifold.*

Assumption E.3. *The target event peak is isolated and nondegenerate, so the crest-time functional is locally stable in a neighborhood of the true trajectory.*

Assumption E.4. *There exists a constant $L_{\mathrm{hyd}} > 0$ such that the discrepancy between the composed continuation trajectory and the corresponding ideal hydrodynamic trajectory is bounded by the latent PDE residual:*

$$\big\|\mathcal{C}_n\big(\mathcal{T}_m(x_t), x_t\big) - h^\star_{m+n}(x_t)\big\|_\infty \le L_{\mathrm{hyd}}\, \|\widehat{\mathcal{R}}_{\mathrm{hydro}}\|_2, \tag{56}$$

where $h^\star_{m+n}(x_t)$ denotes the target long-horizon trajectory.

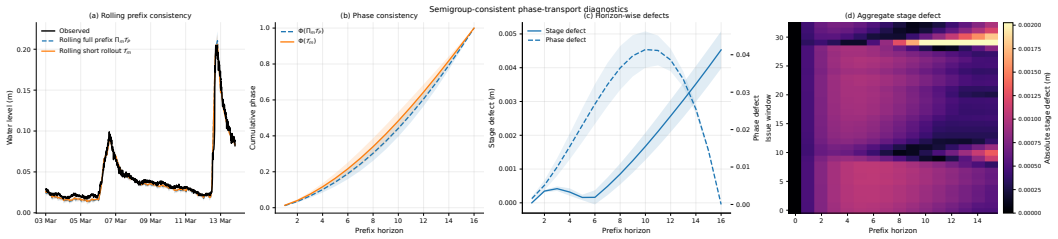


Figure 8: Semigroup-consistent phase-transport diagnostics. (a) Rolling prefix consistency over the selected evaluation period, comparing the short rollout $\mathcal{T}_m(x_t)$ against the corresponding prefix of the full rollout $\Pi_m \mathcal{T}_P(x_t)$, aggregated across issue windows and seeds. (b) Representative-window cumulative phase-map comparison. (c) Horizon-wise stage and phase defects for the representative window. (d) Aggregate absolute stage defect across diagnostic issue windows. The diagnostic matches the implemented regularizer, which compares $\mathcal{T}_m(x_t)$ to the stop-gradient prefix of $\mathcal{T}_P(x_t)$ without teacher-forced future-stage continuation.

Theorem E.5 (Phase drift bound via semigroup inconsistency)**.** *Under Assumptions E.2–E.4, there exist constants $C_1, C_2, C_3 > 0$ such that*

$$\left|\widetilde{\tau}_\beta(\mathcal{T}_{m+n}(x_t)) - \widetilde{\tau}_\beta(h^\star_{m+n}(x_t))\right| \le C_1 \|\mathcal{E}_{m,n}(x_t)\|_\infty + C_2 \|\mathcal{E}^\Phi_{m,n}(x_t)\|_\infty + C_3 \|\widehat{\mathcal{R}}_{\mathrm{hydro}}\|_2 \quad (57)$$

Proof. Let

$$\hat{h}_{m+n} := \mathcal{T}_{m+n}(x_t), \qquad \bar{h}_{m+n} := \mathcal{C}_n\big(\mathcal{T}_m(x_t), x_t\big), \qquad h^\star_{m+n} := h^\star_{m+n}(x_t)$$

By the triangle inequality,

$$\left|\widetilde{\tau}_\beta(\hat{h}_{m+n}) - \widetilde{\tau}_\beta(h^\star_{m+n})\right| \le \left|\widetilde{\tau}_\beta(\hat{h}_{m+n}) - \widetilde{\tau}_\beta(\bar{h}_{m+n})\right| + \left|\widetilde{\tau}_\beta(\bar{h}_{m+n}) - \widetilde{\tau}_\beta(h^\star_{m+n})\right| \quad (58)$$

For the first term, Assumption E.2 implies local Lipschitz continuity of the soft crest-time functional with respect to trajectory perturbations, so there exists $L_\tau > 0$ such that

$$\left|\widetilde{\tau}_\beta(\hat{h}_{m+n}) - \widetilde{\tau}_\beta(\bar{h}_{m+n})\right| \le L_\tau \, \|\hat{h}_{m+n} - \bar{h}_{m+n}\|_\infty = L_\tau \, \|\mathcal{E}_{m,n}(x_t)\|_\infty \quad (59)$$

In addition, by Lemma E.1,

$$\left|\widetilde{\tau}_\beta(\hat{h}_{m+n}) - \widetilde{\tau}_\beta(\bar{h}_{m+n})\right| \le P(P+1) \, \|\Phi(\hat{h}_{m+n}) - \Phi(\bar{h}_{m+n})\|_\infty = P(P+1) \, \|\mathcal{E}^\Phi_{m,n}(x_t)\|_\infty \quad (60)$$

For the second term in (58), Assumption E.4 yields

$$\|\bar{h}_{m+n} - h^\star_{m+n}\|_\infty \le L_{\mathrm{hyd}} \, \|\widehat{\mathcal{R}}_{\mathrm{hydro}}\|_2. \Rightarrow \left|\widetilde{\tau}_\beta(\bar{h}_{m+n}) - \widetilde{\tau}_\beta(h^\star_{m+n})\right| \le L_\tau L_{\mathrm{hyd}} \, \|\widehat{\mathcal{R}}_{\mathrm{hydro}}\|_2. \quad (61)$$

Combining (58), (59), and (60) yields (57), with

$$C_1 = L_\tau, \qquad C_2 = P(P+1), \qquad C_3 = L_\tau L_{\mathrm{hyd}}. \qquad \square$$

Theorem E.5 formalizes the central role of semigroup consistency in long-horizon forecasting. The first term in (57) controls ordinary trajectory-level compositional inconsistency, the second controls event-phase inconsistency, and the third quantifies the extent to which the composed forecast departs from the latent hydrodynamic manifold. In particular, the theorem shows that horizon-dependent phase drift is not merely a by-product of pointwise regression error, but a direct consequence of violating temporal composition and phase transport consistency.

Implementation. In all experiments we set $P = 32$ and use a prefix split $m = 16$. The full rollout $\mathcal{T}_P(x_t)$ is computed first, and its first m steps are detached. A second short rollout $\mathcal{T}_m(x_t)$ is then computed using the same historical input and the first m rainfall-forcing steps.

F Decoupled Inverse Rainfall Ambiguity Modeling

F.1 Purpose and Decoupling Principle

The third contribution is designed to quantify *forecast-time forcing ambiguity*. Let

$$\widehat{\mathbf{h}}_t^{(0)} = F_\Theta\Big(\mathbf{X}_{t-L+1:t}, \mathbf{r}^{\mathrm{pred}}_{t+1:t+P}\Big) \in \mathbb{R}^P, \tag{62}$$

denote the predictor, where $\mathbf{X}_{t-L+1:t}$ is the historical input window, $\mathbf{r}^{\mathrm{pred}}_{t+1:t+P}$ is the rainfall sequence already used by the forecasting model, L is the lookback length, and P is the forecast horizon.

Contribution 3 introduces a separate diagnostic module that receives an *external forecast-side rainfall proxy*

$$\mathbf{r}^{\mathrm{fc}}_{t+1:t+P} = \big(r^{\mathrm{fc}}_{t+1}, \dots, r^{\mathrm{fc}}_{t+P}\big), \tag{63}$$

and estimates a posterior distribution over its unknown discrepancy relative to the realized rainfall. Crucially, this module is *not* fed back into (78). Therefore the hydrograph prediction remains fixed.

Proposition F.1 (Strict decoupling of the hydrograph forecast)**.** *Let $\widehat{\mathbf{h}}_t^{(0)} = F_\Theta(\mathbf{X}_{t-L+1:t}, \mathbf{r}^{\mathrm{pred}}_{t+1:t+P})$ be the predictor, and let q_ψ denote the diagnostic posterior introduced below. If q_ψ is not used as an input to F_Θ, then*

$$\widehat{\mathbf{h}}_t^{\mathrm{C3}} := \widehat{\mathbf{h}}_t^{(0)}, \qquad \partial_\psi \widehat{\mathbf{h}}_t^{\mathrm{C3}} = 0 \tag{64}$$

Hence the module cannot change the base hydrograph forecast.

Proof. The output $\widehat{\mathbf{h}}_t^{\mathrm{C3}}$ is defined to be exactly $\widehat{\mathbf{h}}_t^{(0)}$, and $\widehat{\mathbf{h}}_t^{(0)}$ depends only on the frozen parameters Θ and the predictor-side rainfall input $\mathbf{r}^{\mathrm{pred}}_{t+1:t+P}$. Since ψ does not appear in the computational graph of F_Θ, the derivative with respect to ψ is identically zero. □

Thus, an *inverse-identification and uncertainty-diagnosis layer*, not a forecast-correction layer.

F.2 Inverse Problem Formulation

Let the realized future rainfall and the unknown forecast discrepancy field be

$$\mathbf{r}^{\mathrm{obs}}_{t+1:t+P} = \big(r^{\mathrm{obs}}_{t+1}, \dots, r^{\mathrm{obs}}_{t+P}\big), \delta\mathbf{r}_t := \mathbf{r}^{\mathrm{obs}}_{t+1:t+P} - \mathbf{r}^{\mathrm{fc}}_{t+1:t+P} \in \mathbb{R}^P \tag{65}$$

The goal is to infer a posterior distribution over $\delta\mathbf{r}_t$ conditioned on the base forecast state and its physics diagnostics. We collect all inputs available to the inverse model into

$$\mathcal{Z}_t := \Big(\mathbf{X}_{t-L+1:t}, \mathbf{r}^{\mathrm{fc}}_{t+1:t+P}, \mathbf{r}^{\mathrm{pred}}_{t+1:t+P}, \widehat{\mathbf{h}}_t^{(0)}, \mathbf{s}_t, |\mathbf{R}_t|, |\mathbf{B}_t|\Big), \tag{66}$$

We then define a mean-field heteroscedastic posterior

$$q_\psi(\delta\mathbf{r}_t \mid \mathcal{Z}_t) = \prod_{k=1}^{P} \mathcal{N}\big(\delta r_{t,k}; \mu_{\psi,k}(\mathcal{Z}_t), \sigma^2_{\psi,k}(\mathcal{Z}_t)\big) \tag{67}$$

The posterior mean induces the corrected rainfall mean and pointwise credibility bands

$$\widehat{\mathbf{r}}_t^{\mathrm{corr}} = \mathbf{r}^{\mathrm{fc}}_{t+1:t+P} + \boldsymbol{\mu}_\psi(\mathcal{Z}_t),\ \widehat{r}^{\mathrm{corr},\,q}_{t,k} = r^{\mathrm{fc}}_{t+k} + \mu_{\psi,k} + z_q\sigma_{\psi,k}, \qquad q \in (0,1), \tag{68}$$

where z_q is the standard Gaussian quantile. In our plots, we use $z_{0.05} \approx -1.645$ and $z_{0.95} \approx 1.645$.

F.3 Connection to the Pullback Physics Residual

The base predictor uses the pullback latent operator with gauge residual

$$\mathcal{L}_\Theta[v] = \partial_\tau v - \nu\big(\partial_x\phi\big)^2\partial_{yy}v + B_\Theta\,\partial_y v,\ B_\Theta = \partial_\tau\phi + a_\Theta\partial_x\phi - \nu\partial_{xx}\phi \tag{69}$$

The strong pullback PDE residual used by the base model is

$$\mathcal{R}_\Theta = \mathcal{L}_\Theta[v_\Theta] - f_\Theta^{\text{pred}}, \to f_\Theta^{\text{pred}}(\tau, x) = r^{\text{pred}}(\tau)\, e^{-\kappa x}. \tag{70}$$

It does *not* replace this forcing inside the predictor. Instead, it uses $\mathcal{R}_\Theta$ as evidence about unresolved forcing mismatch. To make this precise, suppose the effective source acting on the frozen latent state were

$$f^\star(\tau, x) = f_\Theta^{\text{pred}}(\tau, x) + \delta f(\tau, x) \tag{71}$$

Then, holding v_Θ and ϕ_Θ fixed, the residual under the modified forcing becomes

$$\mathcal{R}^\star = \mathcal{L}_\Theta[v_\Theta] - f^\star = \mathcal{R}_\Theta - \delta f \tag{72}$$

If the unresolved error is primarily due to source misspecification, then an ideal correction should approximately satisfy $\mathcal{R}^\star \approx 0$, hence

$$\delta f \approx \mathcal{R}_\Theta \tag{73}$$

Now we postulate that rainfall discrepancy propagates to the pullback source through the same exponential projection family as in the base model:

$$\delta f_k = \Pi_\kappa(\delta r_k) := \delta r_k\, e^{-\kappa x_k}, \qquad k = 1, \dots, P, \tag{74}$$

where x_k is the latent coordinate at horizon step k and $\kappa > 0$ is the decay parameter inherited from the predictor. Applying (74) to the posterior mean yields

$$\Pi_\kappa(\boldsymbol{\mu}_\psi)_k = \mu_{\psi,k} e^{-\kappa x_k} \Rightarrow \mathcal{L}_{\text{phys}}^{\text{C3}} = \frac{1}{P} \sum_{k=1}^{P} \left(\mu_{\psi,k} e^{-\kappa x_k} - \mathcal{R}_{\Theta,k}\right)^2 \tag{75}$$

Interpretation. Equation (75) does *not* solve a new forward PDE. Instead, it treats the frozen pullback residual as an inverse signal for the missing source component attributable to uncertain rainfall forcing.

F.4 Heteroscedastic Posterior Regression

When realized rainfall is available offline, we can supervise the inferred discrepancy using

$$\delta \mathbf{r}_t^{\text{obs}} = \mathbf{r}_{t+1:t+P}^{\text{obs}} - \mathbf{r}_{t+1:t+P}^{\text{fc}} \tag{76}$$

Under the diagonal Gaussian posterior (67), the negative log-likelihood is

$$\mathcal{L}_{\text{nll}}^{\text{C3}} = \frac{1}{2P} \sum_{k=1}^{P} \left[\log \sigma_{\psi,k}^2 + \frac{\left(\delta r_{t,k}^{\text{obs}} - \mu_{\psi,k}\right)^2}{\sigma_{\psi,k}^2 + \varepsilon} \right], \tag{77}$$

with $\varepsilon > 0$ for numerical stability.

F.5 Sensitivity-Weighted Ambiguity Functional

The ambiguity score should not measure rainfall uncertainty in isolation. A large posterior variance is operationally important only when the frozen hydrograph predictor is sensitive to rainfall perturbations at the corresponding lead time. Let the frozen base forecast be

$$\widehat{\mathbf{h}}_t^{(0)} = F_\Theta\left(\mathbf{X}_{t-L+1:t}, \mathbf{r}_{t+1:t+P}^{\text{pred}}\right), \tag{78}$$

and define the rainfall-to-stage sensitivity operator in non-fractional Jacobian notation as

$$J_t := \partial_{\mathbf{r}_{t+1:t+P}^{\text{pred}}} \widehat{\mathbf{h}}_t^{(0)} \in \mathbb{R}^{P \times P}. \tag{79}$$

For a small rainfall perturbation $\delta \mathbf{r}_t$, the first-order hydrograph response is

$$\delta \widehat{\mathbf{h}}_t^{(0)} \approx J_t\, \delta \mathbf{r}_t. \tag{80}$$

Since the inverse discrepancy posterior in (15) is diagonal, its covariance is

$$\Sigma_{\psi,t} = \text{diag}\left(\sigma_{\psi,1}^2, \dots, \sigma_{\psi,P}^2\right). \tag{81}$$

Therefore, first-order covariance propagation gives

$$\begin{aligned}\mathrm{Cov}\left[\delta\widehat{\mathbf{h}}_t^{(0)} \,\middle|\, \mathcal{Z}_t\right] &\approx \mathrm{Cov}\left[J_t\delta\mathbf{r}_t \mid \mathcal{Z}_t\right] \\ &= J_t\,\mathrm{Cov}\left[\delta\mathbf{r}_t \mid \mathcal{Z}_t\right] J_t^\top \\ &= J_t\Sigma_{\psi,t}J_t^\top.\end{aligned} \tag{82}$$

Taking the trace yields a scalar ambiguity score:

$$\begin{aligned}\mathcal{A}_t &:= P^{-1}\,\mathrm{tr}\left(J_t\Sigma_{\psi,t}J_t^\top\right) \\ &= P^{-1}\sum_{k=1}^{P}\sigma_{\psi,k}^2\left\|J_{t,:,k}\right\|_2^2.\end{aligned} \tag{83}$$

Thus, ambiguity is large only when two conditions hold simultaneously: the inferred rainfall discrepancy has high posterior variance, and the frozen hydrograph predictor is locally sensitive to rainfall at that lead time.

For implementation and visualization, we define the horizon-wise sensitivity and ambiguity fields as

$$\begin{aligned}s_{t,k}^2 &:= \left\|J_{t,:,k}\right\|_2^2 = \sum_{j=1}^{P}\left(\partial_{r_{t,k}^{\mathrm{pred}}}\widehat{h}_{t,j}^{(0)}\right)^2, \qquad k=1,\dots,P, \\ a_{t,k} &:= \sigma_{\psi,k}^2 s_{t,k}^2.\end{aligned} \tag{84}$$

The issue-level ambiguity is then the horizon average

$$\mathcal{A}_t = P^{-1}\sum_{k=1}^{P}a_{t,k}. \tag{85}$$

Finite-difference implementation. In the implementation, the rainfall-to-stage sensitivity field is computed by automatic differentiation through the frozen predictor. For each output horizon j, we evaluate

$$\partial_{\mathbf{r}_{t+1:t+P}^{\mathrm{pred}}}\widehat{h}_{t,j}^{(0)}$$

while keeping the predictor parameters Θ fixed. This yields the Jacobian rows $J_{t,j,:}$, from which the horizon-wise sensitivity energy and ambiguity field are computed as in (84). No gradient update is applied to the frozen hydrograph predictor during this diagnostic computation.

$$\partial_{r_{t,k}^{\mathrm{pred}}}\widehat{h}_{t,j}^{(0)} \approx (2\epsilon)^{-1}\left[F_{\Theta,j}\left(\mathbf{X}_{t-L+1:t},\mathbf{r}_{t+1:t+P}^{\mathrm{pred}}+\epsilon\mathbf{e}_k\right)-F_{\Theta,j}\left(\mathbf{X}_{t-L+1:t},\mathbf{r}_{t+1:t+P}^{\mathrm{pred}}-\epsilon\mathbf{e}_k\right)\right], \tag{86}$$

where $\epsilon>0$ is a small perturbation magnitude and $\mathbf{e}_k$ is the k-th canonical basis vector over the forecast horizon.

F.6 Error-Consistency Supervision for Ambiguity

The ambiguity field should concentrate on lead times where the frozen base predictor is empirically fragile. We therefore align the predicted ambiguity field with the realized hydrograph error field. Define

$$e_{t,k} = \left(\widehat{h}_{t,k}^{(0)} - h_{t+k}\right)^2, \qquad k=1,\dots,P. \tag{87}$$

Since $a_{t,k}$ and $e_{t,k}$ may have different physical scales, we compare them after applying the same nonnegative sequence normalization:

$$\mathcal{N}(\mathbf{z})_k = \left(z_k - \min_{1\le j\le P} z_j\right)\left[P^{-1}\sum_{j=1}^{P}\left(z_j - \min_{1\le i\le P} z_i\right)+\varepsilon\right]^{-1}. \tag{88}$$

The normalized ambiguity and error fields are

$$\bar{a}_{t,k} = \mathcal{N}(\mathbf{a}_t)_k, \qquad \bar{e}_{t,k} = \mathcal{N}(\mathbf{e}_t)_k. \tag{89}$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as sparse-gauge flood-stage forecasting and summarize the three main contributions. The claims are supported by the benchmark, ablation, uncertainty, and appendix results, and are limited to the evaluated UK catchment setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and future work in Conclusion and Appendix

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The technical appendix states the assumptions for the gauge pullback, semigroup/phase consistency, and ambiguity formulation, and provides numbered lemmas, propositions, theorems, and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the model architecture, loss terms, input features, forecast horizon, train/validation/test protocol, random seeds, baselines, ablations, and evaluation metrics. Additional implementation details and scripts are described in the appendix to enable reproduction of the main experimental claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will make the code public on our GitHub Repository after acceptance (Funding Compliance).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section and appendix report the data splits, lookback and prediction horizons, input features, optimizer, learning rate, loss weights, random seeds, and evaluation protocol. Baseline and ablation settings are described sufficiently to interpret the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main benchmark, ablation, and C3 uncertainty tables report mean values with seed-wise standard deviations over seeds. The paper states that the uncertainty values are standard deviations across independent random-seed runs.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the compute environment used for training and evaluation, including GPU/CPU setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The work uses environmental time-series data for flood forecasting, does not involve human subjects, and does not introduce high-risk generative or surveillance capabilities.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive impacts for flood-risk forecasting and decision support, as well as potential risks from miscalibrated forecasts, over-reliance on automated warnings, and deployment under distribution shift. The paper emphasizes that the method should be validated and calibrated before operational use.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk model such as a generative model etc.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing datasets, baseline methods, and software packages are credited in the paper or supplementary material, with licenses and terms of use reported where available. The paper respects the usage conditions of the hydrological and meteorological data sources.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released supplementary code and configuration files are documented with instructions for training, evaluation, ablation, and figure reproduction. The released assets are anonymized for submission and include limitations and usage notes.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, surveys, or experiments with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research, crowdsourcing, or personally identifiable participant data, so IRB or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core method, experiments, or scientific claims. The proposed model, training procedure, and uncertainty diagnostics do not rely on LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.