DISTRIBUTED HYDROLOGICAL MODELING IN THE FEATURE SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate forecasting of river discharge and floods is very challenging. River dynamics are affected by storage, meteorological forcing, and flow propagation at different spatial and temporal scales. Forecasting thus requires a framework that considers the upstream-to-downstream flow through river networks across grid cells and catchments. This modeling is known in hydrology as distributed modeling and routing. Existing deep learning approaches either ignore this topology, operate on lumped catchments, or route predicted physical quantities through a separate graph or physical routing model. We instead introduce feature-space routing: a topology-aware state-space operator embedded directly in the forecasting dynamics. At every forecast step, the operator gathers latent states from upstream grid cells and causally updates the downstream state according to the known river network. This preserves the physical connectivity of the river system while allowing the propagated state itself to be learned end-to-end and allows the model to predict river discharge considering both local dynamics and neighboring upstream contributions. To address uncertainty and provide probabilistic forecasts, we minimize the fair continuous ranked probability score (fCRPS) as a training objective. Our experiments on the European Flood Awareness System (EFAS) and observational data for river discharge forecasting demonstrate that encoding the physical structure of river networks explicitly in the feature space substantially improves the forecasting skill, particularly in an ungauged setting. Our approach achieves state-of-the-art results on both reanalysis and observational data and is able to forecast maps of river discharge at 1 arcminute and 6-hourly resolution up to 10 days lead time.

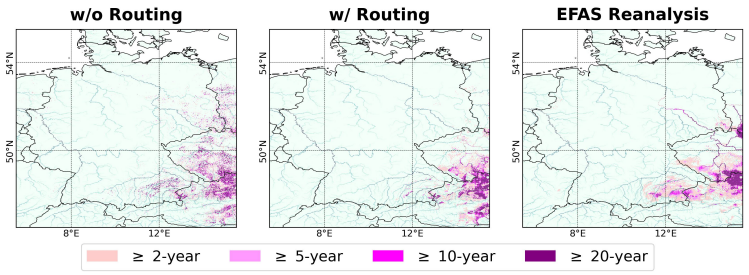


Figure 1: Example river discharge and flood forecasting over mid Europe in Sep. 2024. In the middle is the forecast produced by our model with routing at 4-day lead time, compared to the same model but without routing mechanism on the left and EFAS reanalysis data as a reference on the right side. The flood categories are shown below, e.g., high river discharge with 20-year return period. Without explicit distributed modeling, the routing-free model produces heterogeneous predictions in which neighboring pixels can have inconsistent discharge, leading to many false positives. Our routing module instead yields more spatially coherent predictions that better follow the river network.

1 INTRODUCTION

Accurate flood forecasting in rivers is essential for early warning and protecting citizens. In many operational systems, the forecast must infer the flow magnitude and propagation from uncertain meteorological forcing and limited hydrological initial conditions [Harrigan et al., 2023]. In a nutshell, river discharge forecasting combines two geometries: meteorological forcing and catchment

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

properties are naturally represented as spatial fields, while water propagates through a sparse directed river network. Existing approaches typically privilege one of these representations. Classical hydrological models address this by parametrized calibration and forcing physical laws on the modeling of runoff generation and river routing. However, physics-based models are expensive to run and require extensive calibration efforts. For instance, operating the European Flood Awareness System (EFAS) for real-time operations is computationally expensive and it is only provided for authorized users [Dottori et al., 2017]. Therefore, deep learning approaches have become more attractive for early warning systems of floods [Nearing et al., 2021]. While they achieve promising results, most existing models are lumped models which simplify the modeling and treat the whole catchment as a single entity [Nearing et al., 2024]. Yet, this neglects key drivers needed for accurate flood prediction. Only few works touched the distributed modeling and routing with deep learning at large-scale [Shams Eddin et al., 2025; Yang et al., 2025]. These approaches, however, either do not achieve the required spatio-temporal resolution required for early warnings or are limited in their capabilities in modeling the flow propagation within and through catchments. For instance, Yang et al. [2025] use a physics-based routing module but the approach is limited to 0.25°, not accounting for rapidly evolving dynamics and fine-scale characteristics of the catchments. The routing is also a post-processing step and not part of the training of the LSTM runoff model. While [Shams Eddin et al., 2025] is much more efficient and capable of operating on 0.05° and daily resolution, it does not explicitly represent the directional flow connectivity between upstream and downstream locations.

In this work, we thus propose to encode the routing explicitly in the model by applying a State Space Model (SSM) [Dao & Gu, 2024] directly on the graph structure of the river networks. Using a specialized algorithm to convert the grid points into 1D sequences, the SSM-based routing module propagates upstream-to-downstream information in the feature space while the direction of propagation is constrained by the known river topology. This novel hybrid modeling approach has the advantage that it remains very efficient while capturing flow propagation in a more realistic way than purely machine learning approaches. While the dense grid representation is retained, the learning of the latent representation and its dynamics is guided by the directional flow connectivity, which improves the forecast as shown in Fig. 1. Building upon the proposed routing module, we introduce the first deep learning approach that is capable of forecasting high-resolution probabilistic river discharge and floods at 1-arcminute and sub-daily resolution up to 10 days at continental-scale and in an operationally compatible setting. To address uncertainty, we optimize the model with the fair Continuous Ranked Probability Score (fCRPS) as a training objective. Our contributions can be summarized as follows:

- We introduce a novel feature-space routing, a directed state-space operator that injects known network topology into dense auto-regressive forecasting by propagating learned latent states from upstream to downstream grid points.
- We show through controlled ablations against no routing and standard GNN aggregation that this structural inductive bias substantially improves discharge prediction, flood detection, and particularly spatial generalization to ungauged locations.
- We introduce the first operational deep learning approach that is optimized with a score-based objective function and capable of providing probabilistic maps of river discharge forecast at 1 arcminute and 6-hourly resolution, and up to 10 days lead time over Europe.

2 RELATED WORKS

River Discharge and flood forecasting. We are interested in floods that occur in rivers. Generally, such floods are termed as *Fluvial* or *riverine* flood. They occur when the river flow increases and the water level rises high enough to overflow onto the surrounding land [Bomers & Hulscher, 2023]. Such flood events can be detected from river discharge itself, whenever the flow exceeds a defined threshold. Throughout this work, flood severity refers to exceedance of discharge return-period thresholds instead of inundation depth or extent [Jiang et al., 2024]. Deep learning has become common for forecasting river discharge in hydrology [Slater et al., 2025]. Most existing approaches rely on a lumped model trained jointly across many catchments, for instance an LSTM [Kratzert et al., 2018; 2019; 2024; Ruparell et al., 2025]. A key limitation here is that a lumped model represents an entire catchment as one undifferentiated unit, using averaged meteorological forcing and landscape features over the catchment.Yet catchments are internally heterogeneous, both in their

hydrological processes and in the structure of their river networks. This induces spatio-temporal causal dependencies within and across catchments [Merz et al., 2022; Stein et al., 2025]. Therefore, some studies start adapting Graph Neural Networks (GNNs) to model the topology along river networks [Sun et al., 2021]. But, they are still trained on small scale river networks and limited in term of generalizability [Kirschstein & Sun, 2024]. Most recently, RiverMamba proposed a new methodology based on State Space Model (SSM) to model large spatio-temporal contexts and forecast river discharge at grid-scale [Shams Eddin et al., 2025]. However, RiverMamba has limited forecast horizon due to the fixed number of forecasting heads i.e., it has no auto-regressive mechanism and it does not encode the upstream-to-downstream flow propagation. This process is known in hydrology as distributed modeling and river routing. In this work, we propose a distributed modeling mechanism that is encoded explicitly inside the forecasting engine. In addition, we build our model on SSM and based the forecast on rollout concept to achieve long-term forecast at sub-daily resolution.

Distributed hydrological modeling and routing. Classical hydrological models represent routing and spatial heterogeneity using semi- and fully-distributed formulations [Freeze & Harlan, 1969]. Semi-distributed models split the catchment into sub-catchments and perform routing along the river network, whereas fully-distributed models such as LISFLOOD [Knijff et al., 2010] discretize the catchment into grid cells and solve physics-based governing equations for routing at each grid point. Dedicated distributed routing models such as CaMa-Flood [Yamazaki et al., 2011] similarly propagate runoff through prescribed river networks. Such models, however, remain expensive to run at high resolution and require extensive calibration [Hirpa et al., 2018]. There have been attempts in hydrology to design deep learning models capable of fully spatially distributed modeling. Typical approaches estimate runoff, then route it downstream using either a physics-based routing module [Yang et al., 2025; Kraft et al., 2025] or another neural network [Vischer et al., 2025; Mosaffa et al., 2026]. Hascoet et al. [2026] instead reformulate linear time-invariant routing as a block-sparse convolution, keeping it physically constrained but differentiable and trainable end-to-end with a runoff model. Scholz et al. [2025] propose an LSTM coupled with a CNN to learn runoff embeddings, which a GNN then propagates over the river network topology to route flows between stations. Li et al. [2026] apply the same approach but use CNNs to represent the spatial runoff. Other works use hybrid frameworks to learn the parametrization of physical equations including river routing [Wang et al., 2024a; Bindas et al., 2024; Song et al., 2025]. In our work, we propose a novel method in hydrology based on SSMs to perform fully spatially distributed modeling. Rather than coupling a grid-based runoff model with a physics-based routing or a separated neural network operating on a coarser river graph, we use a single SSM architecture to both process the spatio-temporal grid and route high-dimensional features directly at the native grid resolution within one forecasting engine.

Probabilistic river discharge forecasting. Reliable flood prediction depends on uncertainty estimation [Baste et al., 2026]. This includes, aleatoric and epistemic uncertainties because measurements and initial conditions are inherently uncertain. Physics-based systems such as GloFAS [Harrigan et al., 2023] and EFAS perturb their inputs using ensemble weather forecast to account for these uncertainties. In deep learning for hydrology, Klotz et al. [2022] introduced mixture density networks. This approach is also adopted by Google's flood forecasting systems [Nearing et al., 2024]. Another approach is to model the distribution of discharge with denoising diffusion probabilistic models [Ou et al., 2025; Wang et al., 2025] or variational auto-encoder [Jahangir & Quilty, 2024]. More recently, Kraft et al. [2026] and Ruparell et al. [2026b] proposed a score-based generative approaches and paired it with an LSTM for rainfall-runoff prediction. Ruparell et al. [2026a] demonstrate that moving from lumped to distributed probabilistic hydrology demands explicit modeling of the spatial joint structure underlying runoff uncertainty. In our work, inspired by recent advances in weather forecasting research [Alet et al., 2025; Lang et al., 2026], we propose the fair Continuous Ranked Probability Score (fCRPS) as a training objective To the best of our knowledge, ours is the first deep learning approach optimized with fCRPS in a fully spatially distributed modeling framework.

3 METHOD

In this work, we present a novel approach for distributed hydrological modeling along with the first deep learning model that is able of generating probabilistic river discharge and flood forecasting up to 10 days at 1-arcminute and 6-hourly spatio-temporal resolution (Fig. 1). This is a very challenging task since it requires an approach that should not only learn the catchment dynamics locally, but also considers the upstream-to-downstream flow propagation through river networks to resolve heterogeneity at fine scale. We thus propose a novel feature-space routing. It consists of a directed

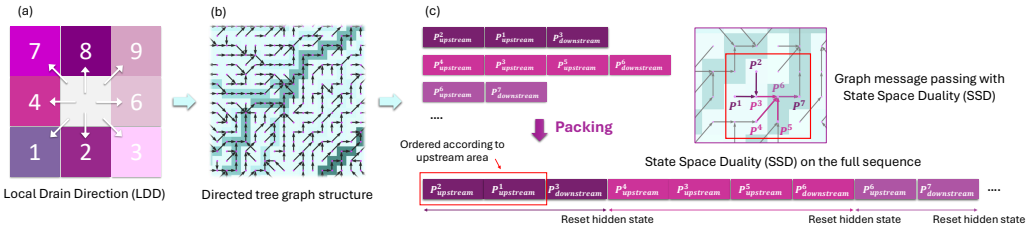


Figure 2: Overview of the routing module. (a) Shows the Local Drain Direction (LDD) which is used to generate the graph of river network (b). Each downstream point forms a sequence with its upstream neighbors and routing is conducted via SSD on concatenated point sequences at once (c).

state-space operator that injects known network topology into a network by propagating learned latent states from upstream to downstream nodes. We will first briefly describe the Structured State Space Duality (SSD) in Sec. 3.1 and the routing in Sec. 3.2. In Sec. 3.3, we describe the entire network.

3.1 Structured State Space Duality (SSD)

Mamba2 or SSD [Dao & Gu, 2024] is a state space model that transforms a 1D sequence of states $x(t)$ into another representation $y(t)$ through an implicit hidden latent state $h(t)$:

$$h_t = \bar{\mathbf{A}} h_{t-1} + \bar{\mathbf{B}} x_t \,, \quad y_t = \mathbf{C} h_t + \mathbf{D} x_t \,, \tag{1}$$

$$\bar{\mathbf{A}} = e^{(\Delta \mathbf{A})} \,, \quad \bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1} (e^{(\Delta \mathbf{A})} - \mathbf{I}) \Delta \mathbf{B} \,, \tag{2}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$ are parametrized matrices discretized with a timescale parameter Δ. $\mathbf{B}(x)$, $\mathbf{C}(x)$, and $\Delta(x)$ depend on the input state x which makes the system selective time-variant, i.e., it is able to control whether to keep or ignore some information along the sequence. SSD is considered as an efficient alternative to Transformers [Vaswani et al., 2017] for processing long sequences of points since it has a linear complexity w.r.t. the input tokens. In SSD, $\mathbf{A}$ is restricted into a scalar-times identity, $\mathbf{A} = a\,\mathbf{I}$. This restriction enables matrix multiplication and the recurrence in Eq. 1 to be state-space dual, equivalently to causal linear attention.

3.2 Routing

In hydrological models, channel routing propagates discharge through the river network, with locally generated surface and subsurface runoff entering the channel as lateral inflow. Each grid point receives upstream contributions, combines them with local inflow, and propagates the resulting flow downstream. We adopt the same structure in our model, but route learned high dimensional *features* x_p rather than physical quantities. For each point p, a routing module aggregates the features of p's upstream neighbors and processes them jointly with p's own features via a selective State Space Duality (SSD). An overview is shown in Fig. 2. The routing structure comes from the Local Drain Direction (LDD), a static field derived from terrain and hydrographic information that assigns each point exactly one downstream direction i.e., one of 9 values (Fig. 2 (a)). Inverting this map gives, for every point, up to 8 upstream neighbors, one per compass direction, which together define a directed drainage graph representing the river network (Fig. 2 (b)). We propose to conduct routing over this graph as message passing implemented via SSD. For each point p, we form a short sequence of p's upstream neighbors followed by p itself. Formally, for point p with $j(p)$ upstream neighbors:

$$s(p) = \big(n_1(p), n_2(p), \ldots, n_{j(p)}(p),\, p\big) \,, \tag{3}$$

$$d_i(p) = \begin{cases} \mathrm{LDD}\big(n_i(p)\big) \,, & i \le j(p) \\ 5 \,, & i = j(p)+1 \,, \end{cases} \tag{4}$$

$$u_i(p) = x_{s(p)_i} + PE_{d_i(p)} \,, \qquad i = 1, \ldots, j(p)+1 \,, \tag{5}$$

$$x_p \leftarrow x_p + \Big[\mathrm{SSD}_\theta\big(u_1(p), \ldots, u_{j(p)+1}(p)\big)\Big]_{j(p)+1} \,, \tag{6}$$

where $n_i(p)$ is the i-th upstream neighbor of p in ascending order of upstream area ($j(p) \le 8$), $s(p)_i$ is the i-th entry of $s(p)$, $x_{s(p)_i}$ is the feature vector of the i-th entry, and $u_i(p)$ is the input token of the i-th entry which will be given to the SSD. $d_i(p)$ is the direction code of the i-th token. For an

upstream neighbor it is the neighbor's own LDD value, i.e. the direction in which it drains into p (one of the 8 codes $\{1, 2, 3, 4, 6, 7, 8, 9\}$), while p itself receives code 5, which marks it as the sequence's downstream point rather than a neighbor. $PE_d \in \mathbb{R}^K$ is a learned direction encoding, added to the raw feature $x_{s(p)_i}$ to form the SSD input token $u_i(p)$. Only SSD$_\theta$'s output at the final position, corresponding to p itself, is added residually to x_p. Since the sequence length $j(p)+1$ varies between points, the sequences of all points are packed into one long sequence rather than run one at a time as illustrated in Fig. 2 (c). To prevent state leakage across segments, every token is tagged with the index of its own segment, $\text{seg}(u_i(p)) = p$, and the hidden state is reset whenever segment changes:

$$h_i = \mathbb{1}\big[\text{seg}(u_i) = \text{seg}(u_{i-1})\big]\, \bar{\mathbf{A}}\, h_{i-1} + \bar{\mathbf{B}}\, u_i\,, \tag{7}$$

where the indicator evaluates to 0 at the first token of every point's sequence and to 1 elsewhere. This makes SSD$_\theta$ applied to the packed sequence exactly equivalent to running it independently on each point's sequence, while naturally handling the variable sequence lengths. The routing update is local and can be evaluated for all grid points in parallel, repeated application at each auto-regressive forecast step allows routed information to propagate downstream over time. A point with no upstream neighbors simply forms a sequence of length one. Because Mamba's scan is causal, a point is only influenced by points earlier in the sequence, so we place p at the end of the sequence which give it a receptive field over all upstream neighbors. Moreover, since the SSD's hidden state decays with distance, neighbors placed closer to p *in the sequence* have a stronger, less-decayed effect on p's aggregated representation than neighbors placed further away. We use upstream drainage area as a hydrologically informed ordering proxy for a neighbor's importance to p, and order neighbors by ascending area, so that the largest upstream area point sits immediately before p and receives the strongest influence. Unlike permutation-invariant neighbour pooling, the SSD operator maintains a learned recurrent state over upstream contributions. Ordering the sequence by upstream area permits to learn nonlinear, context-dependent feature aggregation rather than reducing upstream information through a fixed sum or mean. Note that the LDD we used here is the same as in the LISFLOOD hydrological model, where each point has one downstream direction. Our formulation is nonetheless general and applies unchanged if points were instead assigned multiple downstream directions or a dynamic LDD. We use this proposed routing operator in the forecasting engine of our model to propagate spatio-temporal routing latent states auto-regressively (Sec. 3.3).

3.3 Model architecture

An overview of our entire model, which includes the routing (described in Section 3.2), is shown in Fig. 3. It integrates as initial conditions different sources at daily temporal resolution. We start the forecast at time t at 00:00 UTC and use Analysis data from the ECMWF Integrated Forecast System (IFS) $\{\mathbf{X}^{t-4i}_{ANA}\}^{T-1}_{i=0}$, the IMERG data $\{\mathbf{X}^{t-4i}_{IMERG}\}^{T-1}_{i=0}$ as a satellite-based daily precipitation estimate [Huffman et al., 2020], CPC data as a gauge-based daily precipitation $\{\mathbf{X}^{t-4i}_{CPC}\}^{T}_{i=1}$ [Xie et al., 2010], and the EMO gridded ground weather stations $\{\mathbf{X}^{t-4i}_{EMO}\}^{T+1}_{i=2}$ [Thiemig et al., 2022], where $T = 4$ is the number of daily samples and t is at 6-hourly temporal resolution. Note that each data has a different latency according to their near real-time acquisition.

Additionally, we include AlphaEarth [Brown et al., 2025] embeddings $\mathbf{X}_{AlphaE}$ as static features. Furthermore, since forecasting river discharge is highly dependent on the meteorological forcing leading the forecast, we use medium-range and high-resolution weather forecast from IFS up to 10 days ($t = 40$) and denoted as $\{\mathbf{X}^{i}_{HRES}\}^{t+L}_{i=t+1}$, where $L = 40$ denotes the lead time of the forecast. More details about the input data are described in the suppl. material. The model encodes these initial conditions into $\mathbf{X}^{t}_{features}$ which is then given to the forecasting engine f that generates the next step of probabilistic river discharge forecast at 6-hourly interval conditioned on the weather forecast $\mathbf{X}_{HRES}$, static attributes $\mathbf{X}_{AlphaE}$, and noise vector $\epsilon_z \sim \mathbb{N}(0, \sigma_z)$ for probabilistic forecast. The forecasting engine rolls out up to 40 steps (10 days) in the future including nowcasting ($t = 0$). The auto-regressive forecast for 1 step can be formulated as $\mathbf{X}^{t+1}_{dis06} = f\left(\mathbf{X}^{t}_{feature}; \mathbf{X}^{t+1}_{HRES}, \mathbf{X}_{AlphaE}, \epsilon_z\right)$, where $\mathbf{X}^{t+1}_{dis06}$ is the averaged river discharge over 6 hours between t and $t + 1$. Long-term forecast is generated with rollout. In what follows, we describe the components of the model in details.

Input embedding. First, each grid point p of an input modality is projected into a higher feature dimension with a linear layer and followed with a hyperbolic tangent activation (Tanh). After

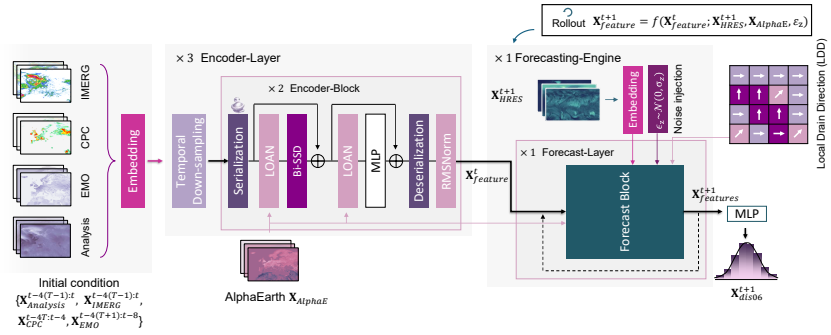


Figure 3: An overview of the proposed model for river discharge forecasting. The model initializes the forecast at time t and forecasts high-resolution river discharge maps $\mathbf{X}^{t:t+L}_{dis06}$ from the initial conditions ($\mathbf{X}^{t-4(T-1):t}_{ANA}$, $\mathbf{X}^{t-4(T-1):t}_{IMERG}$, $\mathbf{X}^{t-4T:t-4}_{CPC}$, $\mathbf{X}^{t-4(T+1):t-8}_{EMO}$), AlphaEarth embedding ($\mathbf{X}_{AlphaE}$), and weather forecasts ($\mathbf{X}^{t+1:t+L}_{HRES}$), conditioned on a noise vector ϵ_z to generate probabilistic forecast.

projection all modalities are combined and normalized:

$$\mathbf{X}^{t}_{embedding}(p) = \text{RMSNorm}\Big(\text{Tanh}\Big(\text{Concat}\big(\text{Linear}(\mathbf{X}^{t}_{ANA}(p)), \text{Linear}(\mathbf{X}^{t}_{IMERG}(p)), \text{Linear}(\mathbf{X}^{t-4}_{CPC}(p)), \text{Linear}(\mathbf{X}^{t-8}_{EMO}(p))\big)\Big)\Big) + \mathbf{X}_{pos}\,, \quad (8)$$

where RMSNorm is the root mean square layer normalization, Linear is a projection layer, $\mathbf{X}_{pos} \in \mathbb{R}^{B\times T\times 1\times K}$ is a learnable positional embedding, $\mathbf{X}^{t}_{embedding}(p) \in \mathbb{R}^{B\times T\times P\times K}$ is the input embedding with $K = 192$, B is the batch size, and P is the number of points (tokens) in the domain. Analysis data have the dimensionality $\mathbf{X}_{ANA} \in \mathbb{R}^{B\times T\times P\times V_a}$ with $V_a = 14$, projected to 96, IMERG and CPC both have $\mathbf{X}_{IMERG}, \mathbf{X}_{CPC} \in \mathbb{R}^{B\times T\times P}$, projected to 16 each, and $\mathbf{X}_{EMO} \in \mathbb{R}^{B\times T\times P\times V_e}$ is projected to a 64d space.

Encoder. The encoder is adapted from RiverMamba [Shams Eddin et al., 2025]. It is used to encode the initial condition into a representation $\mathbf{X}^{t}_{features} \in \mathbb{R}^{B\times 1\times P\times K}$ via spatio-temporal modeling inside Mamba modules. The encoder consists of 3 encoder layers and each one of them consists of 2 encoder blocks. The first layer models the full temporal resolution of $T = 4$. Afterward, T is down-sampled to $T/2$ using a linear layer at the beginning of each subsequent layer. Unlike bidirectional Mamba blocks in RiverMamba, where the forward and backward passes use two separately-parametrized SSMs, we use a shared SSM for bi-directionality. Each block incorporates further AlphaEarth embedding as static features via Location-aware Adaptive Normalization (LOAN) [Shams Eddin et al., 2023]. The output $\mathbf{X}^{t}_{features}$ is then given to the forecasting engine.

Forecasting Engine. The forecasting engine takes $\mathbf{X}^{t}_{features}$ and predicts the next $\mathbf{X}^{t+1}_{features}$ autoregressively and conditioned on the weather forecast $\mathbf{X}^{t+1}_{HRES}$, static attributes $\mathbf{X}_{AlphaE}$, and noise vector ϵ_z, as shown in Fig. 3. The structure of the block is similar to the encoder block but has 3 more processes (Fig. 4). First, it incorporates the weather forecast $\mathbf{X}^{t+1}_{HRES}$ by projecting it to 64d, serializing it and concatenating it with $\mathbf{X}^{t}_{features}$. Second, it uses noise injection with ϵ_z to generate an ensemble of predictions. Finally, the routing module operates in the end of the forecasting layer to distribute the learned spatio-temporal features through the river network. The output $\mathbf{X}^{t+1}_{features}$ for each lead time l produced by the forecasting engine is then given to an MLP which projects $\mathbf{X}^{t+1}_{features}$ to 64d and then 1d to estimate the river discharge $\mathbf{X}^{t+1}_{dis06}$.

Noise injection and training. The initial condition and weather forecast have systematic biases and are inherently uncertain. To model uncertainty, we use a score-based function that trains the model only on the marginal distribution (i.e., optimizing on few ensemble members), however, still be able to approximate the joint distribution over the predicted discharge. For this we incorporated a learned perturbation via noise injection in the forecasting engine. The noise is sampled from a Gaussian distribution $\epsilon_z \sim \mathbb{N}(0, \sigma_z = 0.5)^{32}$ with 32 dimension and mapped via 2 linear layers to $\epsilon_{z1} \in \mathbb{R}^{B\times 1\times 1\times K}$ and $\epsilon_{z2} \in \mathbb{R}^{B\times 1\times 1\times K}$. These perturbation are then added to both LOAN layers in the forecasting block, e.g., for the first part of the block $\text{LOAN}(\mathbf{X}) + \epsilon_{z1}$. Note that we share

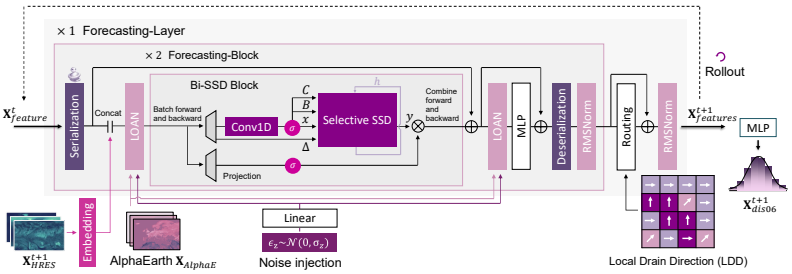


Figure 4: The structure of the forecast layer. It has a similar structure as the encoder layer, but it additionally incorporates meteorological forecasts (HRES) by concatenation, the noise ϵ_z, and the routing module. The bi-directional SSD block is shown in more details.

the noise vector across all grid points and lead time. As a score-based function, we use the fair Continuous Ranked Probability Score (CRPS) as an objective [Gneiting & Raftery, 2007]. For $\mathbf{X}^{1:M}_{dis06}$ samples, an unbiased estimate [Zamo & Naveau, 2018] can be written as:

$$\mathcal{L} = \frac{1}{BLP} \sum_{b,t,p} \left[\frac{1}{M} \sum_{m=1}^{M} \left| x^{m,b,t}_{\text{dis06}}(p) - \hat{x}^{b,t}_{\text{dis06}}(p) \right| - \frac{1}{2M(M-1)} \sum_{m,m'=1}^{M} \left| x^{m,b,t}_{\text{dis06}}(p) - x^{m',b,t}_{\text{dis06}}(p) \right| \right] \tag{9}$$

where M is the number of ensemble members, $x^{e,b,t}_{dis06}(p)$ is the prediction, and $\hat{x}^{b,t}_{dis06}(p)$ is the ground truth. During training, we set $M = 4$ ensemble members per GPU node. For experiments in a deterministic setting, we used the mean squared error as an objective function. All target river discharge values are transformed with the formula $\text{sign}(\hat{x})\log(1 + |\hat{x}|)$ as in RiverMamba.

For fine-tuning on observations, we keep the number of input P and optimize only on points with available observations. Note that our model still produces 6-hourly output for both reanalysis and observational data. During fine-tuning on observational data, we average the prediction over 4 six-hour steps to match the daily resolution before computing the loss. We assume that the learned sub-daily dynamics primarily originates from the pretraining stage. For inference, we can produce any number of ensemble flood forecasting and densely as in Fig. 1.

Warmup and nowcasting. For the first forecast step, $\mathbf{X}^{t}_{features}$ is produced by the encoder, whereas at every subsequent step, $\mathbf{X}^{t+1:t+L}_{features}$ is the forecasting engine's own previous prediction fed back as input. This auto-regressive feedback can induce a distribution shift. Inspired by physics-based systems, we use warmup to stabilize the forecast. Similar to spin-up in numerical weather and hydrological models, we run the forecasting engine for 4 steps before the actual forecast horizon. The first three warmup steps are driven by an earlier-initialized IFS-HRES forecast cycle at its short lead times, i.e., the same near-term forecast that would genuinely be available in an operational setting for that recent past. The fourth and final warmup step is driven by the latest available analysis. Only the internal state after this final warmup step is kept to initialize the rollout over $\mathbf{X}^{t+1:t+L}_{features}$.

4 Experimental Results

Data for pretraining on river discharge are obtained from EFAS reanalysis. The EFAS dataset is provided as a 6-hourly discharge over Europe at 1 arcminute resolution. We use the EFAS reanalysis as a target discharge for training and testing in Sec. 4.1. For fine-tuning on observations, we test the model on observational GRDC data in Sec. 4.2 and CAMELS in the suppl. material. We train up to 40 steps (10 days) on the years 1992-2021, validate on 2022-2023, and test on 2024-2025. We evaluate the performance on both EFAS reanalysis and observations, where diagnostic stations are available (1971 GRDC and 2206 CAMELs stations). Metrics are calculated by taking the median on individual grid points. We compare our model to the state-of-the-art diffusion-based DRUM [Ou et al., 2025], Encoder-Decoder LSTM (ED-LSTM) [Nearing et al., 2024], Mean-Embedding LSTM (ME-LSTM) Cohen et al. [2026] models, RiverMamba [Shams Eddin et al., 2025], and the operational EFAS forecasting system operated by ECMWF Dottori et al. [2017]. We used $M = 12$ ensemble members for all probabilistic models (RiverMamba is deterministic). More details about dataset, evaluation metrics, and baselines are in the suppl. material.

Table 1: Results on EFAS Reanalysis (CAMELS split).

		Validation (2022-2023)			Test (2024-2025)		
	Model	$R2^r$ (↑)	KGE^r (↑)	$CRPS^r$ (↓)	$R2^r$ (↑)	KGE^r (↑)	$CRPS^r$ (↓)
Probabilistic	DRUM (mean)	0.1496	0.1505	2.0872	0.0984	0.1352	2.3581
	ED-LSTM (mean)	0.1726	0.2431	1.9121	0.1347	0.2333	2.1971
	ME-LSTM (mean)	0.2093	0.3071	1.8395	0.1210	0.2640	2.2249
	Ours (mean)	**0.4515**	**0.5751**	**1.4720**	**0.3907**	**0.5474**	**1.7619**

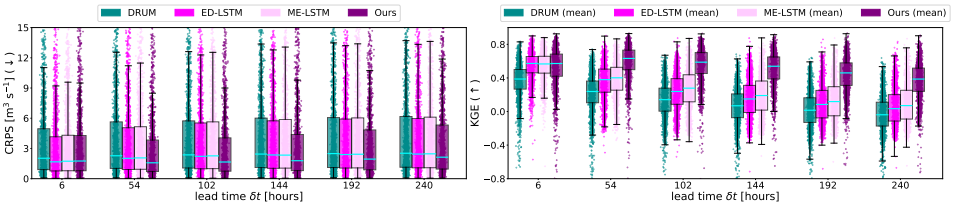


Figure 5: Results on EFAS reanalysis with lead time.

4.1 Experiments on EFAS river discharge reanalysis

Fig. 5 plots the median of CRPS and KGE metrics with 6 selected lead time. Boxes represent distribution quartiles, with evaluation grid points plotted along the y-axis. LSTM has a competitive performance but its prediction skill drops with long-term forecast. Our model achieves the best results with big margins. We attribute this to the spatio-temporal modeling and routing in the proposed model. Table 1 shows that our model achieves also better results quantitatively. Table 2 shows a comparison with RiverMamba in a deterministic setting over the Mid Europe domain (281K points) for 16 steps lead time. Our model consistently outperforms RiverMamba on F1 score. The RiverMamba baseline has no explicit routing mechanism inside the model. It relies on the spatio-temporal modeling of SSM without considering local drain directions, while our model explicitly models this within the routing module which improves flood detection. More results are in the suppl. material. In the following, we discuss the main ablation studies that were performed over a mid European domain for 16 steps lead time in a deterministic setting.

Table 2: F1 comparison to RiverMamba.

	Val	Test
RiverMamba	0.141±0.01	0.138±0.01
Ours	**0.334**±0.02	**0.314**±0.01

Table 3: Ungauged results on Reanalysis.

	$R2^r$	KGE^r	$F1^r$
✗ Routing	-0.501	-0.002	0.015
✓ Routing	**0.202**	**0.338**	**0.105**

Routing. In Table 4 (a), we evaluate the impact of adding explicit routing (Eq. 5-6) in the forecasting engine. In fact, routing introduces a hydrologically informed structural prior into the model which improves the forecasting skill substantially on both reanalysis and observational data. Without explicit routing, the model produces heterogeneous predictions in which neighboring pixels can have inconsistent discharge. The routing module instead yields more homogeneous, spatially clustered predictions (Fig. 1). In Table 3, we show that routing has even a bigger positive impact on spatial generalization for unseen catchments during training (ungauged setting). On average it improves R2 form -0.5 to +0.2. In Table 4 (b), we compare our proposed SSM message passing for routing with GNNs. GS uses the GraphSAGE model [Hamilton et al., 2017]. While GS-mean uses mean

Table 4: Ablation studies on the validation set over Mid Europe. Shown are KGE on reanalysis (KGE^r) and observational (KGE^o) CAMELS data along with $F1^r$@(1.5-500) for reanalysis over the full domain (281K points).

(a) Routing and Warmup

Routing	Warmup	KGE^r | KGE^o | $F1^r$(↑)
✗	✗	0.666 | 0.692 | 0.096
✓	✗	0.672 | 0.739 | 0.114
✗	✓	0.687 | 0.721 | 0.162
✓	✓	**0.740** | **0.747** | **0.339**

(b) Routing Module

Module	KGE^r | KGE^o | $F1^r$(↑)
GS-mean	0.680 | 0.721 | 0.310
GS-sum	0.702 | 0.702 | 0.285
GAT	0.714 | 0.633 | 0.268
SSM	**0.740** | **0.747** | **0.339**

(c) Upstream Aggregation

Type	KGE^r | KGE^o | $F1^r$(↑)
Summation	0.738 | 0.682 | 0.335
Mean	0.730 | 0.739 | 0.332
GraphMamba	0.722 | 0.720 | 0.276
Ours	**0.740** | **0.747** | **0.339**

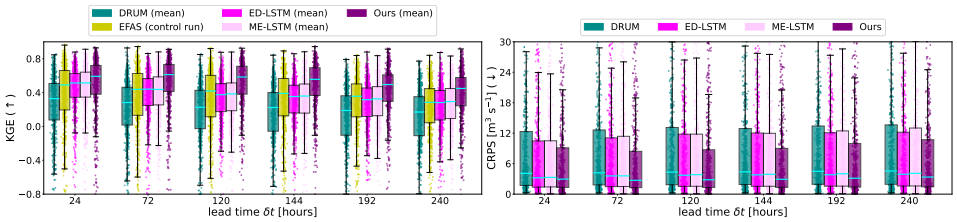


Figure 6: Results on GRDC observations with lead time.

aggregation over upstream neighbors, GS-sum uses summation over the neighbors. GAT [Veličković et al., 2017] is a graph transformer. As can be seen, our SSM message passing achieves the best result compared to the GNN variants. In Table 4 (c), we show the impact of different aggregation techniques for routing. In the first row, we sum up the upstream features and give them as input to the SSD as a sequence of length 2 for each point. In the second row, we do the same but take the mean instead of summation. GraphMamba [Wang et al., 2024b] sorts the tokens by their feature norms where the neighbor with highest norm is placed last in the sequence. Our technique works the best.

Warmup. In Table 4 (a), we also show the benefit of doing warmup on the forecast. As can be seen, warmup improves the results since it closes the distribution shift between $\mathbf{X}^{t}_{features}$ and $\mathbf{X}^{t+1:t+L}_{features}$ making the forecast more stable. More ablation studies are in the suppl. material.

4.2 Experiments on observational river discharge data

Table 5: Results on gauged GRDC observations.

		Validation (2022-2023)			Test (Jun 2024- Dec 2025)		
	Model	R2^o (↑)	KGE^o (↑)	CRPS^o (↓)	R2^o (↑)	KGE^o (↑)	CRPS^o (↓)
	EFAS (control run)	-	-	-	0.0410	0.3964	-
Probabilistic	DRUM (mean)	0.1750	0.2834	4.5375	-0.0089	0.2355	4.3474
	ED-LSTM (mean)	0.2841	0.3957	3.9024	0.1214	0.3788	3.6818
	ME-LSTM (mean)	0.3339	0.4771	3.4928	0.0913	0.3843	3.7962
	Ours (mean)	**0.6042**	**0.6689**	**2.4592**	**0.3693**	**0.5520**	**2.9708**

Performance on gauged GRDC river discharge observations is summarized in Table 5, which includes the EFAS model with explicit physics-based routing with forecast starting from mid 2024. As before, Fig. 6 evaluates these models across lead times. The diffusion-based DRUM model performs poorly, whereas EFAS performs on par with the ED-LSTM and ME-LSTM baselines despite its advantage of using discharge as initial condition. Nevertheless, EFAS has lower R2 values, likely because it models naturalized discharge without considering realistic human interventions such as dams and reservoirs. Our model consistently outperforms the other approaches for all metrics on observational data. More results are in the suppl. material.

5 Conclusions and limitations

We presented a novel deep learning approach for fully spatially distributed modeling and routing in hydrology. Our approach leverages State Space Duality (SSD) for upstream-to-downstream flow modeling. It packs local upstream-neighbor sequences into one long segmented sequence and then adapts the SSD to route high-dimensional features rather than physical quantities. For each point, a routing SSD module aggregates the learned upstream features from neighbors and processes them jointly with the downstream point's features. We showed that explicitly encoding the river network in the feature space substantially improves forecast skill. Built upon this, we introduced an operational model that forecasts gridded river discharge at very high spatio-temporal resolution for up to 10 days, achieving new state-of-the-art results. While this is a major advancement, our approach still has some limitations. The local drain directions are inferred from topology and we use a static directed graph in this work. The routing therefore does not represent backwater effects or flow reversals. Furthermore, satellite observations could be helpful to define a more accurate and dynamic graph. We also do not consider human interventions with natural flow such as reservoirs, abstractions, and diversions. Despite these, our model showed a very promising step towards accurate and high-resolution distributed river discharge forecasting.

AI USE STATEMENT

In this work, we used generative AI tools for very limited editing (e.g., grammar, spelling, word choice). We also used a generative AI tool for finding and fixing errors in the codebase. Additionally, we used generative AI tools to optimize and convert the dataset into Zarr format suitable for the timeseries baseline models such as LSTMs and to fine-tune the hyper-parameters of the LSTM models.

ETHICS STATEMENT

Not applicable.

REPRODUCIBILITY STATEMENT

The paper fully discloses all the information needed to reproduce the main experimental results. We described the method in details in Sec. 3 and implementation details in Sec.K. Instructions on how the data and code can be obtained and used are provided in the appendix in Sec. O and Sec. P.

REFERENCES

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.

S. Baste, S. Lerch, D. Klotz, and R. Loritz. The need for uncertainty: why probabilistic lstms are key to improving flood predictions and enabling learned warning rules. *EGUsphere*, 2026:1–33, 2026. doi: 10.5194/egusphere-2026-469. URL https://egusphere.copernicus.org/preprints/2026/egusphere-2026-469/.

Tadd Bindas, Wen-Ping Tsai, Jiangtao Liu, Farshid Rahmani, Dapeng Feng, Yuchen Bian, Kathryn Lawson, and Chaopeng Shen. Improving river routing using a differentiable muskingum-cunge model and physics-informed machine learning. *Water Resources Research*, 60(1):e2023WR035337, 2024. doi: https://doi.org/10.1029/2023WR035337. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR035337. e2023WR035337 2023WR035337.

Anouk Bomers and Suzanne J. M. H. Hulscher. Neural networks for fast fluvial flood predictions: Too good to be true? *River Research and Applications*, 39(8):1652–1658, 2023. doi: https://doi.org/10.1002/rra.4144. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rra.4144.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

D. Cohen, R. Amira, R. Aschner, Y. Carny, B. Feinstein, H. Fester, S. Fronman, M. Gauch, O. Gilon, R. Green, A. Hassidim, D. Klotz, F. Kratzert, D. Korenfeld, G. Loike, A. Markel, Y. Matias, R. Mayo, A. Metzger, B. Mosheyev, A. Niego, S. Rees, E. Reinstein, A. Sicherman, G. Shalev, O. Shefi, Y. Shildan, I. Zemach, O. Zlydenko, and G. Nearing. Extending medium-range global flood forecasts: The google global flood forecasting model version 2. *EGUsphere*, 2026:1–31, 2026. doi: 10.5194/egusphere-2026-2283. URL https://egusphere.copernicus.org/preprints/2026/egusphere-2026-2283/.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

F. Dottori, M. Kalas, P. Salamon, A. Bianchi, L. Alfieri, and L. Feyen. An operational procedure for rapid flood risk assessment in europe. *Natural Hazards and Earth System Sciences*, 17(7): 1111–1126, 2017. doi: 10.5194/nhess-17-1111-2017. URL https://nhess.copernicus.org/articles/17/1111/2017/.

R.Allan Freeze and R.L. Harlan. Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of Hydrology*, 9(3):237–258, November 1969. ISSN 0022-1694. doi: 10.1016/0022-1694(69)90020-1. URL http://dx.doi.org/10.1016/0022-1694(69)90020-1.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

S. Harrigan, E. Zsoter, H. Cloke, P. Salamon, and C. Prudhomme. Daily ensemble river discharge reforecasts and real-time forecasts from the operational global flood awareness system. *Hydrology and Earth System Sciences*, 27(1):1–19, 2023. doi: 10.5194/hess-27-1-2023. URL https://hess.copernicus.org/articles/27/1/2023/.

Tristan Hascoet, Victor Pellet, Satoru Oishi, and Takemasa Miyoshi. Differentiable river routing for end-to-end learning of hydrological processes. *Journal of Geophysical Research: Machine Learning and Computation*, 3(1):e2025JH000760, 2026. doi: https://doi.org/10.1029/2025JH000760. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JH000760. e2025JH000760 2025JH000760.

Feyera A. Hirpa, Peter Salamon, Hylke E. Beck, Valerio Lorini, Lorenzo Alfieri, Ervin Zsoter, and Simon J. Dadson. Calibration of the global flood awareness system (glofas) using daily streamflow data. *Journal of Hydrology*, 566:595–606, 2018. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2018.09.052. URL https://www.sciencedirect.com/science/article/pii/S0022169418307467.

George J. Huffman, David T. Bolvin, Dan Braithwaite, Kuo-Lin Hsu, Robert J. Joyce, Christopher Kidd, Eric J. Nelkin, Soroosh Sorooshian, Erich F. Stocker, Jackson Tan, David B. Wolff, and Pingping Xie. *Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG)*, pp. 343–353. Springer International Publishing, Cham, 2020. ISBN 978-3-030-24568-9. doi: 10.1007/978-3-030-24568-9_19. URL https://doi.org/10.1007/978-3-030-24568-9_19.

Mohammad Sina Jahangir and John Quilty. Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. *Journal of Hydrology*, 629:130498, 2024. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2023.130498. URL https://www.sciencedirect.com/science/article/pii/S0022169423014403.

Shijie Jiang, Larisa Tarasova, Guo Yu, and Jakob Zscheischler. Compounding effects in flood drivers challenge estimates of extreme river floods. *Science Advances*, 10(13), March 2024. ISSN 2375-2548. doi: 10.1126/sciadv.adl4005. URL http://dx.doi.org/10.1126/sciadv.adl4005.

Nikolas Kirschstein and Yixuan Sun. The merit of river network topology for neural flood forecasting. In *International Conference on Machine Learning*, pp. 24713–24725. PMLR, 2024.

D. Klotz, F. Kratzert, M. Gauch, A. Keefe Sampson, J. Brandstetter, G. Klambauer, S. Hochreiter, and G. Nearing. Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, 26(6):1673–1693, 2022. doi: 10.5194/hess-26-1673-2022. URL https://hess.copernicus.org/articles/26/1673/2022/.

J. M. Van Der Knijff, J. Younis, and A. P. J. De Roo and. Lisflood: a gis-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24(2):189–212, 2010. doi: 10.1080/13658810802549154. URL https://doi.org/10.1080/13658810802549154.

Basil Kraft, Martina Kauzlaric, William Hugo Aeberhard, Massimiliano Zappa, and Lukas Gudmundsson. Drop: A scalable deep learning approach for runoff simulation and river routing. 2025.

Basil Kraft, Steven Stalder, William H. Aeberhard, Nicolás Harrington Ruiz, Nicolai Meinshausen, Xinwei Shen, and Lukas Gudmundsson. Modeling uncertainty with engression: A deep generative time-series approach. *Geophysical Research Letters*, 53(2):e2025GL120122, 2026. doi: https://doi.org/10.1029/2025GL120122. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025GL120122. e2025GL120122 2025GL120122.

F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018. doi: 10.5194/hess-22-6005-2018. URL https://hess.copernicus.org/articles/22/6005/2018/.

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

F. Kratzert, M. Gauch, D. Klotz, and G. Nearing. Hess opinions: Never train a long short-term memory (lstm) network on a single basin. *Hydrology and Earth System Sciences*, 28(17):4187–4201, 2024. doi: 10.5194/hess-28-4187-2024. URL https://hess.copernicus.org/articles/28/4187/2024/.

Simon Lang, Mihai Alexe, Mariana CA Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D Dueben, Sara Hahner, et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2(1):18, 2026. doi: 10.1038/s44387-026-00073-7.

Jinyang Li, Kuolin Hsu, Soroosh Sorooshian, and Dan Lu. From lumped to spatially distributed hydrologic modeling: A data-driven framework evaluated across north american catchments. *Water Resources Research*, 62(9):e2025WR043068, 2026. doi: https://doi.org/10.1029/2025WR043068. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025WR043068. e2025WR043068 2025WR043068.

Bruno Merz, Stefano Basso, Svenja Fischer, David Lun, Günter Blöschl, Ralf Merz, Björn Guse, Alberto Viglione, Sergiy Vorogushyn, Elena Macdonald, Luzie Wietzke, and Andreas Schumann. Understanding heavy tails of flood peak distributions. *Water Resources Research*, 58(6), 2022. ISSN 1944-7973. doi: 10.1029/2021wr030506. URL http://dx.doi.org/10.1029/2021WR030506.

H. Mosaffa, F. Pappenberger, C. Prudhomme, M. Chantry, C. Rüdiger, and H. Cloke. A gnn routing module is all you need for lstm rainfall–runoff models. *Hydrology and Earth System Sciences*, 30(7):2079–2092, 2026. doi: 10.5194/hess-30-2079-2026. URL https://hess.copernicus.org/articles/30/2079/2026/.

Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, et al. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563, 2024. doi: https://doi.org/10.1038/s41586-024-07145-1.

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.

Zhigang Ou, Congyi Nai, Baoxiang Pan, Yi Zheng, Chaopeng Shen, Peishi Jiang, Xingcai Liu, Qiuhong Tang, Wenqing Li, and Ming Pan. Probabilistic diffusion models advance extreme flood forecasting. *Geophysical Research Letters*, 52(15):e2025GL115705, 2025. doi: https://doi.org/10.1029/2025GL115705. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025GL115705. e2025GL115705 2025GL115705.

Karan Ruparell, Robert J. Marks, Andy Wood, Kieran M. R. Hunt, Hannah L. Cloke, Christel Prudhomme, Florian Pappenberger, and Matthew Chantry. Hydra-lstm: A semi-shared machine learning architecture for prediction across watersheds. *Artificial Intelligence for the Earth Systems*, 4(3):240103, 2025. doi: 10.1175/AIES-D-24-0103.1. URL https://journals.ametsoc.org/view/journals/aies/4/3/AIES-D-24-0103.1.xml.

Karan Ruparell, Tristan Hascoet, Takemasa Miyoshi, Kieran MR Hunt, Hannah L Cloke, Christel Prudhomme, and Florian Pappenberger. Joint distribution of upstream runoff governs downstream river-discharge prediction uncertainty in distributed ml models. *arXiv preprint arXiv:2607.03217*, 2026a.

Karan Ruparell, Kieran MR Hunt, Hannah Cloke, Christel Prudhomme, Florian Pappenberger, and Matthew Chantry. Ai-generated ensemble river flow forecasting: Using rollout and an additional noise input to build ensemble forecasts. 2026b.

Fedor Scholz, Manuel Traub, Christiane Zarfl, Thomas Scholten, and Martin V. Butz. Fully differentiable, fully distributed rainfall-runoff modeling. *Hydrology and Earth System Sciences*, 29 (21):6257–6283, November 2025. ISSN 1607-7938. doi: 10.5194/hess-29-6257-2025. URL http://dx.doi.org/10.5194/hess-29-6257-2025.

Mohamad Hakam Shams Eddin, Ribana Roscher, and Juergen Gall. Location-aware adaptive normalization: A deep learning approach for wildfire danger forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023. doi: 10.1109/TGRS.2023.3285401.

Mohamad Hakam Shams Eddin, Yikui Zhang, Stefan Kollet, and Jürgen Gall. Rivermamba: A state space model for global river discharge and flood forecasting. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 133401–133423. Curran Associates, Inc., 2025. doi: 10.52202/085713-4446. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c1501f836c75dd1da19974a38c64ee1e-Paper-Conference.pdf.

Louise Slater, Georgios Blougouras, Liangkun Deng, Qimin Deng, Emma Ford, Anne Hoek van Dijke, Feini Huang, Shijie Jiang, Yinxue Liu, Simon Moulds, Andrew Schepen, Jiabo Yin, and Boen Zhang. Challenges and opportunities of ml and explainable ai in large-sample hydrology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2302), 2025. ISSN 1471-2962. doi: 10.1098/rsta.2024.0287. URL http://dx.doi.org/10.1098/rsta.2024.0287.

Yalan Song, Tadd Bindas, Chaopeng Shen, Haoyu Ji, Wouter J. M. Knoben, Leo Lonzarich, Martyn P. Clark, Jiangtao Liu, Katie van Werkhoven, Sam Lamont, Matthew Denno, Ming Pan, Yuan Yang, Jeremy Rapp, Mukesh Kumar, Farshid Rahmani, Cyril Thébault, Richard Adkins, James Halgren, Trupesh Patel, Arpita Patel, Kamlesh Arun Sawadekar, and Kathryn Lawson. High-resolution national-scale water modeling is enhanced by multiscale differentiable physics-informed machine learning. *Water Resources Research*, 61(4):e2024WR038928, 2025. doi: https://doi.org/10.1029/2024WR038928. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024WR038928. e2024WR038928 2024WR038928.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers–scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025.

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: https://doi.org/10.1029/2021WR030394. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021WR030394. e2021WR030394 2021WR030394.

Vera Thiemig, Goncalo N. Gomes, Jon O. Skøien, Markus Ziese, Armin Rauthe-Schöch, Elke Rustemeier, Kira Rehfeldt, Jakub P. Walawender, Christine Kolbe, Damien Pichon, Christoph Schweim, and Peter Salamon. EMO-5: a high-resolution multi-variable gridded meteorological dataset for Europe. *Earth System Science Data*, 14(7):3249–3272, 2022. doi: 10.5194/essd-14-3249-2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

M. A. Vischer, N. Otero, and J. Ma. Spatially resolved rainfall streamflow modeling in central europe. *EGUsphere*, 2025:1–26, 2025. doi: 10.5194/egusphere-2024-3649. URL https://egusphere.copernicus.org/preprints/2025/egusphere-2024-3649/.

Chao Wang, Shijie Jiang, Yi Zheng, Feng Han, Rohini Kumar, Oldrich Rakovec, and Siqi Li. Distributed hydrological modeling with physics-encoded deep learning: A general framework and its application in the amazon. *Water Resources Research*, 60(4):e2023WR036170, 2024a. doi: https://doi.org/10.1029/2023WR036170. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR036170. e2023WR036170 2023WR036170.

Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*, 2024b.

Yihan Wang, Annan Yu, Lujun Zhang, Charuleka Varadharajan, and N Benjamin Erichson. Hydrodiffusion: diffusion-based probabilistic streamflow forecasting with a state space backbone. *arXiv preprint arXiv:2512.12183*, 2025.

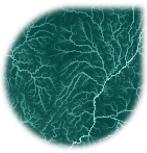
Pingping Xie, M Chen, and W Shi. Cpc unified gauge-based analysis of global daily precipitation. In *Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc*, volume 2, 2010.

Dai Yamazaki, Shinjiro Kanae, Hyungjun Kim, and Taikan Oki. A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, 47(4), April 2011. ISSN 1944-7973. doi: 10.1029/2010wr009726. URL http://dx.doi.org/10.1029/2010WR009726.

Yuan Yang, Dapeng Feng, Hylke E. Beck, Weiming Hu, Ather Abbas, Agniv Sengupta, Luca Delle Monache, Robert Hartman, Peirong Lin, Chaopeng Shen, and Ming Pan. Global daily discharge estimation based on grid long short-term memory (lstm) model and river routing. *Water Resources Research*, 61(6):e2024WR039764, 2025. doi: https://doi.org/10.1029/2024WR039764. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024WR039764. e2024WR039764 2024WR039764.

Michaël Zamo and Philippe Naveau. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, 50(2):209–234, 2018. doi: 10.1007/s11004-017-9709-7.

Distributed Hydrological Modeling in the Feature Space



Technical Appendix and Supplementary Material

Table of Contents

A Ablation Studies **18**

A.1 Feature importance 18

A.2 LISFLOOD vs AlphaEarth 18

A.3 Routing 19

A.4 Bi-directional Mamba 21

B Consistency with River Network Topology **21**

C Performance Distribution of distributed modeling **22**

D Performance in an Ungauged Setting **24**

D.1 Impact of distributed modeling on ungauged EFAS reanalysis 24

D.2 performance on ungauged CAMELS observations 25

D.3 performance on ungauged GRDC observations 25

E Comparison to RiverMamba **27**

E.1 Comparison to RiverMamba on gauged EFAS Reanalysis 27

E.2 Comparison to RiverMamba on gauged CAMELS Observations 31

F Comparison to operational EFAS forecasts **33**

F.1 Comparison to EFAS control run (deterministic forecast) 33

F.2 Comparison to EFAS ensemble run (probabilistic forecast) 34

G Hydrographs **36**

G.1 Example hydrographs from training on EFAS Reanalysis data 36

G.2 Example hydrographs from fine-tuning on CAMELS observational data 38

G.3 Example hydrographs from fine-tuning on GRDC observational data 40

H Case Studies of Floods Across Europe **41**

H.1 2024 Augsburg Germany Flood 41

H.2 2024 Central European floods 43

H.3 2024 Spanish floods 45

H.4 2025 United Kingdom Floods . . . 47
H.5 2025 Balkan Rivers Floods . . . 49

I Non-linear reservoir **51**

J Additionl Results **52**
J.1 Experiments on EFAS Reanalysis . . . 52
J.1.1 Experiments on EFAS Reanalysis for CAMELS split . . . 52
J.1.2 Experiments on EFAS Reanalysis for GRDC split . . . 55
J.2 Experiments on GRDC Observations . . . 59
J.3 Experiments on CAMELS Observations . . . 63

K Implementation details **67**
K.1 Computational time . . . 69
K.2 Serialization . . . 69
K.3 Location-aware Adaptive Normalization (LOAN) . . . 70

L Evaluation Metrics **70**

M Floods return periods **71**

N Baselines **72**
N.1 Climatology . . . 72
N.2 Persistence . . . 72
N.3 EFAS Forecast . . . 72
N.4 RiverMamba . . . 72
N.5 DRUM . . . 72
N.6 ED-LSTM (Encoder-Decoder LSTM . . . 73
N.7 ED-LSTM (Mean Embedding LSTM) . . . 73

O Dataset **73**
O.1 EFAS reanalysis river discharge data . . . 73
O.2 Observational river discharge data . . . 73
O.2.1 GRDC . . . 73
O.2.2 CAMELS . . . 76
O.3 IFS-HRES data . . . 76
O.4 Analysis data . . . 78
O.5 ERA5-Land reanalysis data . . . 78
O.6 EMO data . . . 78
O.7 IMERG data . . . 78
O.8 CPC data . . . 79

O.9 Static data 79
O.9.1 AlphaEarth 79
O.9.2 LISFLOOD 79

P Code and data availability **80**

A Ablation Studies

For the ablation studies, we conducted experiments over a Mid European domain ($55.5°$N $46.5°$S, $5.5°$E $15.5°$E) which has $280,820$ points on land and includes 480 GRDC and 892 CAMELS stations for evaluation. Fig. 85 gives an overview of the domain. For the experiments, we run only deterministic model for $L = 16$ lead time including training on nowcasting (t=0) and evaluate on the validation set 2022-2023.

A.1 Feature importance

In Table 6, we study the impact of different input features on the performance of our model. In general, removing any input feature degrades the performance. However, the model is still functional whenever we remove one of the input. Removing the observational IMERG has less effect than removing CPC. Although removing the IFS-HRES weather forecast has the biggest impact, the model is still produces a skillful forecast estimating only from the initial conditions. Finally, the best performance is achieved when we use all input variables.

Table 6: Ablation studies for feature importance on the validation set over Europe. Shown are KGE for CAEMLS (KGE_c^r) and GRDC (KGE_g^r) splits on reanalysis. along with the $F1^r$@(1.5-500) for reanalysis over the full domain (281K points). In the second half of the table, we report the results of $R2_c^o$, KGE_c^o, and $F1_c^o$@(1.5-20) on the fine-tunning on CAMELS observations.

Results on EFAS Reanalysis						
Static	CPC	IMERG	EMO	Analysis	IFS-HRES	$R2_c^r$ | KGE_c^r | $R2_g^r$ | KGE_g^r | $F1^r(\uparrow)$
✗	✓	✓	✓	✓	✓	0.6963 | 0.7338 | 0.6919 | 0.7443 | 0.2734
✓	✗	✓	✓	✓	✓	0.6741 | 0.6972 | 0.6674 | 0.7145 | 0.3053
✓	✓	✗	✓	✓	✓	0.7021 | 0.7303 | 0.7041 | 0.7411 | 0.3100
✓	✓	✓	✗	✓	✓	0.6863 | **0.7474** | 0.6637 | 0.7116 | 0.3101
✓	✓	✓	✓	✗	✓	0.6869 | 0.6979 | 0.6835 | 0.7107 | 0.3069
✓	✓	✓	✓	✓	✗	0.5065 | 0.5817 | 0.5262 | 0.6125 | 0.2549
✓	✓	✓	✓	✓	✓	**0.7061** | 0.7396 | **0.7043** | **0.7493** | **0.3394**
Results on CAMELS observational data						
Static	CPC	IMERG	EMO	Analysis	IFS-HRES	$R2_c^o$ | KGE_c^o | $F1_c^o(\uparrow)$
✗	✓	✓	✓	✓	✓	0.6863 | 0.7163 | 0.2691
✓	✗	✓	✓	✓	✓	0.6668 | 0.7308 | 0.3069
✓	✓	✗	✓	✓	✓	**0.7006** | **0.7493** | 0.2857
✓	✓	✓	✗	✓	✓	0.6964 | 0.7100 | 0.2101
✓	✓	✓	✓	✗	✓	0.6704 | 0.7157 | 0.2133
✓	✓	✓	✓	✓	✗	0.5376 | 0.5646 | 0.1008
✓	✓	✓	✓	✓	✓	0.6990 | 0.7472 | **0.3109**

A.2 LISFLOOD vs AlphaEarth

In this experiment, we study the impact of replacing LISFLOOD static river attribute features (Sec. O.9.2) by the embedding of AlphaEarth from Google DeepMind. LISFLOOD is considered an important source for modeling in hydrology and it is used in state-of-the-art physics-based EFAS and GloFAS systems. In contrast to a learned embedding like AlphaEarth, LISFLOOD static features have physical meaning. We report the result in Table 7. In the first row, we remove all static features and keep only the Cartesian coordinates as a positional encoding (Sec. O.9.2). removing the static features reduces the performance for flood detection the most and has a bigger impact on fine-tunning on observations. More interesting is that AlphaEarth embedding can be used to replace the LISFLOOD static maps in our model. The advantage is that, AlphaEarth embedding is learned from radar, satellites, and 3D laser measurements. It exists globally at 10 meter resolution and does not need the efforts to generate over 100 static maps and calibration parameters globally. As a comparison, in this experiment, we used all 64 features from AlphaEarth compared to 119 LISFLOOD static features.

Table 7: Ablation studies for static feature importance on the validation set over Europe. Shown are KGE for CAEMLS (KGE_c^r) and GRDC (KGE_g^r) splits on reanalysis. along with the $F1^r$@(1.5-500) for reanalysis over the full domain (281K points). In the second half of the table, we report the results of $R2_c^o$, KGE_c^o, and $F1_c^o$@(1.5-20) on the fine-tunning on CAMELS observations.

Results on EFAS Reanalysis		
LISFLOOD Static	AlphaEarth	$R2_c^r$ | KGE_c^r | $R2_g^r$ | KGE_g^r | $F1^r$(↑)
✗	✗	0.6963 | 0.7338 | 0.6919 | 0.7443 | 0.2734
✓	✗	0.7045 | **0.7404** | 0.6943 | 0.7462 | 0.3360
✗	✓	**0.7061** | 0.7396 | **0.7043** | **0.7493** | **0.3394**
Results on CAMELS observational data		
LISFLOOD Static	AlphaEarth	$R2_c^o$ | KGE_c^o | $F1_c^o$(↑)
✗	✗	0.6863 | 0.7163 | 0.2691
✓	✗	0.6821 | 0.7101 | 0.2059
✗	✓	**0.6990** | **0.7472** | **0.3109**

A.3 ROUTING

In Table 8, we show if an explicit routing can replace the learned features by Bi-SSD in the forecasting block. Note that in this experiment, we keep the Bi-SSD in the encoder and remove only the spatio-temporal modeling in the forecasting block by removing Bi-SSD in the forecasting block. From the Table 8, we can see a clear benefit to keep the spatio-temporal modeling with Bi-SSD even if we have an explicit routing module. The reason could be that Bi-SSD is useful to capture features over the whole domain which improves the forecasting over long horizon and high-resolution i.e., Bi-SSD can capture if a heavy rain could happen in the next steps based on its large receptive field. This spatio-temporal information over the domain is only captured on a small-scale with routing.

Table 8: Ablation studies for the importance of Bi-SSD in the forecasting block. Shown are KGE for CAEMLS (KGE_c^r) and GRDC (KGE_g^r) splits on reanalysis. along with the $F1^r$@(1.5-500) for reanalysis over the full domain (281K points). We also report the results of $R2_c^o$, KGE_c^o, and $F1_c^o$@(1.5-20) on the fine-tunning on CAMELS observations.

	EFAS Reanalysis	CAMELS Observations
Bi-SSD	$R2_c^r$ | KGE_c^r | $R2_g^r$ | KGE_g^r | $F1^r$(↑)	$R2_c^o$ | KGE_c^o | $F1_c^o$(↑)
✗	0.6715 | 0.7057 | 0.6307 | 0.6915 | 0.3225	0.6793 | 0.7324 | 0.2947
✓	**0.7061** | **0.7396** | **0.7043** | **0.7493** | **0.3394**	**0.6990** | **0.7472** | **0.3109**

In Table 9, we study the impact of having more than one routing module or having multi-step routing in the forecasting block. We noticed that adding more than one routing module reduces generalization of the model and a better performance was obtained by using only one routing module for each forecasting step i.e., doing routing or distributed modeling every 6 hours.

Table 9: Ablation studies for the number of routing module/distributed steps in the forecasting block. Shown are KGE for CAEMLS (KGE_c^r) and GRDC (KGE_g^r) splits on reanalysis. along with the $F1^r$@(1.5-500) for reanalysis over the full domain (281K points).

Results on EFAS Reanalysis	
# Routing modules	$R2_c^r$ | KGE_c^r | $R2_g^r$ | KGE_g^r | $F1^r$(↑)
×1	**0.7061** | **0.7396** | **0.7043** | **0.7493** | **0.3394**
×2	0.6547 | 0.7057 | 0.6289 | 0.6650 | 0.2839

In Table 10, we show supplementary results for Table 4 (c) with different upstream aggregation techniques. In "Connection per Neighbor", we run SSD from each upstream point to its downstream (SSD on small blocks with a sequence of 2 each) e.g., a point with, 3 valid upstream neighbors

appears as the token in 3 separate 2-token sequences. The point final routing vector is the mean of its self token outputs across all its incoming upstream edges. In "Binary Mask", we study if routing propagates useful upstream-to-downstream information or if it just suppresses points with no upstream to reduce false positives. For this in the second row of Table 10, we check if a point has at least 1 upstream neighbor, we double its own feature vector i.e., no information is propagated from upstream neighbors. If a point has no neighbors, we do not add anything to the feature vector. Like this "Binary Mask" adds an inductive bias to the model informing it explicitly about the topology without propagating information between grids. As can be seen from the first row, better results are achieved when we use all upstream points and sort them with upstream area as a one long sequence. The second row experiment verifies that the routing in the model propagates useful upstream-to-downstream information between the gird points and learns features beyond the topology of the graph. In Fig. 7, we show the impact of adding routing on the flood severity detection qualitatively.

Table 10: Ablation studies for different upstream aggregation techniques as a supplementary to Table 4 (c). Shown are KGE for CAEMLS (KGE_c^r) and GRDC (KGE_g^r) splits on reanalysis. along with the $F1^r$@(1.5-500) for reanalysis over the full domain (281K points). We also report the results of $R2_c^o$, KGE_c^o, and $F1_c^o$@(1.5-20) on the fine-tunning on CAMELS observations.

	EFAS Reanalysis	CAMELS Observations
Type	$R2_c^r$ | KGE_c^r | $R2_g^r$ | KGE_g^r | $F1^r$(↑)	$R2_c^o$ | KGE_c^o | $F1_c^o$(↑)
Connection per Neighbor	0.7027 | 0.7264 | **0.7064** | 0.7456 | 0.2950	0.6936 | 0.7318 | 0.2461
Binary Mask	0.6622 | 0.6816 | 0.6404 | 0.6751 | 0.1122	0.6245 | 0.7158 | 0.2954
All+Sort(area)	**0.7061** | **0.7396** | 0.7043 | **0.7493** | **0.3394**	**0.6990** | **0.7472** | **0.3109**

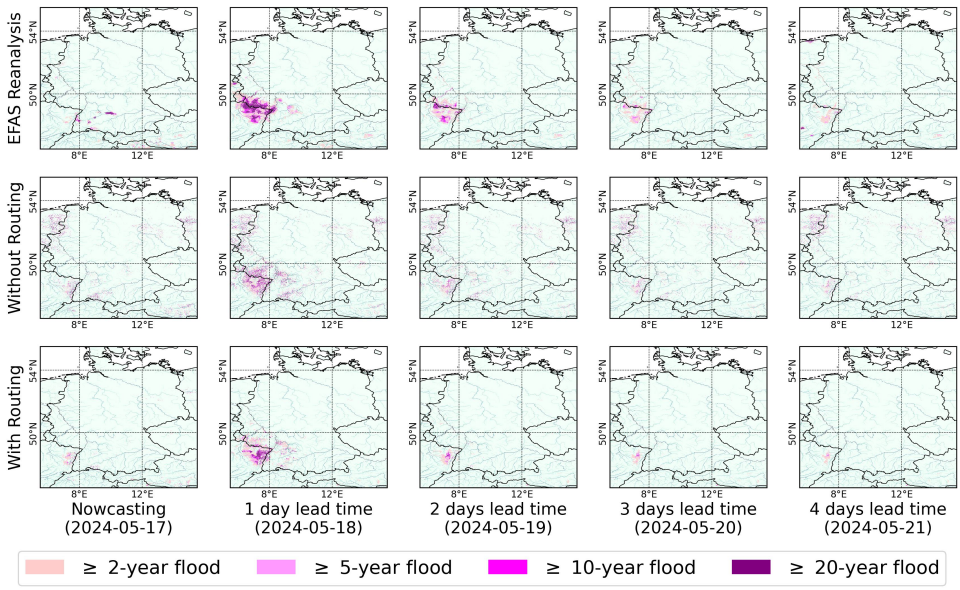


Figure 7: Comparison between EFAS reanalysis (top, used as a reference), a model without routing (middle row), a model with routing (bottom row) during an extreme flood in Saarland Germany , mid May 2024. Shown are flood severity maps at 1 arcminute resolution where each panel shows flood extent at different lead times from nowcasting to 4 days lead time. Without explicit distributed modeling, the routing-free model produces heterogeneous predictions in which neighboring pixels can have inconsistent discharge, leading to many false positives. Our routing module instead yields more homogeneous and spatially clustered predictions. In addition, the routing module makes the predicted river discharge consistent with topology across the river skeletons (Sec. B).

A.4 Bi-directional Mamba

In this experiment, we compare the Mamba as in RiverMamba [Shams Eddin et al., 2025] to the Mamba2 variant used in this work. Bi-directional Mamba blocks in RiverMamba use two separately-parametrized SSMs for the forward and backward passes. In our work, we use a shared SSM for bi-directionality. Inside each block, we build a second copy of the sequence by reversing the order of points along the space-filling curve, batch it together with the forward-order sequence, and pass both through the same SSD block in one call. Table 11 compares the two approaches. From the table we can see that the SSD algorithm of Mamba2 utilizes the GPU more efficiently. With Mamba2, we can train and run inference faster than RiverMamba. This is important since we have a rollout mechanism in the forecasting engine and we need an efficient algorithm to run the forward and backward passes simultaneously.

Table 11: Ablation studies for bi-directional Mamba in comparison to RiverMamba. Shown are training time for in epoch in minutes and inference time in seconds.

Results on EFAS Reanalysis		
Bi-directionality	Training time* (↓) 1 epoch 16 steps	Inference time* (↓) 280,820 points
×2 Mamba (RiverMamba)	∼159 min	6.20±0.005 sec
×1 Mamba2 (Ours)	∼49 min	4.10±0.008 sec

* tested on NVIDIA A40 GPU.

B Consistency with River Network Topology

In this experiments, we study the effect of adding explicit routing to the model on having consistent forecast along the river network topology. For this, we train a deterministic model with and without routing on the Mid European domain for 16 time steps + 1 nowcasting.

It is assumed that every upstream pixel should have less or equal river discharge value than its downstream direction since catchment area only grows moving downstream. For every pixel with a defined local drain direction (LDD) direction, we evaluate this constraint on the forecast river discharge value. We perform the test over 2024-2025 with 728 forecasting issued dates for 277,805 checkable pixel pairs. We also consider a small tolerance of 5% relative $+\,0.1\,\mathrm{m^3 s^{-1}}$ absolute per pixel pair. Table 12 shows that adding the routing improves consistency at every lead time. The improvement when adding routing is up to $+0.54\%$ at 96 h. In Fig. 8, we show the spatial distribution of constraint violation over the Mid European domain for a random date. As can be seen the routing provides more consistent forecast along the topology of LDD. It is also notable that the errors of the model without routing are not spread uniformly but concentrate at specific, presumably topologically difficult regions like river confluences, narrow channels, or steep local gradients.

Table 12: LDD consistency rate (ratio of pixel pairs satisfying downstream $\geq$ upstream) for the EFAS reanalysis as a reference and the two forecast models. Evaluation is averaged for the years 2024-2025 over the Mid European domain for 16 time steps and 1 nowcasting step.

	Overall	Lead Time t			
Source	Mean	24 h	48 h	72 h	96 h
EFAS Reanalysis (GT)	99.97%	-	-	-	-
✗ routing	99.09%	99.19%	99.34%	99.37%	99.35%
✓ routing	**99.62%**	**99.73%**	**99.87%**	**99.89%**	**99.89%**
Δ Benefit of adding routing	+0.53%	+0.54%	+0.53%	+0.52%	+0.54%

Note that even the EFAS reanalysis which is produced by a physics-based hydrological model (LISFLOOD) and uses routing with the same LDD is 99.97% consistent with the strict (downstream $\geq$ upstream) rule at a single 6-hourly timestep. Although, the difference in magnitude is typically small (median $\approx 3\%$ of the upstream value) and constraint violations are not spread uniformly across

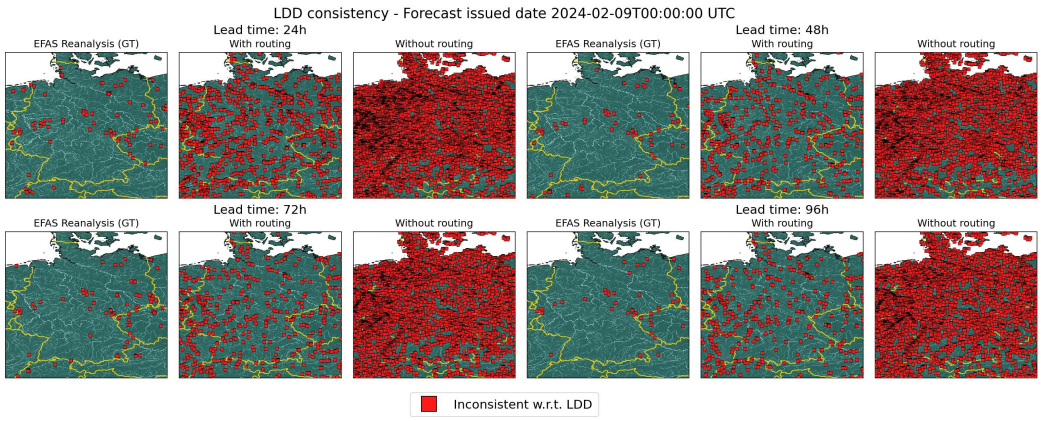


Figure 8: Local Drain Direction (LDD) consistency check (downstream $\geq$ upstream) for the forecast issued on 2024-02-09T00:00:00 UTC, at four lead times (24, 48, 72, and 96 h. The magenta (GT) or red (model) markers flag pixels where the LDD upstream point carries more discharge than its downstream neighbour.

the domain. We suspect that due to the landscape i.e., degree of slope and the finite propagation speed of the kinematic wave scheme between adjacent grid cells ($\sim$ 1.5 km apart at 1-arcminute EFAS resolution). At a single instant in time, a flood may have reached an upstream cell but not yet fully propagated into its downstream neighbour. Second Reservoirs can act as a temporary storage that can delay discharge propagation. Finally, the evaporation, irrigation, and withdrawals can also play different roles along the flow path.

C Performance Distribution of Distributed Modeling

In this experiment, we study the impact of routing spatially in a gauged setting. We trained for 16 steps + 1 step nowcasting and evaluated on 1458 daily forecasts spanning the years 2022-2025 against EFAS reanalysis as a reference. Each grid point is assigned an upstream-drainage-area class from the EFAS upstream area static field (Fig. 9). At every grid point and lead time we compute on the 4-years time series the metrics MAE, RMSE, R2, KGE, and F1.

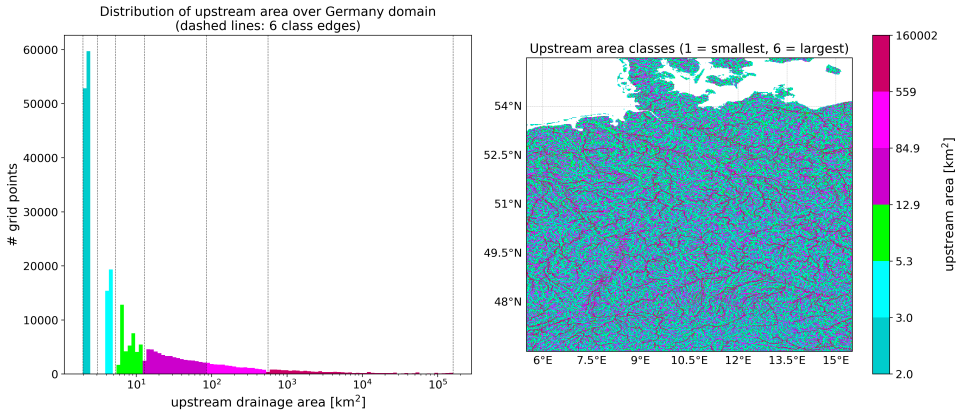


Figure 9: Upstream drainage area distribution over the Mid European domain with 280, 820 EFAS grid points and the 6 classification classes we used in Sec. C, Table 13, and Fig. 10. On the left side is the histogram of upstream area. On the right is the spatial map of the assigned class at every grid point starting from small headwater streams (class 1) to the largest rivers (class 6).

Table 13 reports the performance per class as the median over the class's grid points and averaged across lead times. Cells are colored Cyan where adding routing is better than without routing and Purple where it is worse. Routing improves every metric in every class except one the MAE for the

largest basins (class 6) which is marginally worse with routing (7.916 → 7.930), even though RMSE, R2 and KGE all improve for that class. The gain is also uneven across classes. It is largest for the smallest headwater cells (class 1: R2 0.016 → 0.417, KGE 0.236 → 0.532) and generally shrinks through the mid-sized classes, reaching its minimum at class 5 (R2 +0.029, KGE +0.041) before a small uptick at class 6 (R2 +0.060, KGE +0.052). For F1, routing improves detection at essentially every class and return period combination. Fig. 10 shows the spatial impact of the routing module on the forecasting skill.

Table 13: Gauged performance of the model with routing against a model without it by upstream-area class. Evaluation is done over 280, 820 EFAS grid points for the validation and test sets combined (2022-2025). For each class and metric, values are the class-median of the per-pixel MAE/RMSE/R2/KGE and averaged across the 17 lead times. Cyan marks cells where routing improves the metric and Purple marks cells where it is worse.

Class (Uparea km^2)	MAE (↓)	RMSE (↓)	R2 (↑)	KGE (↑)	F1@(1.5-500) (↑)
1: 1.95-3.04	0.0143 → 0.0078	0.0213 → 0.0133	0.016 → 0.417	0.236 → 0.532	0.1218 → 0.1279
2: 3.04-5.28	0.0180 → 0.0144	0.0306 → 0.0244	0.287 → 0.447	0.429 → 0.554	0.0687 → 0.1249
3: 5.28-12.88	0.0349 → 0.0257	0.0589 → 0.0437	0.243 → 0.476	0.399 → 0.581	0.0471 → 0.1215
4: 12.88-84.87	0.0960 → 0.0860	0.1681 → 0.1542	0.422 → 0.521	0.540 → 0.626	0.0578 → 0.1071
5: 84.87-559.26	0.5460 → 0.5327	1.0069 → 0.9718	0.559 → 0.588	0.630 → 0.671	0.0503 → 0.0851
6: 559.26-160002	7.9160 → 7.9300	14.408 → 14.003	0.557 → 0.617	0.641 → 0.693	0.0410 → 0.0840

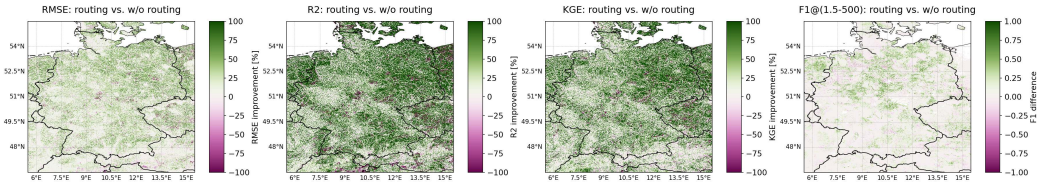


Figure 10: Per pixel percentage improvement from the model with routing over a model without routing, for RMSE^r, R2^r, KGE^r and F1^r@(1.5-500). For each grid point, the percentage improvement is computed as $\Delta\text{RMSE} = \frac{\text{RMSE}_{\text{(w/o routing)}} - \text{RMSE}_{\text{(routing)}}}{\text{RMSE}_{\text{(w/o routing)}}} \times 100$, $\Delta\text{R2} = \frac{\text{R2}_{\text{(routing)}} - \text{R2}_{\text{(w/o routing)}}}{|\text{R2}_{\text{(w/o routing)}}|} \times 100$, $\Delta\text{KGE} = \frac{\text{KGE}_{\text{(routing)}} - \text{KGE}_{\text{(w/o routing)}}}{|\text{KGE}_{\text{(w/o routing)}}|} \times 100$, where subscripts "routing" and "w/o routing" denote a model with routing and without it. F1 is computed per pixel and the plotted quantity is the plain difference $\Delta\text{F1} = \text{F1}_{\text{(routing)}} - \text{F1}_{\text{(w/o routing)}}$. The values are bounded in $[-1, 1]$. Positive (green) values indicate routing improves the metric at that pixel and negative (magenta) values indicate it is worse.

D PERFORMANCE IN AN UNGAUGED SETTING

D.1 IMPACT OF DISTRIBUTED MODELING ON UNGAUGED EFAS REANALYSIS

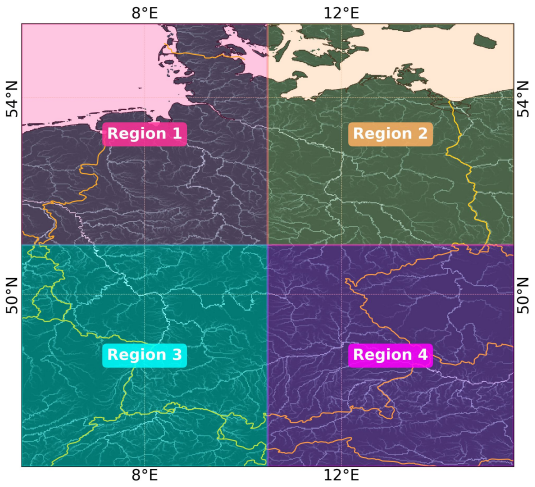


Figure 11: The 4 spatial-holdout regions used to construct the ungauged evaluation in Table 14 and Fig. 12. The Mid European domain is split into 4 equal quadrants at its geometric midpoint (51°N, 10.5°E). Region 1 (top-left), Region 2 (top-right), Region 3 (bottom-left) and Region 4 (bottom-right).

We repeat the routing ablation from Sec. C in a fully ungauged setting, using the same 6 classification classes of upstream area (Fig. 9). The Mid European domain is split into 4 quadrants (top-left, top-right, bottom-left, and bottom-right). For each experiments (with routing and without routing), we train 4 models with one holdout region. Training is done until 2021 so every point in the final map is genuinely out-of-sample in space and time in an ungauged setting on reanalysis. For evaluation, we evaluate on all $280,820$ grid points for 16 lead time + 1 nowcasting for the years 2022-2025. Table 14 reports the results for MAE, RMSE, R2, KGE, and F1 per class. Colored Cyan indicates where a model with routing is better than a model without routing and Purple where it is worse. The effect of routing is far larger here than in the gauged setting (Sec. C).

Without routing, the ungauged model is inaccurate. Adding routing turns R2 positive for classes 1-5 (e.g. class 1: $-6.758 \rightarrow 0.089$; class 3: $-1.034 \rightarrow 0.109$) and makes KGE positive for every class, including class 6 ($-0.036 \rightarrow 0.239$). MAE and RMSE improve at every class with no exceptions. For F1, routing improves detection at every class. Fig. 12 shows the spatial impact of the routing module on the forecasting skill in an ungauged setting.

Table 14: Ungauged (spatial-holdout) performance of routing against without routing by upstream-area class. The evaluation is at all $280,820$ EFAS grid points over the years 2022-2025. For each class and metric, values are the class-median of the per pixel metric and averaged across the 17 lead times. Cyan marks points where routing improves the metric and Purple marks points where it is worse.

Class (Uparea km^2)	MAE (↓)	RMSE (↓)	R2 (↑)	KGE (↑)	F1@(1.5-500) (↑)
1: 1.95-3.04	0.0346 → 0.0126	0.0536 → 0.0183	-6.758 → 0.089	-1.159 → 0.237	0.0539 → 0.0543
2: 3.04-5.28	0.0502 → 0.0228	0.0835 → 0.0338	-2.085 → 0.105	-0.349 → 0.268	0.0391 → 0.0519
3: 5.28-12.88	0.0843 → 0.0409	0.1348 → 0.0616	-1.034 → 0.109	-0.098 → 0.291	0.0334 → 0.0531
4: 12.88-84.87	0.2624 → 0.1469	0.4032 → 0.2264	-0.633 → 0.119	-0.031 → 0.300	0.0288 → 0.0454
5: 84.87-559.26	1.4830 → 0.9720	2.1890 → 1.4710	-0.602 → 0.143	-0.044 → 0.319	0.0266 → 0.0390
6: 559.26-160002	20.564 → 14.658	28.943 → 22.410	-0.918 → -0.144	-0.036 → 0.239	0.0186 → 0.0223

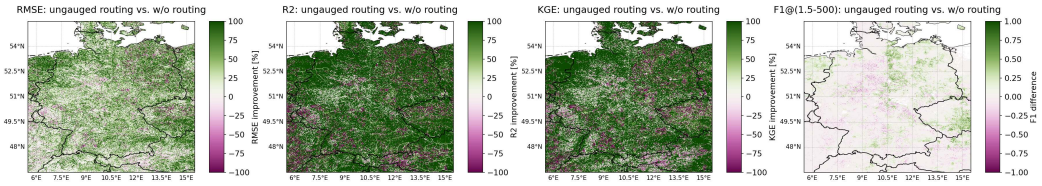


Figure 12: Per pixel percentage improvement for routing over without routing in the ungauged (spatial-holdout) setting. Metrics RMSE^r, R2^r, KGE^r, and F1^r@(1.5-500) are averaged over lead time for the years 2022-2025. For each grid point, the percentage improvement is computed as $\Delta\text{RMSE} = \frac{\text{RMSE}_{\text{(w/o routing)}} - \text{RMSE}_{\text{(routing)}}}{\text{RMSE}_{\text{(w/o routing)}}} \times 100$, $\Delta\text{R2} = \frac{\text{R2}_{\text{(routing)}} - \text{R2}_{\text{(w/o routing)}}}{|\text{R2}_{\text{(w/o routing)}}|} \times 100$, $\Delta\text{KGE} = \frac{\text{KGE}_{\text{(routing)}} - \text{KGE}_{\text{(w/o routing)}}}{|\text{KGE}_{\text{(w/o routing)}}|} \times 100$, where subscripts "routing" and "w/o routing" denote a model with routing and without it. F1 is computed per pixel and the plotted quantity is the plain difference $\Delta\text{F1} = \text{F1}_{\text{(routing)}} - \text{F1}_{\text{(w/o routing)}}$. The F1 values are bounded in $[-1, 1]$. Positive (green) values indicate routing improves the metric at that pixel and negative (magenta) values indicate it is worse.

D.2 Performance on ungauged CAMELS observations

We further test the fully ungauged setting against real station observations rather than EFAS reanalysis. The 892 CAMELS stations are split into two random halves 446 stations each (Fig. 13). We trained 2 models on each half of the stations for 4 days lead time + 1 nowcasting step against real observed discharge. For every station, we then keep only the prediction from the model that never saw it during training and evaluate for the years 2010 to 2024. Table 15 and Fig. 14 show that the model transfers well to unseen gauges. KGE is positive at the large majority of stations (mean 0.563, median 0.602), R2 is positive at most stations (mean 0.316, median 0.581), and floods are detected well across all return periods (F1 mean 0.185).

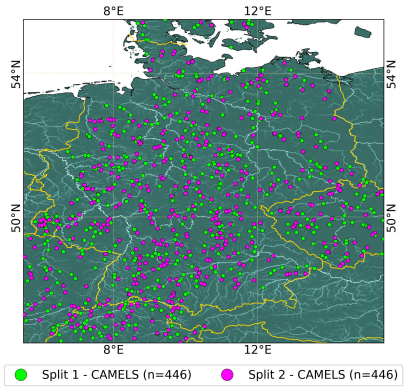


Figure 13: The two random holdout splits of the 892 CAMELS stations (446 each) used for the fully ungauged evaluation.

Table 15: Per station skill of the ungauged CAMELS prediction against observations.

	KGE (↑)	R2 (↑)	F1@(1.5-500) (↑)
mean	0.563	0.316	0.185
median	0.602	0.581	0.159

D.3 Performance on ungauged GRDC observations

We repeat the same protocol in Sec. D.2 for the 480 GRDC stations. The stations are split into 2 random halves 240 each (Fig. 15), We trained 2 models on each half of the stations for 4 days

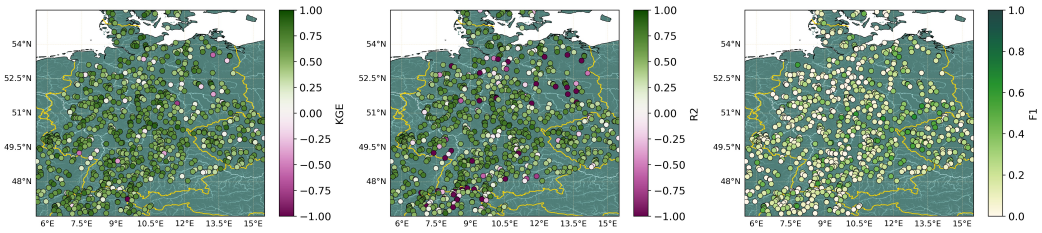


Figure 14: Per station KGE, R2, and F1 of the ungauged CAMELS prediction. Evaluation is done for the years 2010 to 2024 and for 4 days lead time + 1 nowcasting.

lead time + 1 nowcasting step against real observed discharge. We evaluate on years 2010 to 2025, such that every station is evaluated by the model that did not see it during training. The model also generalizes well to unseen GRDC gauges. Table 16 and Fig. 16 show the results. Mean KGE is 0.542 (median 0.595), mean R2 is 0.202 (median 0.545), and F1 remains skillful across return periods (mean 0.200).

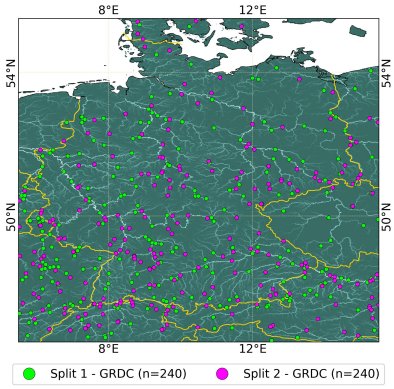


Figure 15: The two random holdout splits of the 480 GRDC stations (240 each) used for the fully ungauged evaluation.

Table 16: Per station skill of the ungauged GRDC prediction against observations.

	KGE (↑)	R2 (↑)	F1@(1.5-500) (↑)
mean	0.542	0.202	0.200
median	0.595	0.545	0.175

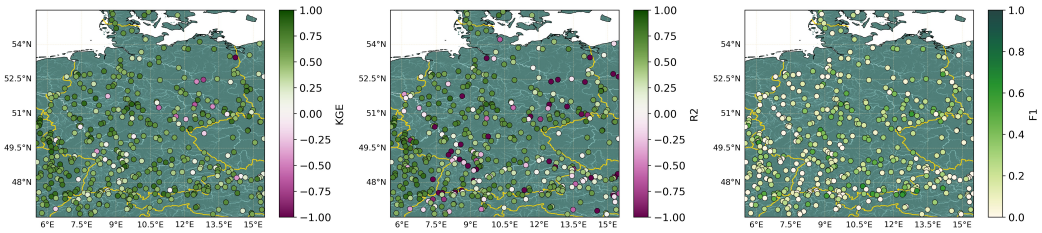


Figure 16: Per station KGE, R2, and F1 of the ungauged GRDC prediction. Evaluation is done for the years 2010 to 2025 and for 4 days lead time + 1 nowcasting.

E COMPARISON TO RIVERMAMBA

E.1 COMPARISON TO RIVERMAMBA ON GAUGED EFAS REANALYSIS

In this experiment, we compare to RiverMamba developed by Shams Eddin et al. [2025]. RiverMamba does not provide forecast for EFAS domain. In addition, it was trained with different spatio-temporal resolution and input data. For this, we train RiverMamba on our dataset over EFAS domain and with the same input data. For the expieremnt, we train over the Mid European domain up to 16 lead time steps (96 hours). Note that in principle RiverMamba is not scalable to long-term forecast. It has fixed learned horizon for the forecast (fixed lead time). Another difference is that RiverMamba does not provide probabilistic forecast nor it provides nowcasting. Our model in contrast, uses rollout for the forecast and can be scaled up to longer lead time. We could even increase the rollout step during inference compared to the training setting. In addition, our model provide nowcasting and probabilistic forecast (Table 17). Tables 18 and 19 show the results for EFAS reanalysis on both CAMELS and GRDC splits. As can be seen, our probabilistic model achieves the best results overall, while our deterministic model is consistently outperforming RiverMamba on F1 score. This is notable with the evaluation on the full domain with $280,820$ points (Table 20). The RiverMamba baseline has no explicit routing mechanism inside the model. It relies mainly on the spatio-temporal modeling of SSM to solve the routing implicitly, while our model does this explicitly with the routing module which improves flood detection. This is also consistent with the results in Table 4 (a) for the study of the routing impact. More results are provided in Figs. 17-20.

Table 17: Comparison to RiverMamba [Shams Eddin et al., 2025].

Model	# Param.	auto-regressive	Probabilistic	Nowcasting
RiverMamba	4.4 million	✗	✗	✗
Ours	2.3 million	✓	✓	✓

Table 18: Results on EFAS Reanalysis (CAMELS split). (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024-2025)		
	Model	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)
Determ	RiverMamba	0.7075±0.0138	0.7255±0.0077	0.1242±0.0160	0.5806±0.0221	0.6580±0.0118	0.0611±0.0104
Determ	Ours	0.7075±0.0056	0.7410±0.0134	0.1627±0.0143	0.5676±0.0097	0.6589±0.0162	0.0841±0.0037
Prob	Ours (mean)	**0.7247**	**0.7695**	**0.2129**	**0.6280**	**0.7284**	**0.2266**

Table 19: Results on EFAS Reanalysis (GRDC split). (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024-2025)		
	Model	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)
Determ	RiverMamba	0.7326±0.0099	0.7579±0.0052	0.1106±0.0125	0.6161±0.0222	0.6927±0.0141	0.0990±0.0151
Determ	Ours	0.6957±0.0273	0.7507±0.0273	0.1326±0.0272	0.5631±0.0237	0.6645±0.0241	0.1132±0.0136
Prob	Ours (mean)	**0.7359**	**0.7876**	**0.1998**	**0.6505**	**0.7552**	**0.2125**

Table 20: Results on EFAS Reanalysis (Full mid Europe domain). (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)		Test (2024-2025)	
	Model	F1@(1.5-20) (↑)	F1@(1.5-500) (↑)	F1@(1.5-20) (↑)	F1@(1.5-500) (↑)
Determ	RiverMamba	0.1670±0.0122	0.1407±0.0099	0.1713±0.0134	0.1379±0.0108
Determ	Ours	**0.3542**±0.0199	**0.3337**±0.0176	**0.3183**±0.0191	**0.3143**±0.0129

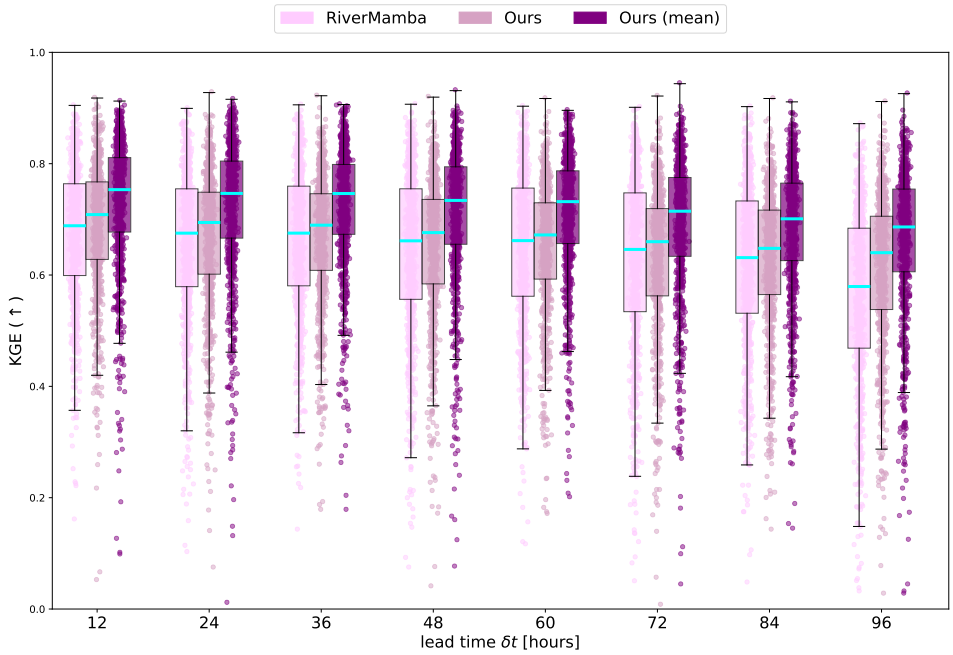


Figure 17: KGE of the river discharge forecasting with different lead time on CAMELS reanalysis split (test set 2024-2025 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

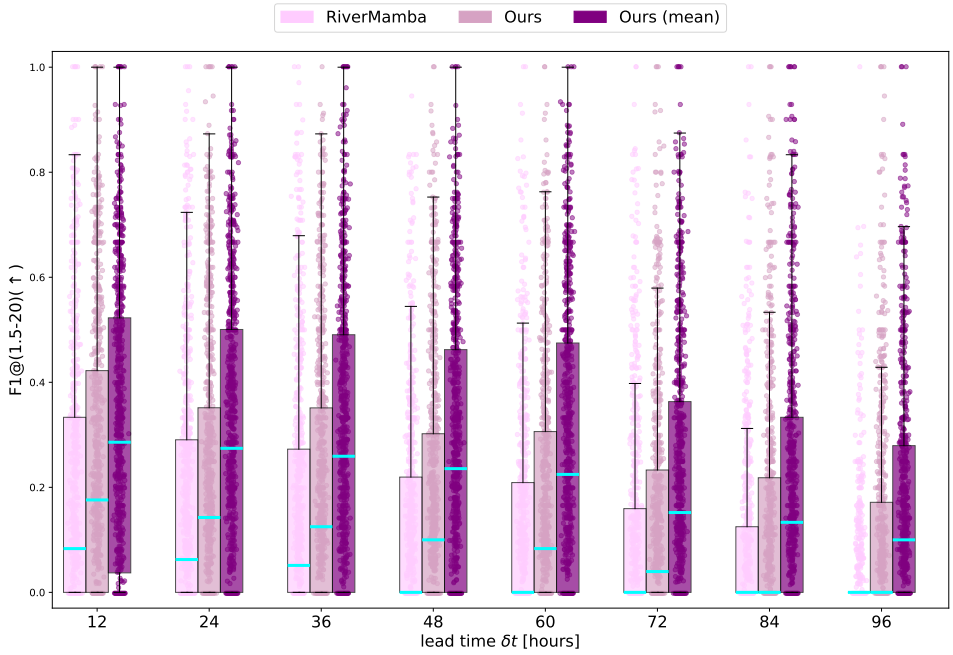


Figure 18: Averaged F1-score for RPs 1.5-20 with lead time on CAMELS reanalysis split (test set 2024-2025 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

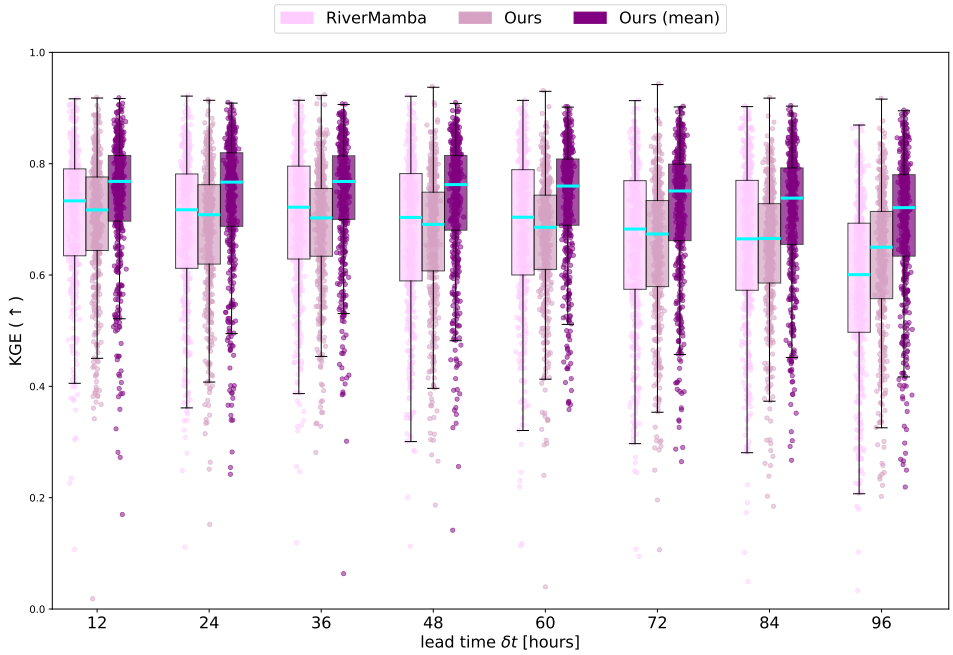


Figure 19: KGE of the river discharge forecasting with different lead time on GRDC reanalysis split (test set 2024-2025 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

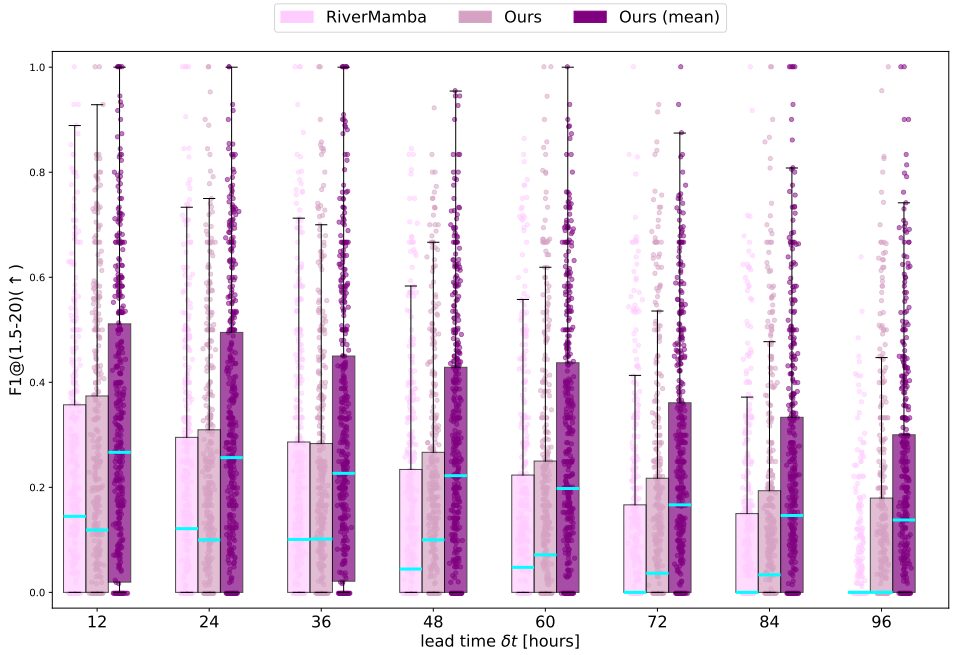


Figure 20: Averaged F1-score for RPs 1.5-20 with lead time on CAMELS reanalysis split (test set 2024-2025 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

In Figs. 21-24, we show the empirical cumulative distribution functions (CDFs) of KGE and F1-score@(1.5-10) across lead time for both CAMELS (N=892) and GRDC (N=480) reanalysis splits. A curve shifted to the right indicates a better performing model. Ours (mean) model's curve lies uniformly to the right of both the deterministic model (Ours) and the RiverMamba baseline. This is consistent across lead times, metrics, and splits. This indicates that the gain is broad across the catchments rather than concentrated in a few ones.

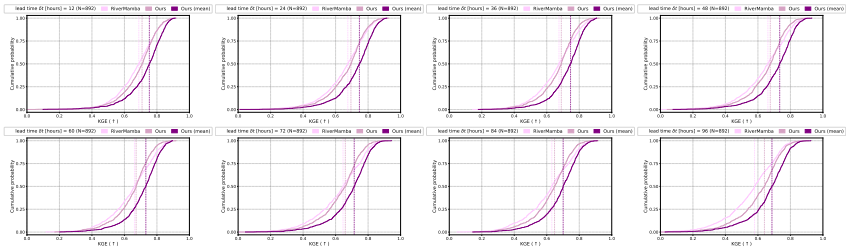


Figure 21: Cumulative distribution function (CDF) of station-level KGE on the CAMELS reanalysis split (N=892), across lead times. Dashed vertical lines shows each model's media.

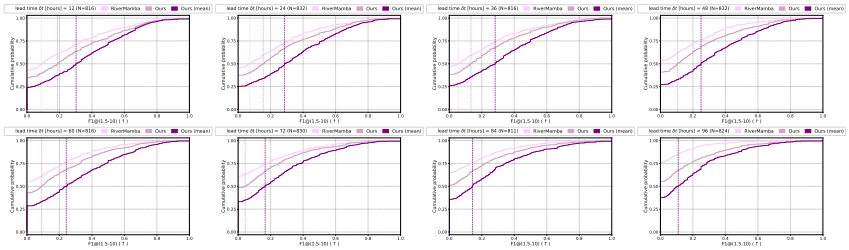


Figure 22: Cumulative distribution function (CDF) of station-level F1@(1.5-10) on the CAMELS reanalysis split (N=892), across lead times. Dashed vertical lines shows each model's media.

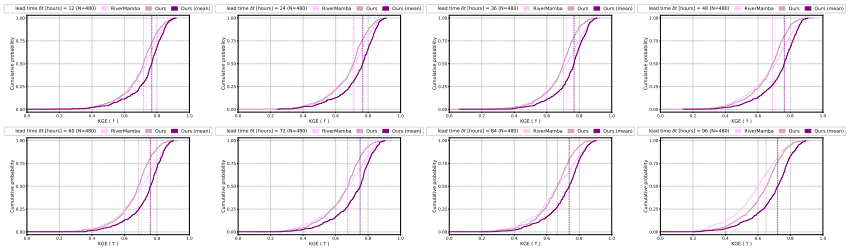


Figure 23: Cumulative distribution function (CDF) of station-level KGE on the GRDC reanalysis split (N=480), across lead times. Dashed vertical lines shows each model's media.

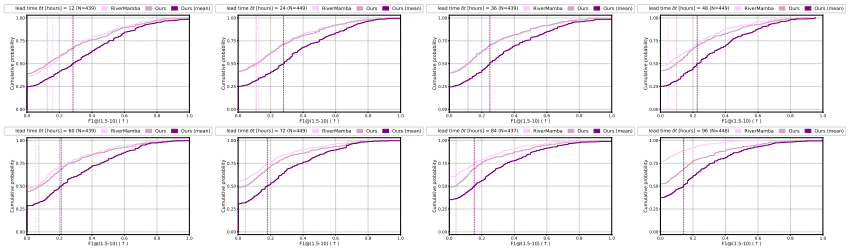


Figure 24: Cumulative distribution function (CDF) of station-level F1@(1.5-10) on the GRDC reanalysis split (N=480), across lead times. Dashed vertical lines shows each model's media.

In Figs. 25-26, we compute the per-station difference in KGE and F1@(1.5) between each of our two model variants, the deterministic model (Ours) and the probabilistic model (Ours (mean)) and the RiverMamba baseline. Positive (green) values indicate our model outperforms the baseline, while negative (magenta) values indicate the opposite.

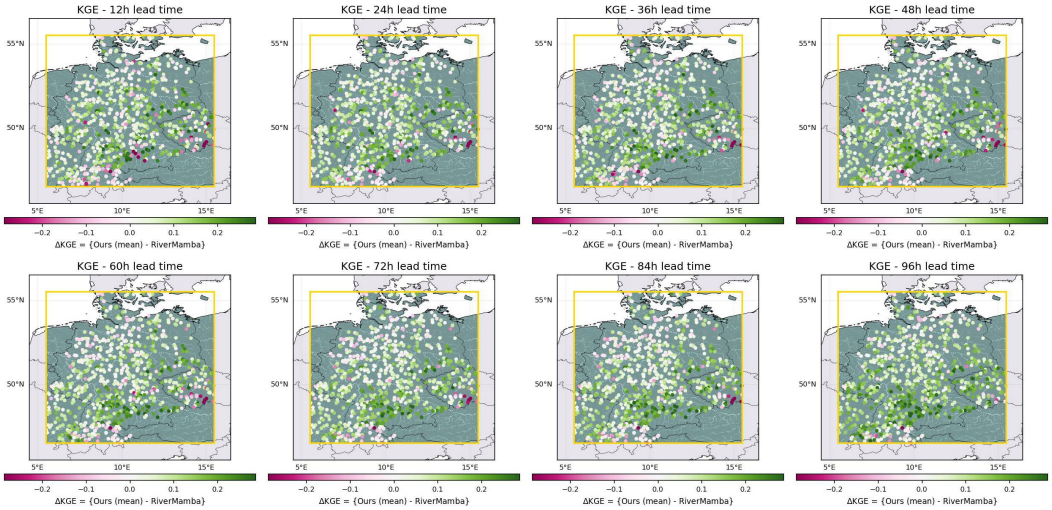


Figure 25: Comparison of KGE between Ours (mean) and RiverMamba on CAMELS EFAS reanalysis split across lead times (test set 2024-2025 temporally out-of-sample).

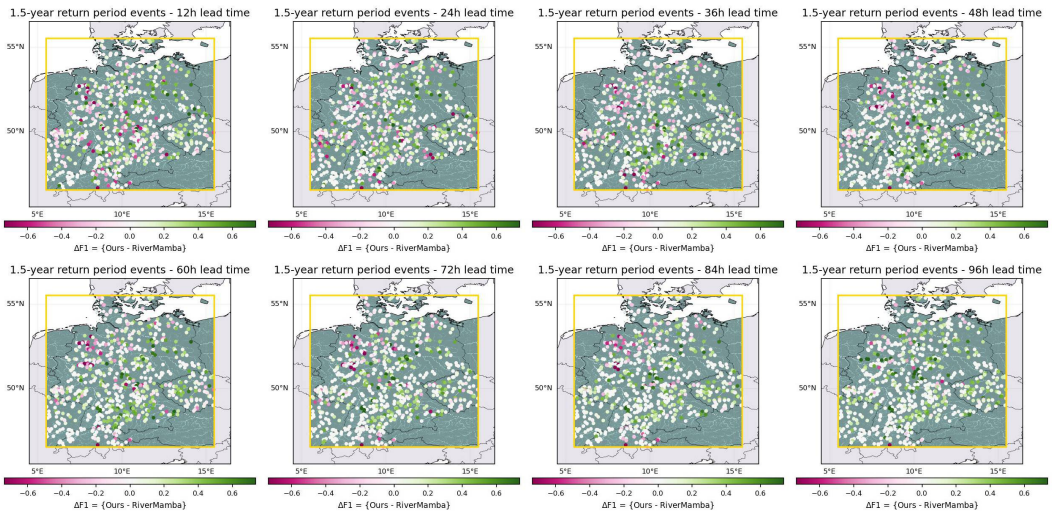


Figure 26: Comparison of F1-score between Ours and RiverMamba on CAMELS reanalysis EFAS split for the 1.5-year return period events across lead times (test set 2024-2025 temporally out-of-sample).

E.2 Comparison to RiverMamba on gauged CAMELS Observations

In this section, we compare to RiverMamba on observational CAMELS data. For this experiments, we used the models trained on reanalysis in Sec. E.1 and finetune them CAMELS observations for 4 days over the Mid European domain. In Table 21 we show the quantitative results and more results are provided in Figs. 27 and 28.

Table 21: Results on gauged CAMELS Observations. (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024)		
	Model	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)
Determ	RiverMamba	0.7236±0.0000	0.7563±0.0113	**0.3181**±0.0041	0.5428±0.0151	0.6257±0.0235	0.2864±0.0675
	Ours	**0.7286**±0.0011	0.7716±0.0065	0.2883±0.0008	**0.6216**±0.0313	0.6761±0.0256	0.2805±0.0107
Prob	Ours (mean)	0.7175	**0.7744**	0.3151	0.6134	**0.6941**	**0.3580**

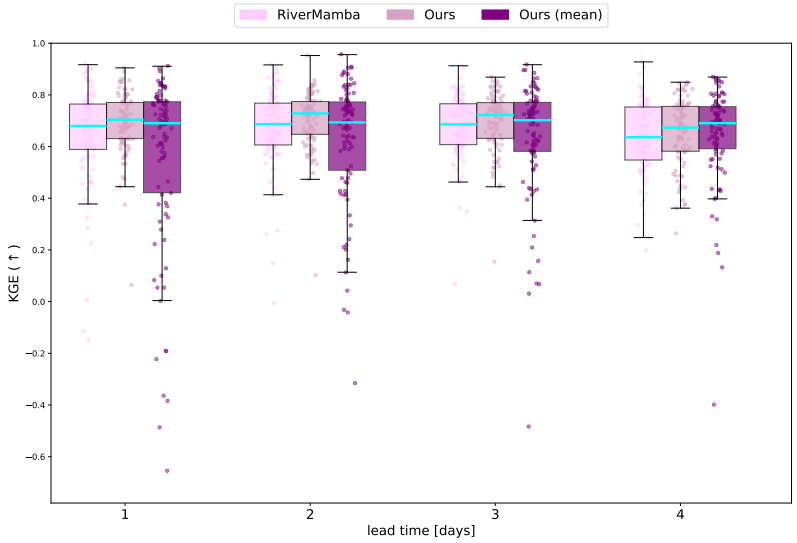


Figure 27: KGE of the river discharge forecasting with different lead time on CAMELS observations (test set 2024 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

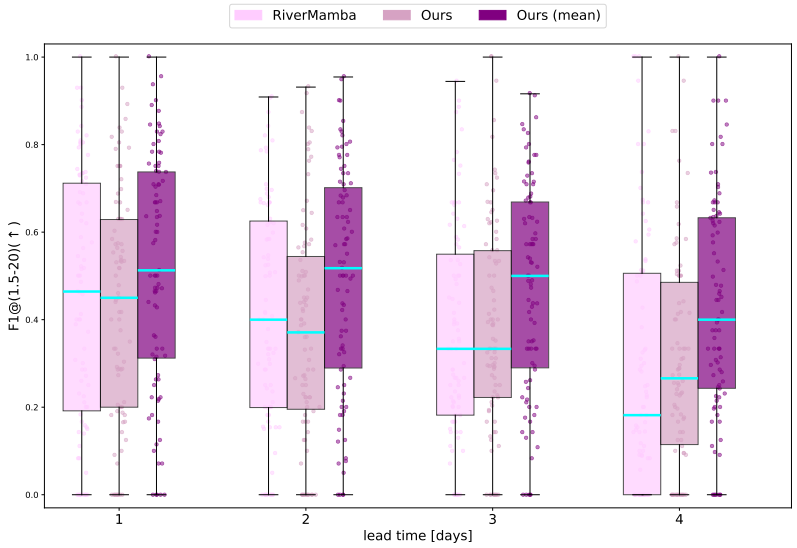


Figure 28: Averaged F1-score for 1.5-20 year return periods with lead time on CAMELS observations (test set 2024 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

F Comparison to operational EFAS forecasts

F.1 Comparison to EFAS control run (deterministic forecast)

In this section, we compare to the operational EFAS forecasts (Sec. N.3). We use the deterministic forecast from EFAS starting from mid May 2024 until end of 2025 and we validate on GRDC observations. Note that EFAS is a very strong baseline over the European domain. EFAS uses LISFLOOD hydrological model [Knijff et al., 2010] which has explicit kinematic wave equations for routing. It uses also river discharge from reanalysis as an initial condition while in our work we do not use any past river state as input to the model. The quantitative results are shown in Table 22 and qualitative results are shown in Fig. 29 and 30. A comparsion to the probabilsitic forecast from EFAS is shown in Sec. F.2.

Table 22: Results on gauged GRDC Observations. (±) denotes the standard deviation for 3 runs.

		Test set (2024-05-30 to 2025-12-21)				
	Model	RMSE (↓)	R2 (↑)	KGE (↑)	F1@(1.5-20) (↑)	F1@(1.5-500) (↑)
Determ	EFAS (control run)	9.8394	0.0481	0.3967	0.0945	0.0525
Determ	Ours	7.2889±0.1628	0.3413±0.0135	0.5180±0.0073	0.0703±0.0159	0.0512±0.0175
Prob	Ours (mean)	**7.2860**	**0.3673**	**0.5515**	**0.1015**	**0.1515**

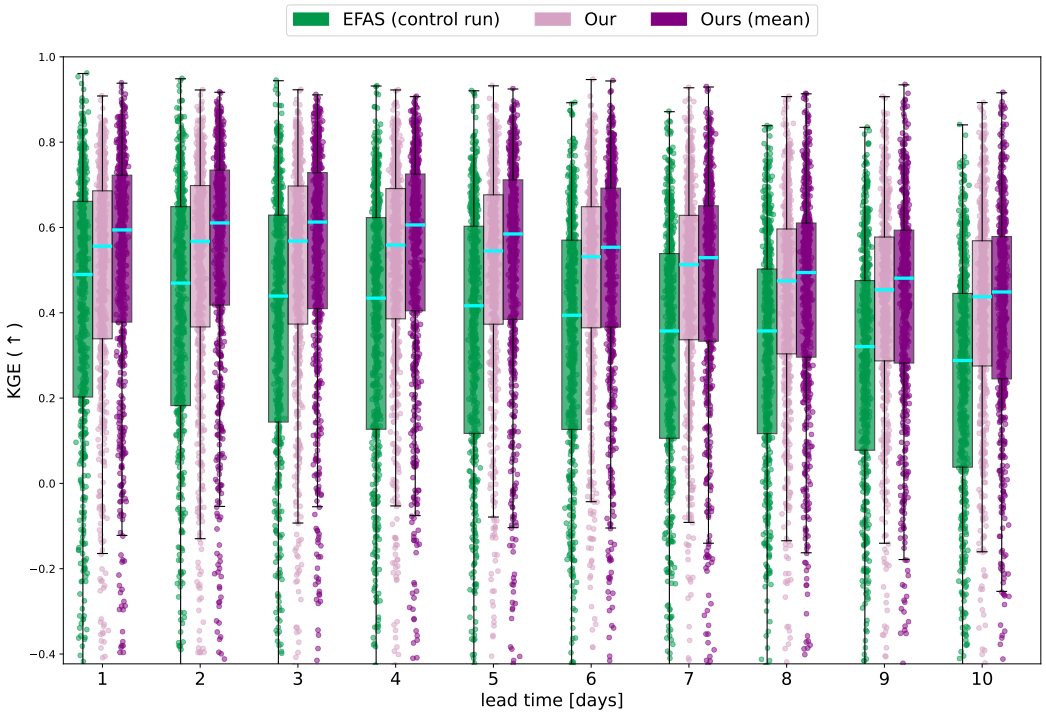


Figure 29: KGE of the river discharge forecasting with different lead time on GRDC observations (test set (2024-05-30 to 2025-12-21) temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

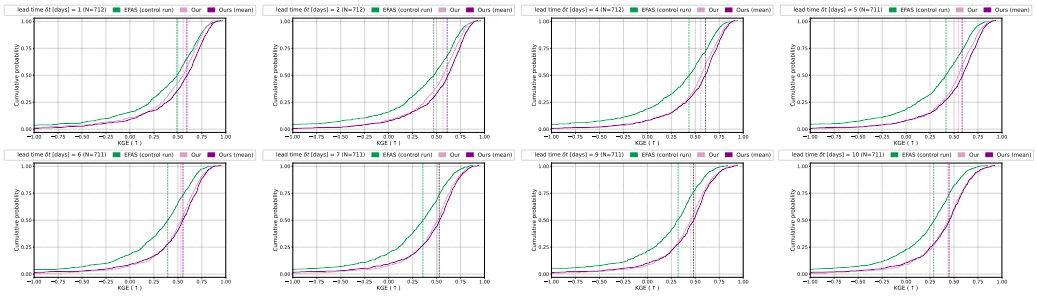


Figure 30: Cumulative distribution function (CDF) of station-level KGE on the GRDC observations, across lead times. Dashed vertical lines shows each model's media.

F.2 Comparison to EFAS ensemble run (probabilistic forecast)

Table 23: Results on gauged GRDC Observations.

		Test set (2024-06-01 to 2024-08-31)					
	Model	RMSE (↓)	R2 (↑)	KGE (↑)	CRPS (↓)	CRPS_{mae} (↓)	CRPS_{spread} (↑)
Prob	EFAS (mean of 51 ens members)	6.8122	-0.8517	0.1667	3.8862	5.4043	1.5180
	Ours (mean of 12 ens members)	**4.5627**	**0.0519**	**0.3718**	**2.1599**	**4.1280**	**1.9682**

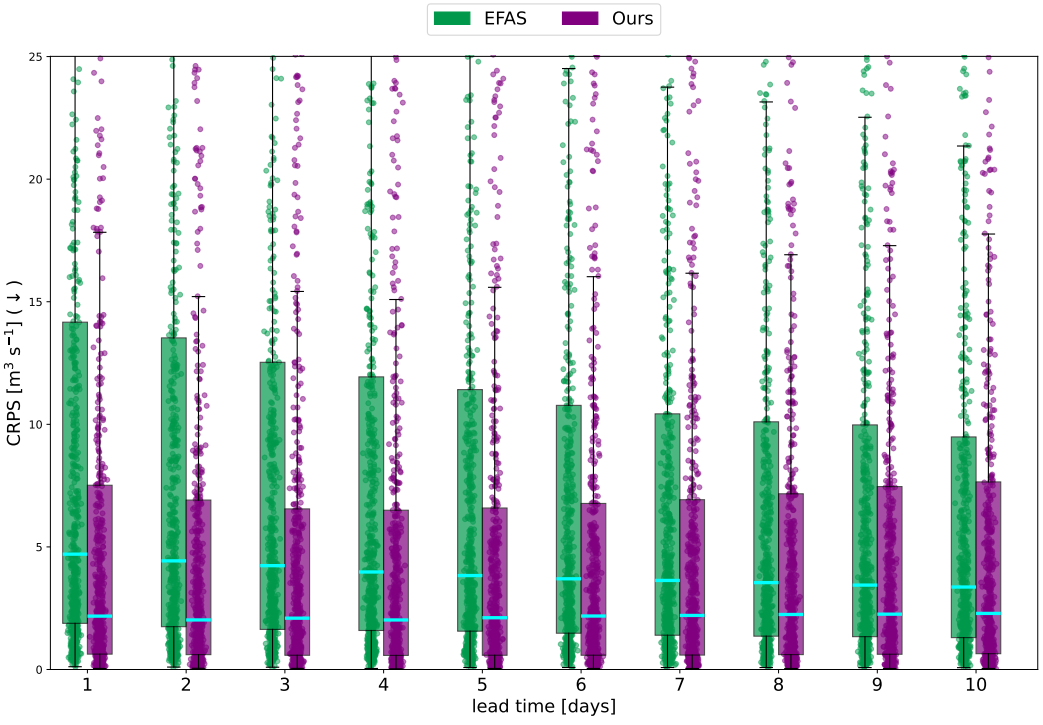


Figure 31: CRPS of the river discharge forecasting with different lead time on GRDC observations (test set (2024-06-01 to 2024-08-31) temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

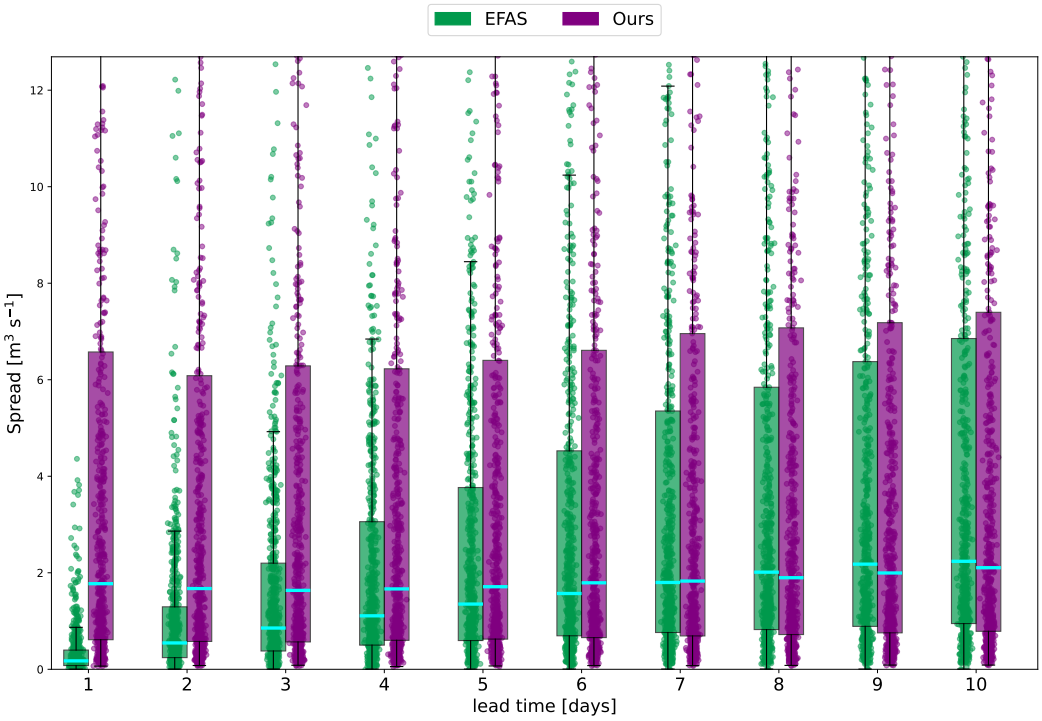


Figure 32: Ensemble spread of the river discharge forecasting with different lead time on GRDC observations (test set (2024-06-01 to 2024-08-31) temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

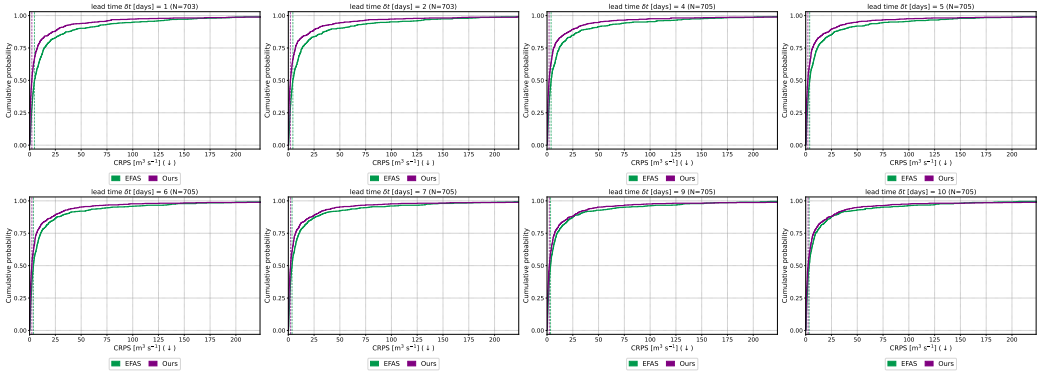


Figure 33: Cumulative distribution function (CDF) of station-level CRPS on the GRDC observations, across lead times. Dashed vertical lines shows each model's media.

G HYDROGRAPHS

G.1 EXAMPLE HYDROGRAPHS FROM TRAINING ON EFAS REANALYSIS DATA

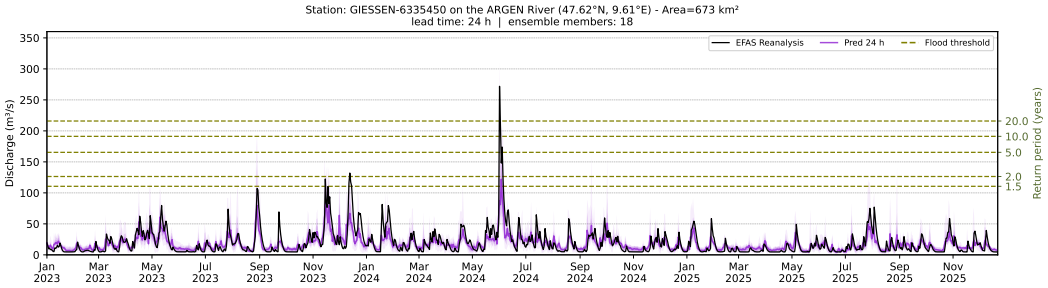


Figure 34: Ensemble forecast against EFAS reanalysis at Giessen (GRDC 6335450) on the Argen River (47.62°N, 9.61°E, catchment area 673 km^2), for lead time 24 h. For forecast, the ensemble median (solid line) is shown with nested uncertainty bands spanning the 25th–75th percentile (darkest), 10th–90th percentile, and min–max range (lightest) across all 18 ensemble members, against the EFAS reanalysis (black). Dashed olive lines are the EFAS flood return period thresholds (1.5-20.0 years).

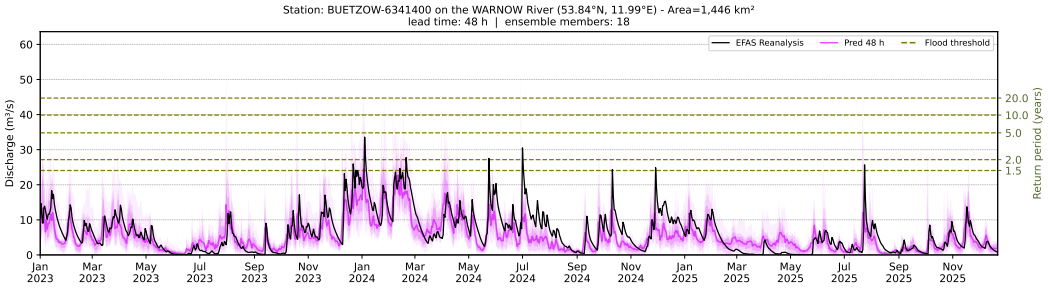


Figure 35: Ensemble forecast against EFAS reanalysis at Buetzow (GRDC 6341400) on the Warnow River (53.84°N, 11.99°E, catchment area 1,446 km^2), for lead time 48 h. For forecast, the ensemble median (solid line) is shown with nested uncertainty bands spanning the 25th–75th percentile (darkest), 10th–90th percentile, and min–max range (lightest) across all 18 ensemble members, against the EFAS reanalysis (black). Dashed olive lines are the EFAS flood return period thresholds (1.5-20.0 years).

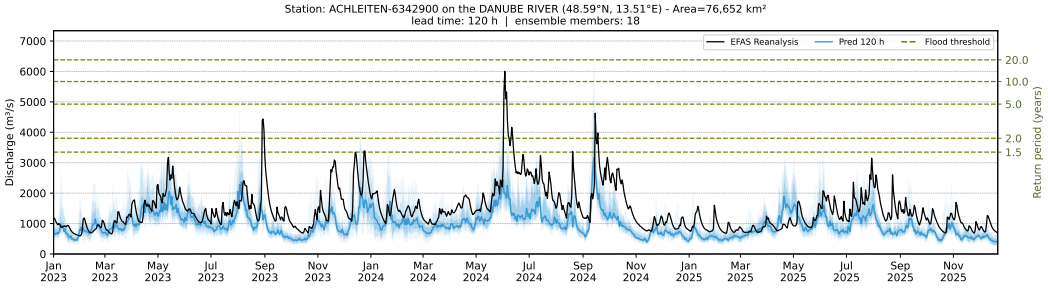


Figure 36: Ensemble forecast against EFAS reanalysis at Achleiten (GRDC 6342900) on the Danube River (48.59°N, 13.51°E, catchment area 76,652 km^2), for lead time 120 h. For forecast, the ensemble median (solid line) is shown with nested uncertainty bands spanning the 25th–75th percentile (darkest), 10th–90th percentile, and min–max range (lightest) across all 18 ensemble members, against the EFAS reanalysis (black). Dashed olive lines are the EFAS flood return period thresholds (1.5-20.0 years).

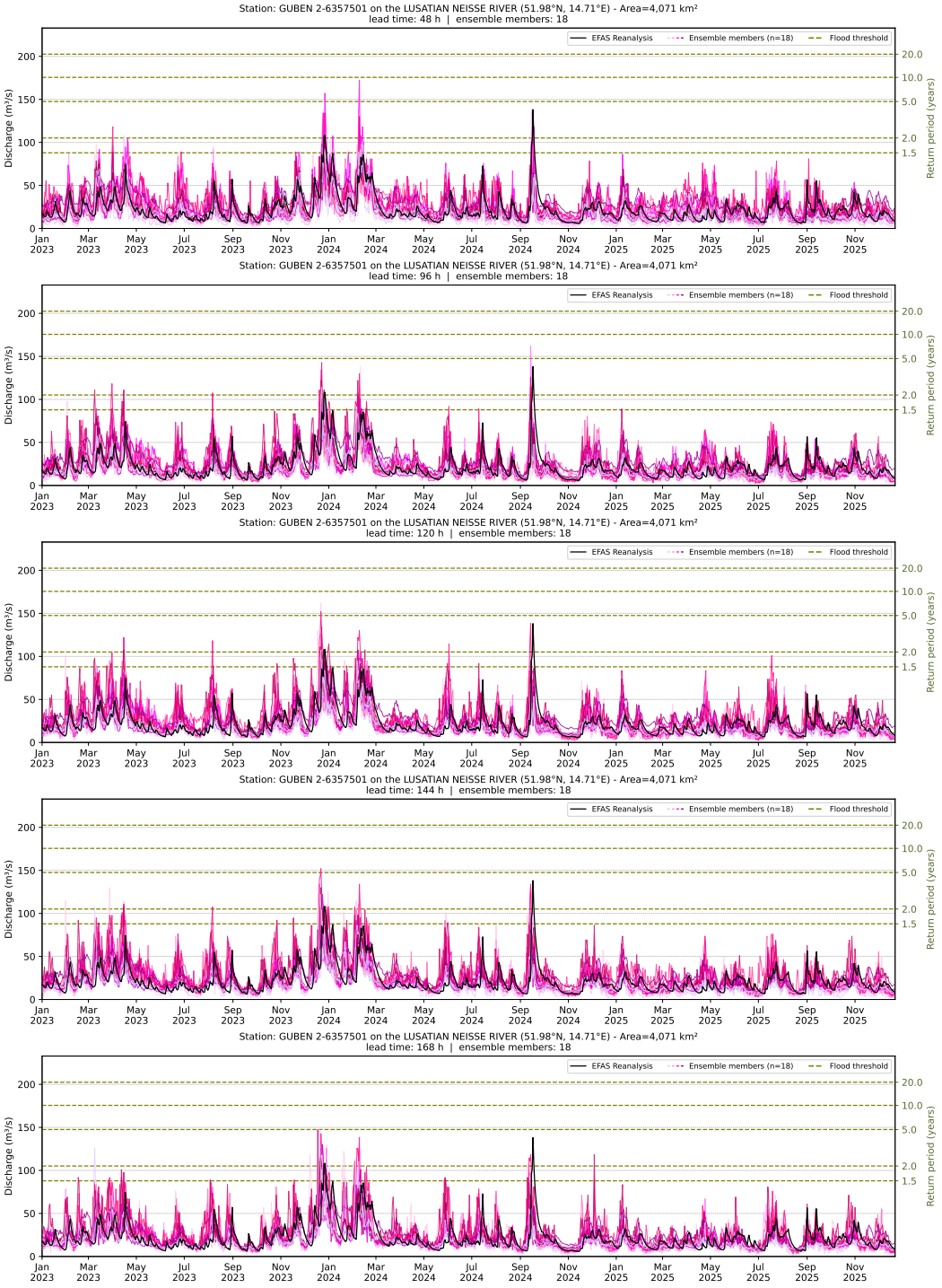


Figure 37: Ensemble forecasts against EFAS reanalysis at Guben (GRDC 6357501) on the Lech River (51.98°N, 14.71°E, catchment area 1,184 km^2), for five lead times (48 h, 96 h, 120 h, 144 h, 168). Each row shows the EFAS reanalysis discharge (black) against individual ensemble members (colored lines) for that lead time. Dashed olive lines are the EFAS flood return period thresholds (1.5-20.0 years).

G.2 EXAMPLE HYDROGRAPHS FROM FINE-TUNING ON CAMELS OBSERVATIONAL DATA

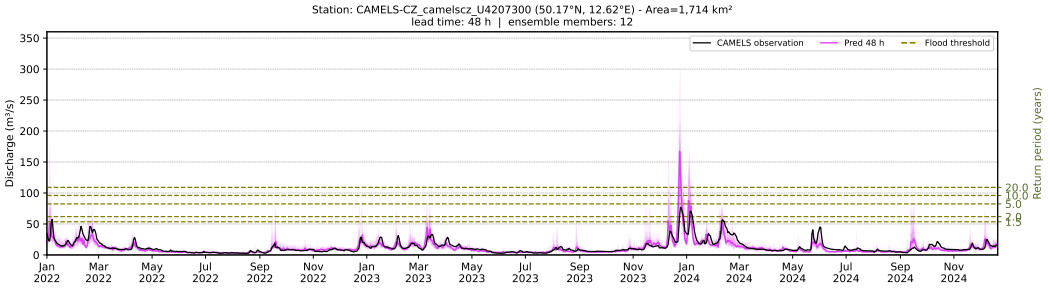


Figure 38: Ensemble forecast against CAMELS-CZ in Czechia for the station U420730 on the Ohře River (50.17°N, 12.62°E, catchment area $1,714$ km^2), for lead time 48 h. For forecast, the ensemble median (solid line) is shown with nested uncertainty bands spanning the 25th–75th percentile (darkest), 10th–90th percentile, and min–max range (lightest) across all 12 ensemble members, against the CAMELS-CZ observations (black). Dashed olive lines are the flood return period thresholds (1.5-20.0 years) computed from observations.

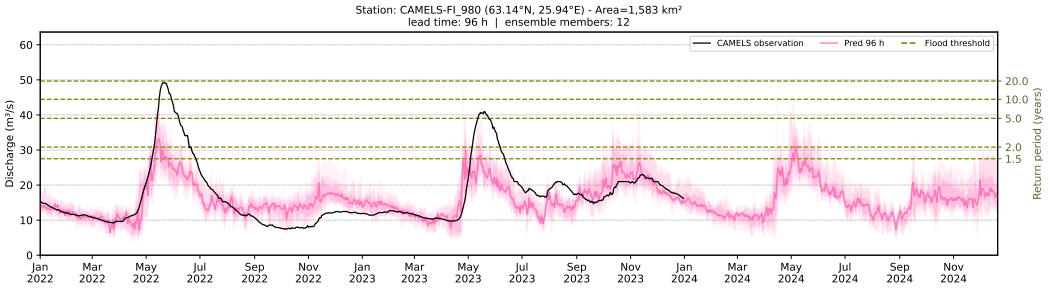


Figure 39: Ensemble forecast against CAMELS-FI in Finland for the station Kärnäjärvi-Kellankoski (FI_980) at the lake Kymijoki (63.14°N, 25.94°E, catchment area $1,583$ km^2), for lead time 96 h. For forecast, the ensemble median (solid line) is shown with nested uncertainty bands spanning the 25th–75th percentile (darkest), 10th–90th percentile, and min–max range (lightest) across all 12 ensemble members, against the CAMELS-FI observations (black). Dashed olive lines are the flood return period thresholds (1.5-20.0 years) computed from observations.

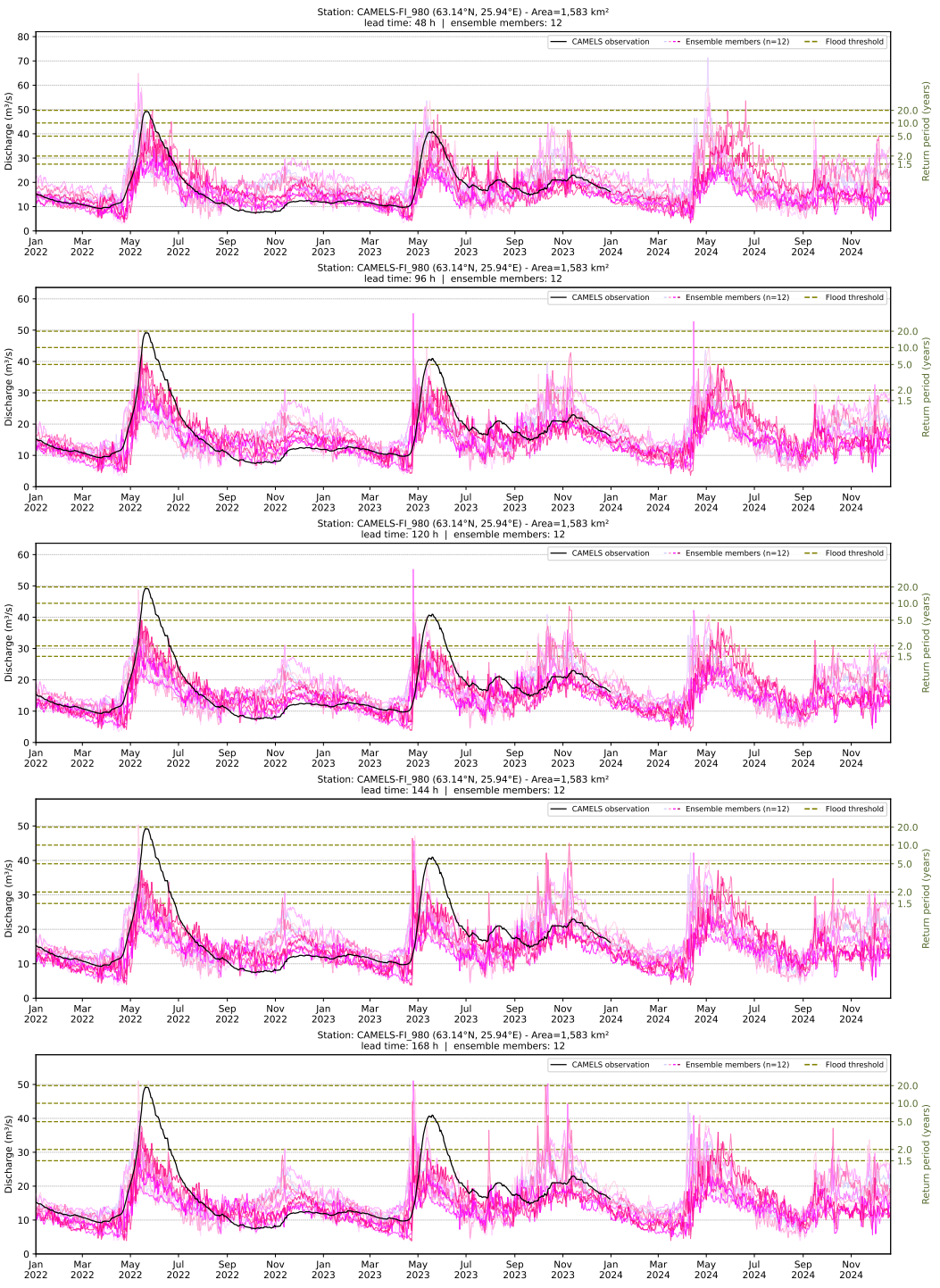


Figure 40: Ensemble forecasts against CAMELS-FI in Finland for the station Kärnäjärvi-Kellankoski (FI_980) at the lake Kymijoki (63.14°N, 25.94°E, catchment area $1,583$ km^2), for five lead times (48 h, 96 h, 120 h, 144 h, 168). Each row shows the CAMELS-CZ observational discharge (black) against individual ensemble members (colored lines) for that lead time. Dashed olive lines are the flood return period thresholds (1.5-20.0 years) computed from observations.

G.3 EXAMPLE HYDROGRAPHS FROM FINE-TUNING ON GRDC OBSERVATIONAL DATA

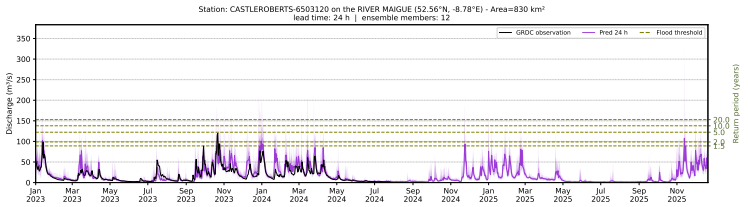


Figure 41: Ensemble forecast against GRDC observations at Castle Roberts (GRDC 6503120) on the Maigue River (52.56°N, 8.78°W, catchment area 830 km^2), for lead time 24 h. For forecast, the ensemble median (solid line) is shown with nested uncertainty bands spanning the 25th–75th percentile (darkest), 10th–90th percentile, and min–max range (lightest) across all 18 ensemble members, against the EFAS reanalysis (black). Dashed olive lines are the flood return period thresholds (1.5-20.0 years) computed from observations.

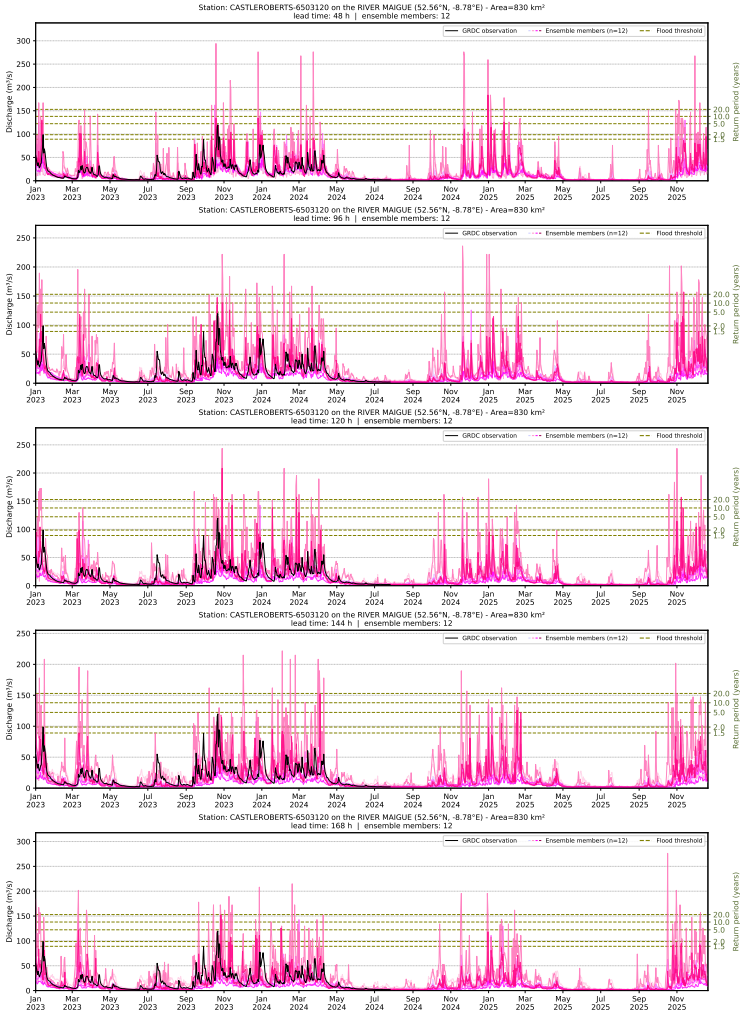


Figure 42: Ensemble forecasts against GRDC observations at Castle Roberts (GRDC 6503120) on the Maigue River (52.56°N, 8.78°W, catchment area 830 km^2), for five lead times (48 h, 96 h, 120 h, 144 h, 168). Each row shows the GRDC observational discharge (black) against individual ensemble members (colored lines) for that lead time. Dashed olive lines are the flood return period thresholds (1.5-20.0 years) computed from observations.

H CASE STUDIES OF FLOODS ACROSS EUROPE

H.1 2024 AUGSBURG GERMANY FLOOD

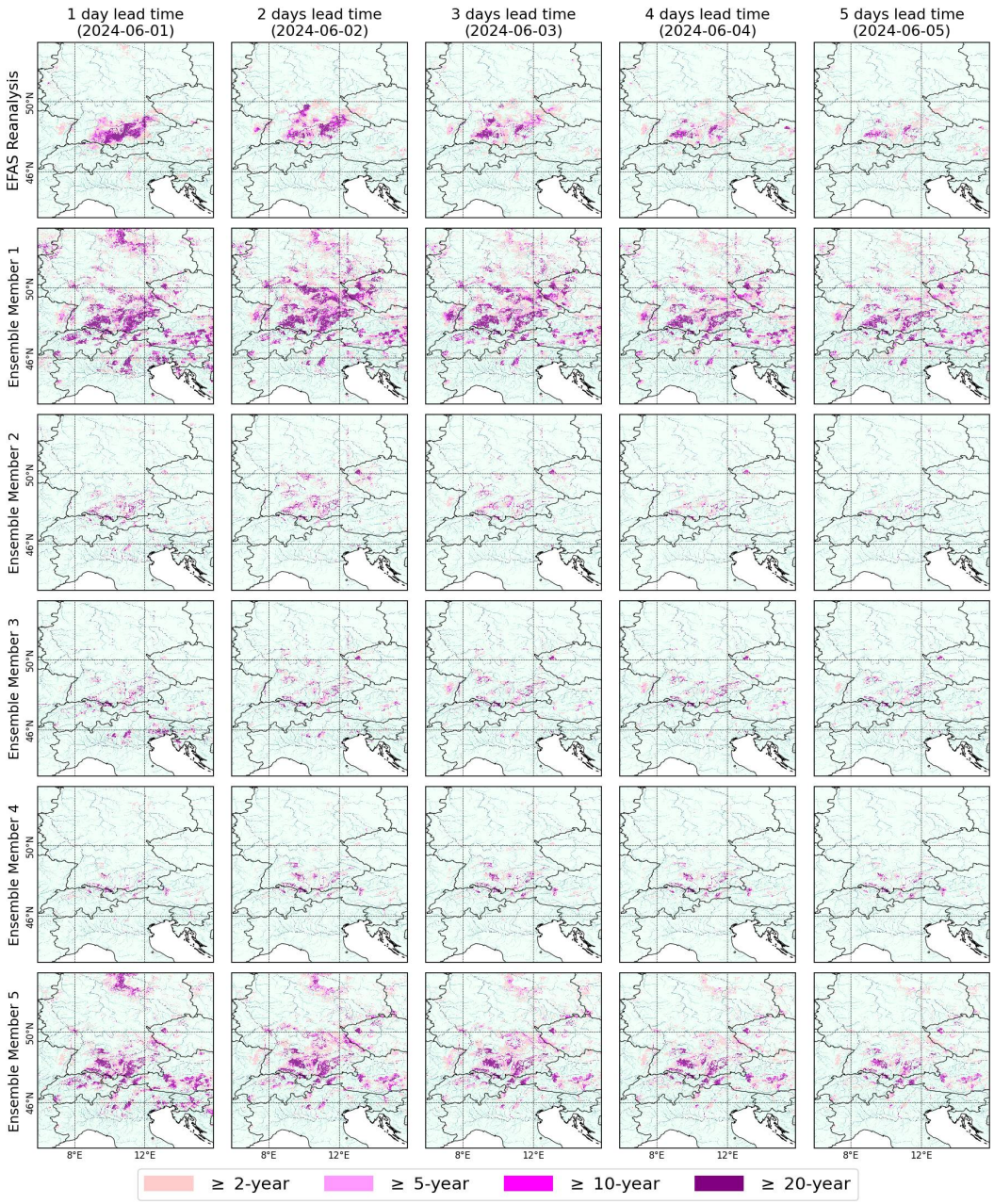


Figure 43: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Augsburg flood in Germany, June 2024. Shown are flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days.

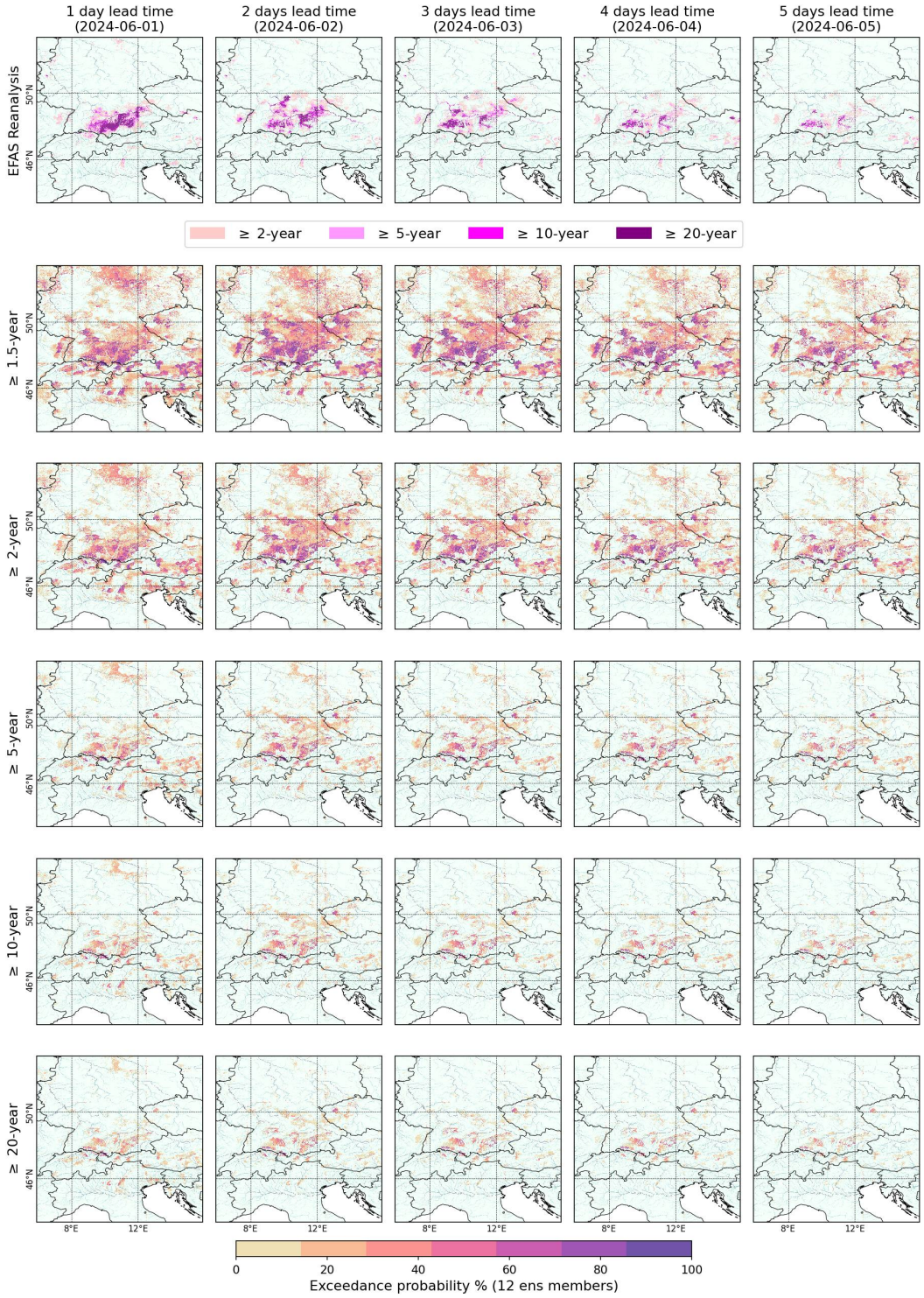


Figure 44: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Augsburg flood in Germany, June 2024. First row shows flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days. Rows 2 to 6 show the exceedance probabilities for flood return periods (1.5-20.0) computed from 12 ensemble members.

H.2 2024 CENTRAL EUROPEAN FLOODS

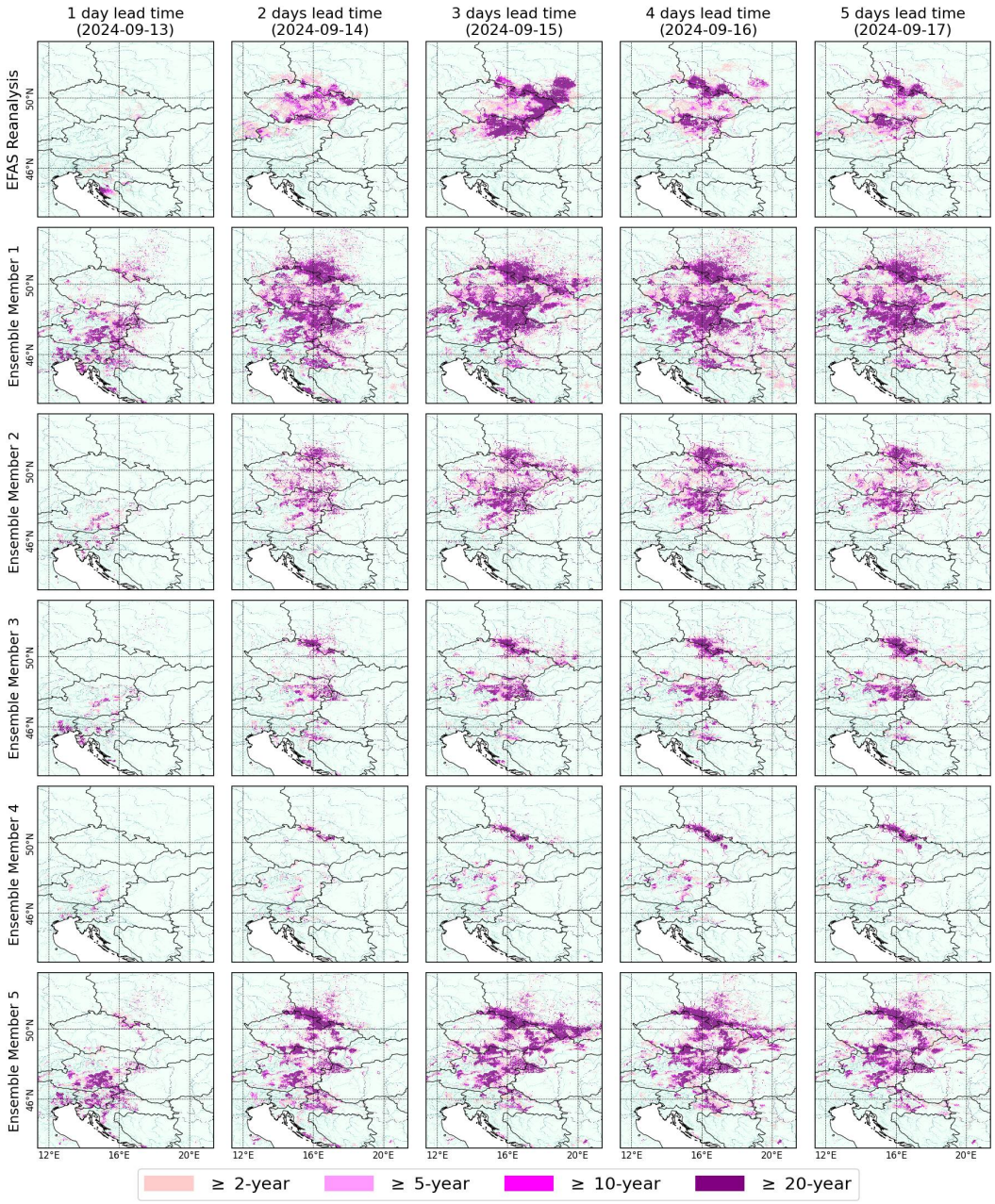


Figure 45: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Central European flood, September 2024. Shown are flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days.

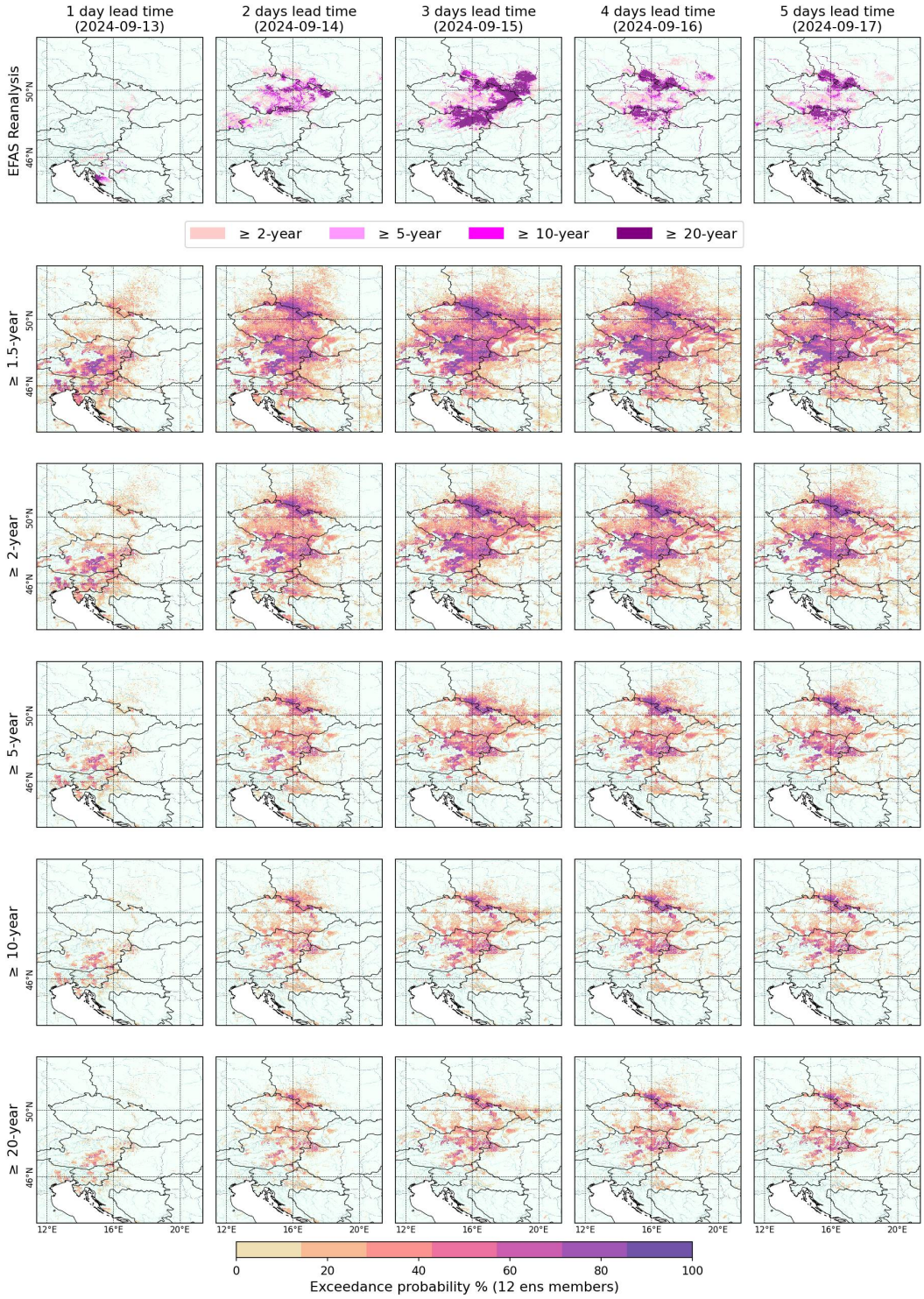


Figure 46: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Central European flood, September 2024. First row shows flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days. Rows 2 to 6 show the exceedance probabilities for flood return periods (1.5-20.0) computed from 12 ensemble members.

H.3 2024 SPANISH FLOODS

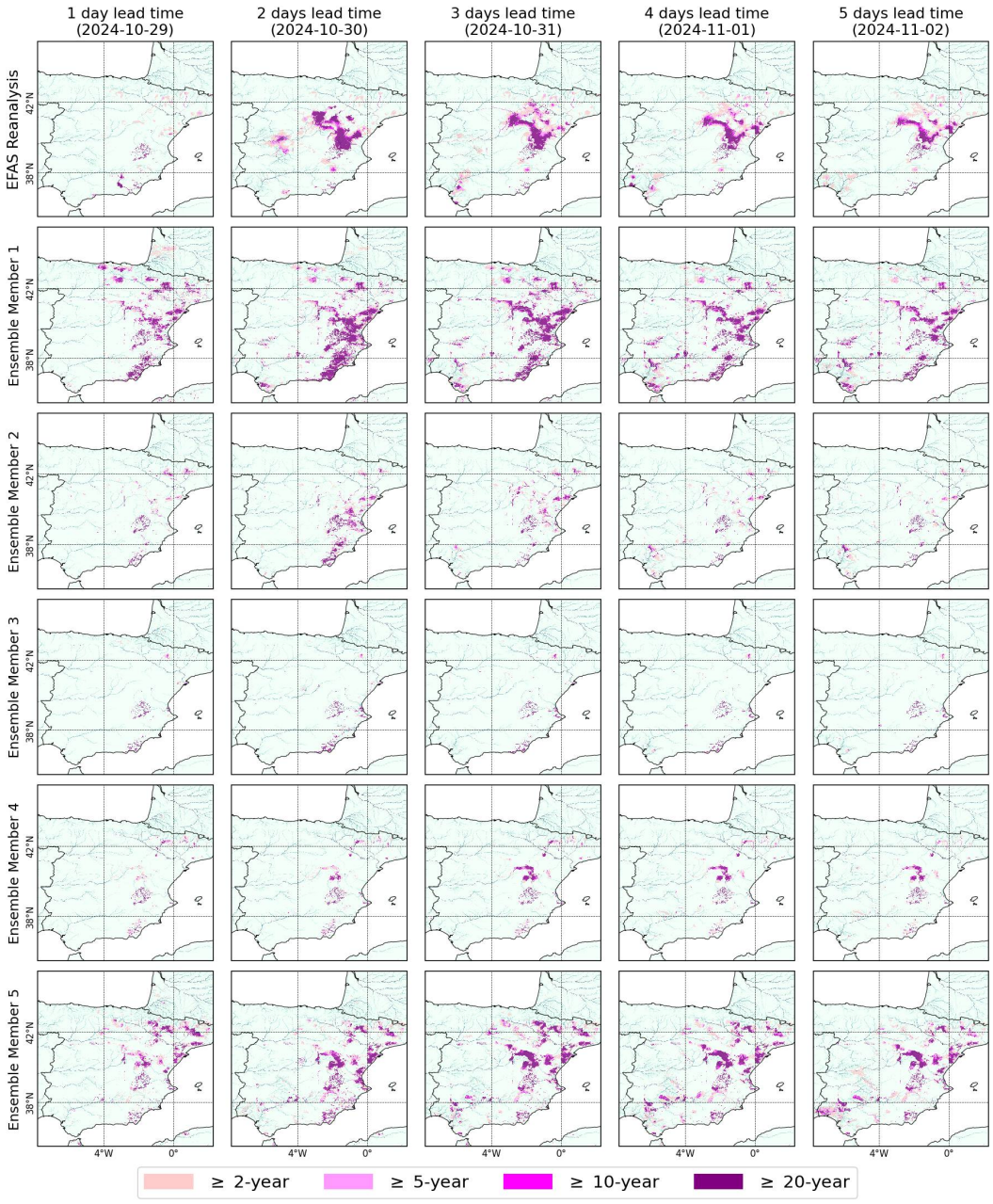


Figure 47: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Spanish flood, October 2024. Shown are flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days.

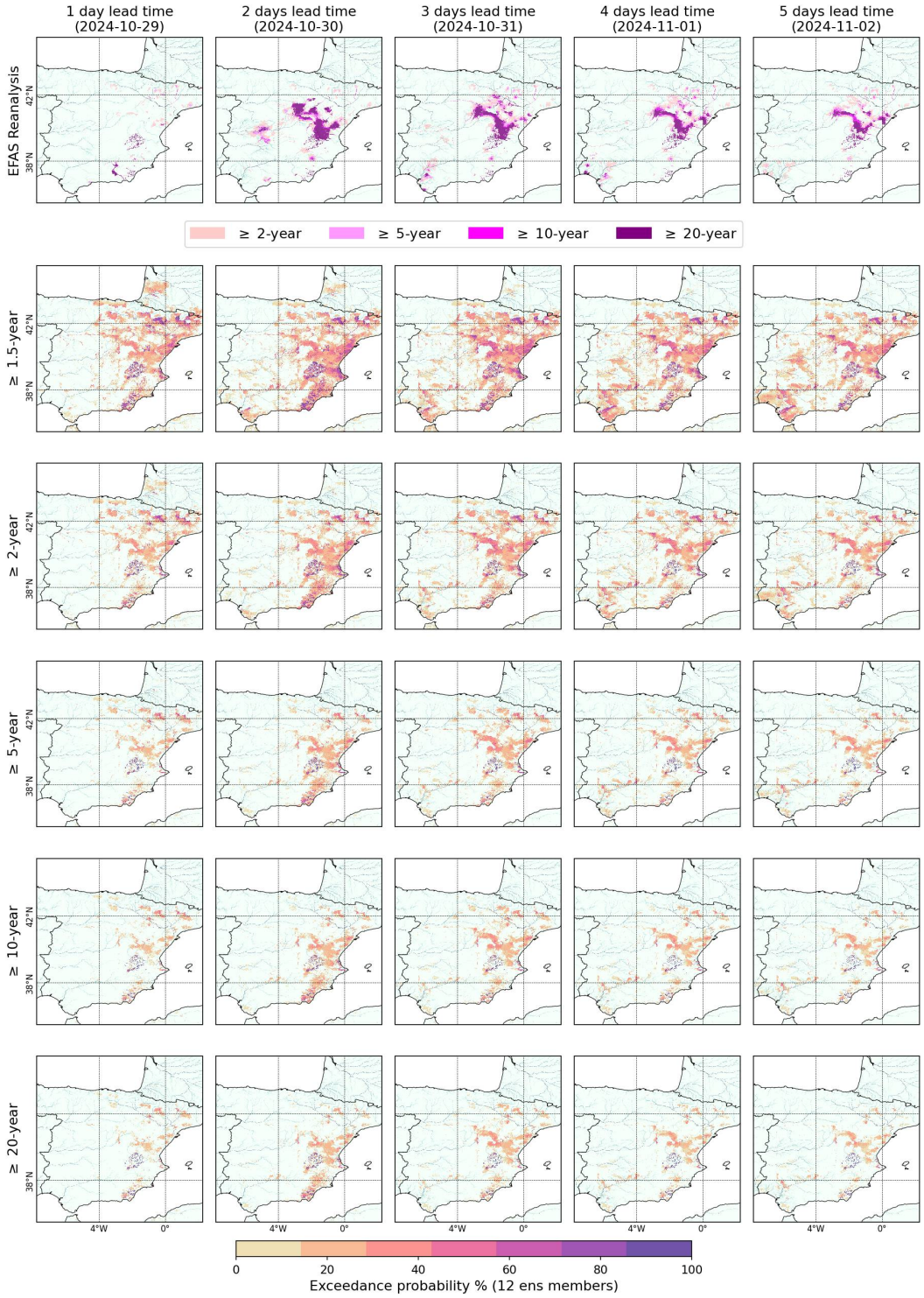


Figure 48: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Spanish flood, October 2024. First row shows flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days. Rows 2 to 6 show the exceedance probabilities for flood return periods (1.5-20.0) computed from 12 ensemble members.

H.4 2025 United Kingdom Floods

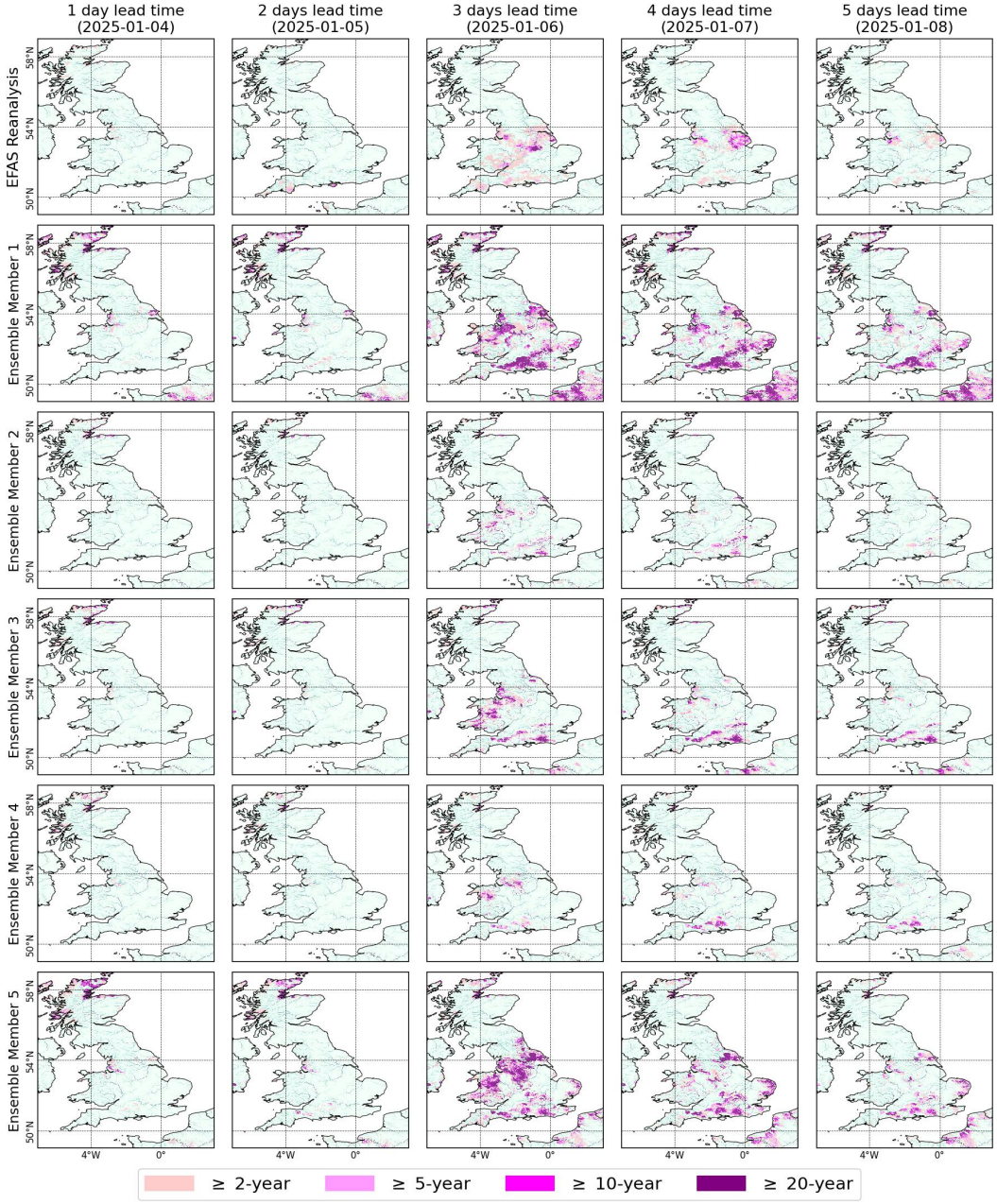


Figure 49: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the United Kingdom flood, January 2025. Shown are flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days.

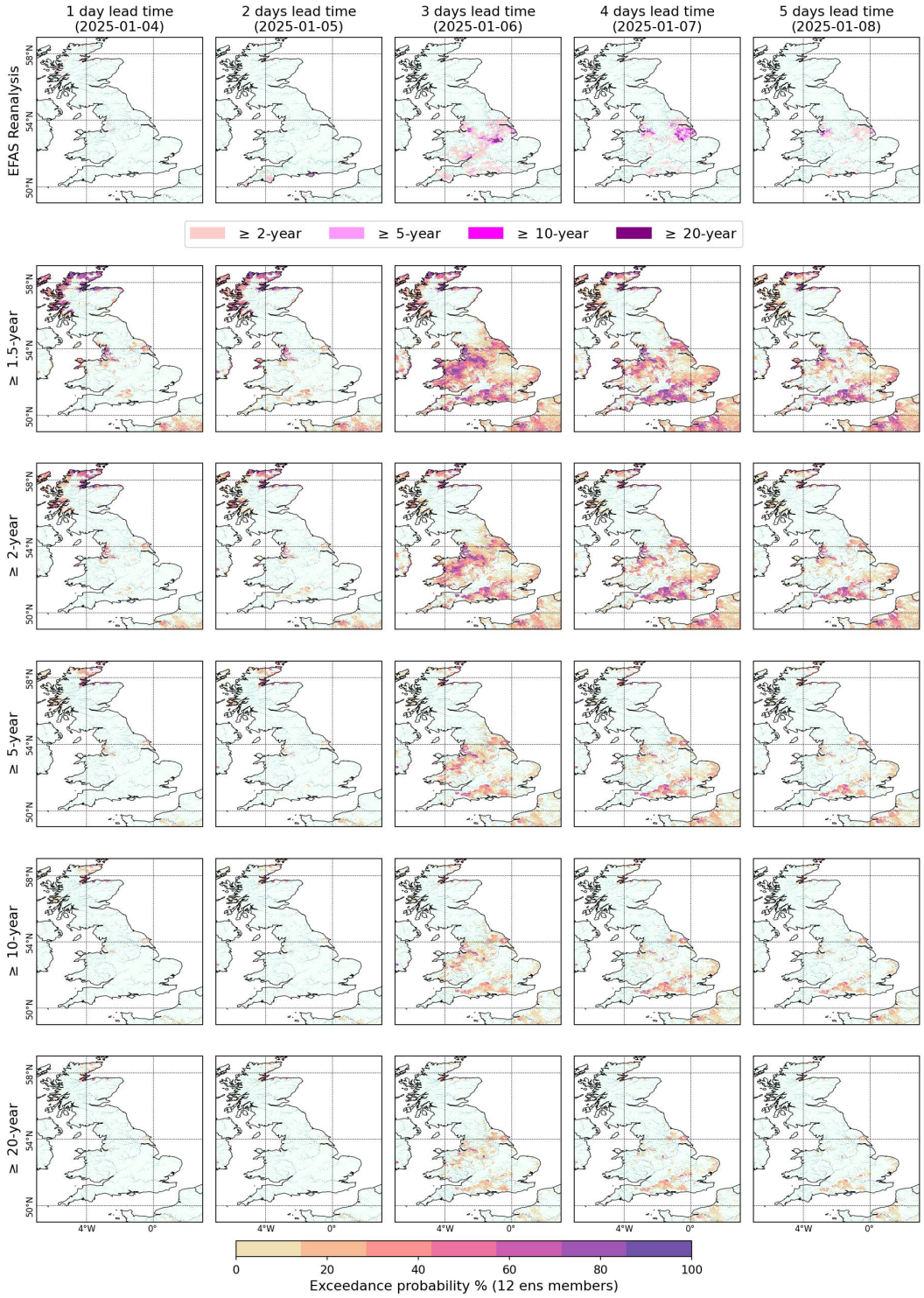


Figure 50: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the United Kingdom flood, January 2025. First row shows flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days. Rows 2 to 6 show the exceedance probabilities for flood return periods (1.5-20.0) computed from 12 ensemble members.

H.5 2025 Balkan Rivers Floods

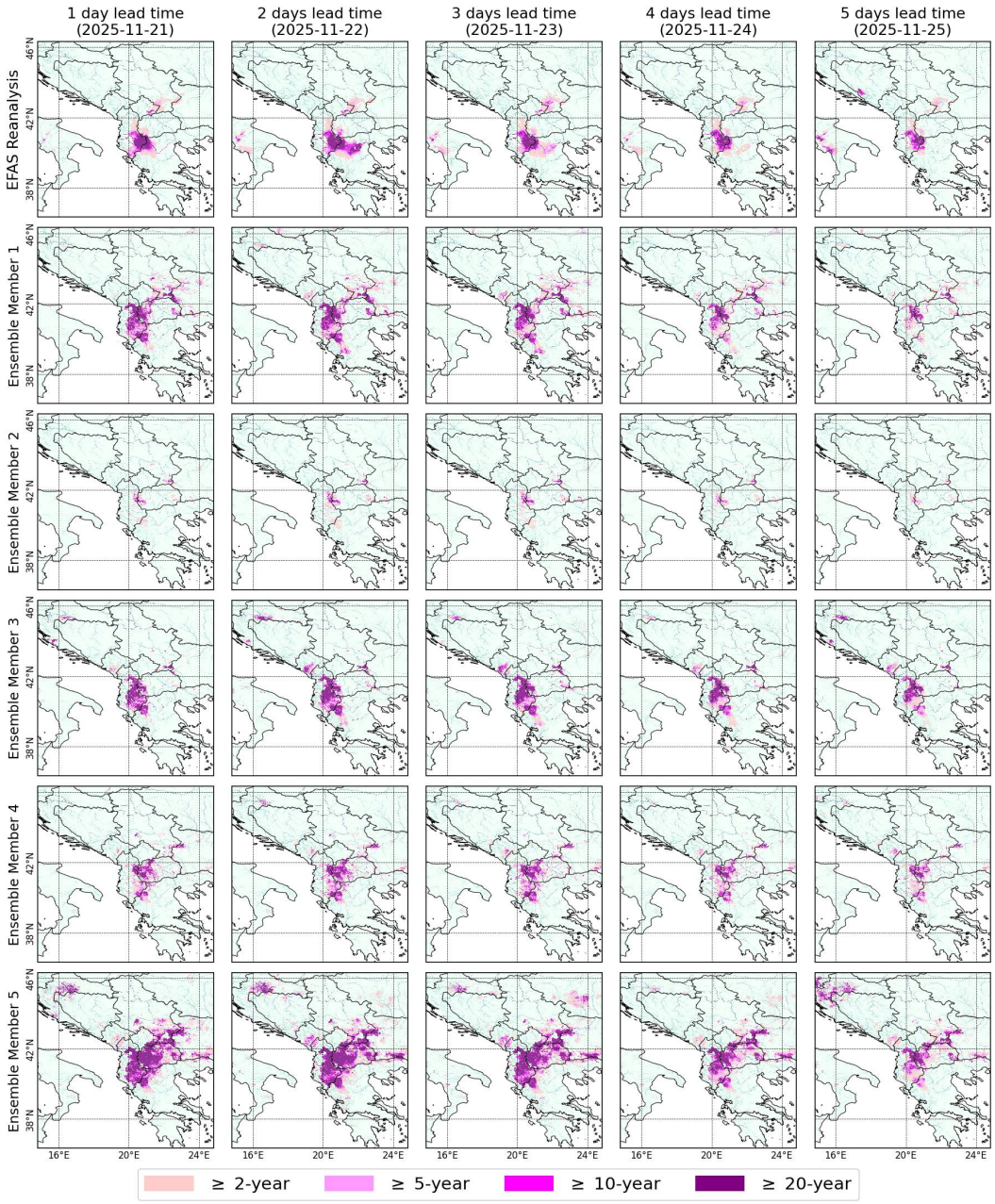


Figure 51: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Balkan Rivers flood, November 2025. Shown are flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days.

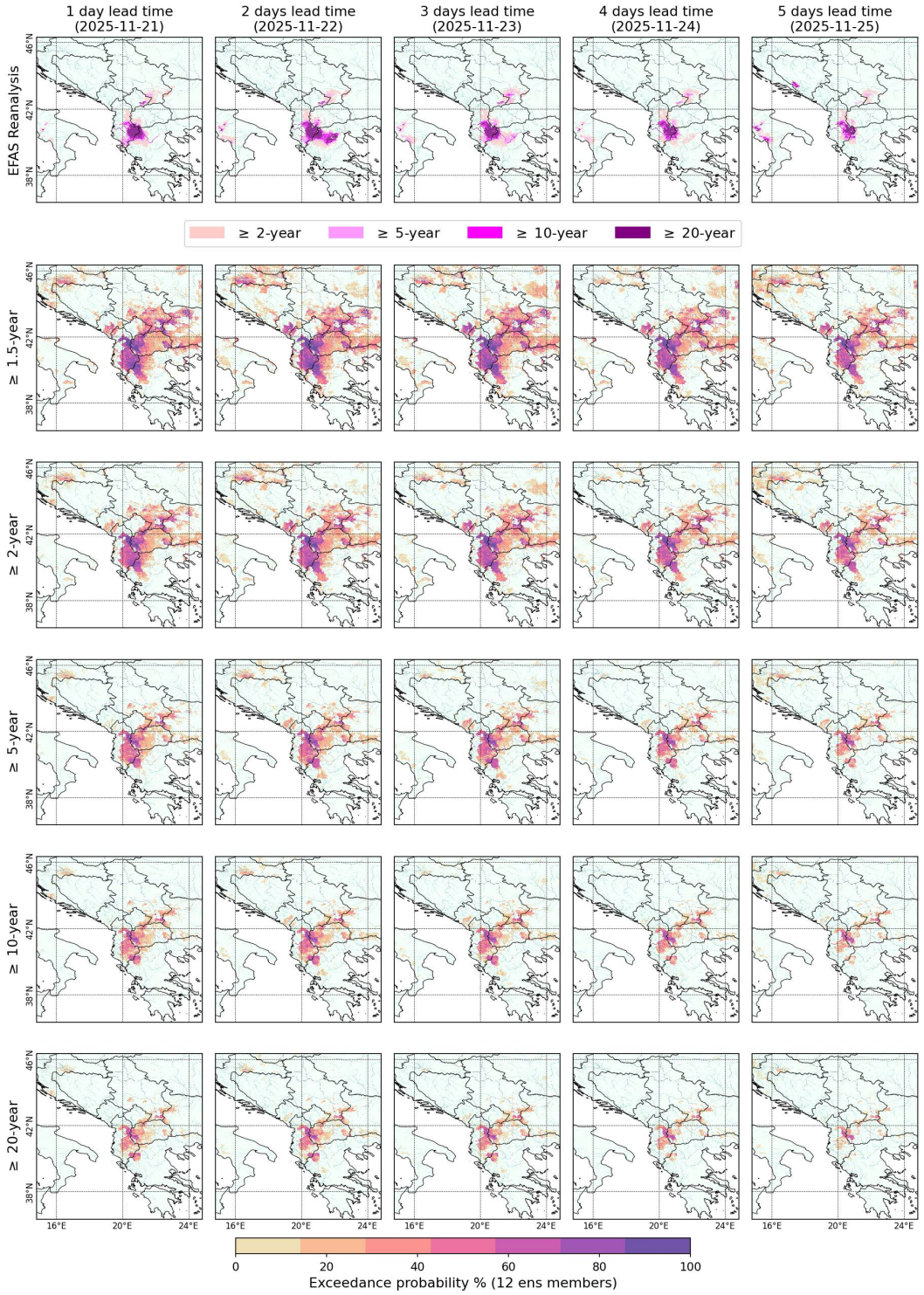


Figure 52: Comparison between EFAS reanalysis (first row, used as a reference) and our ensemble forecast (second to sixth rows) during the Balkan Rivers flood, November 2025. First row shows flood severity maps at 1-arcminute resolution where each panel shows flood extent at different lead times from 1 to 5 days. Rows 2 to 6 show the exceedance probabilities for flood return periods (1.5-20.0) computed from 12 ensemble members.

I NON-LINEAR RESERVOIR

In this experiment, we test whether the learned routing features behave as a non-linear reservoir i.e. an internal state that is carried forward and updated over time rather than a static representation per location. For this we inspect the internal state of the routing module in the forecasting block for random dates from the years 2022-2025.

The routing SSD block produces, at every grid point p, lead time l, and forecast time t, a routing feature vector $x(p)_l^t \in \mathbb{R}^{192}$ (the residual in Eq. 6). This is the state that the SSM propagates across the auto-regressive rollout. As a scalar proxy for the magnitude of this state, we use its L2-norm, $\bar{x}(p)_l^t = \|x(p)_l^t\|_2$.

In Fig. 53, we show how state evolves across independent events. For this we fix the lead time at 48 h and compare the routing state across independent forecast times. We selected four random times for the forecast 2022-07-23, 2023-02-24, 2024-05-27, and 2025-10-17. Each subplot shows the norm vector $\bar{x}^t(p)$. All four subplots share the same coarse spatial structure of the river network. This shows, that the learned state is spatially organized like a hydrological routing system and it is not static but varies with time.

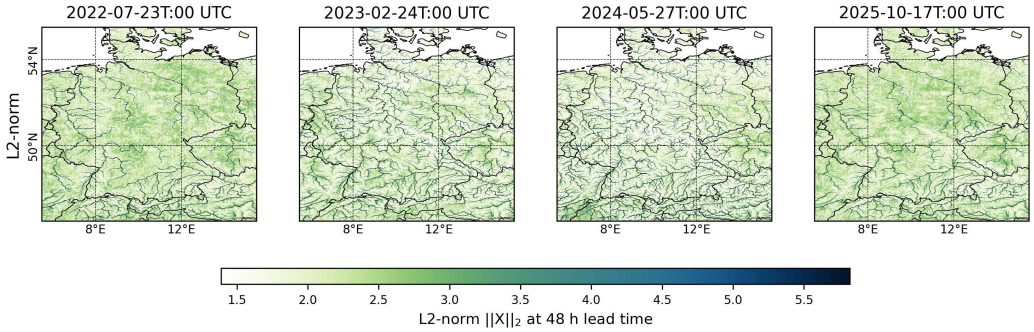


Figure 53: Routing state at a fixed 48 h lead time across four independent forecasts, one randomly sampled per year from 2022 to 2025. Top row shows the L2-norm of the 192-dimensional routing feature vector across the years.

J ADDITIONL RESULTS

J.1 EXPERIMENTS ON EFAS REANALYSIS

J.1.1 EXPERIMENTS ON EFAS REANALYSIS FOR CAMELS SPLIT

Table 24: Results on EFAS Reanalysis (CAMELS split). (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024-2025)		
	Model	RMSE (↓)	R2 (↑)	KGE (↑)	RMSE (↓)	R2 (↑)	KGE (↑)
Deterministic	Persistence	4.949	0.086	0.533	6.013	-0.118	0.456
	ED-LSTM	4.622±0.016	0.229±0.007	0.300±0.009	5.316±0.010	0.157±0.007	0.264±0.013
	ME-LSTM	4.830±0.034	0.181±0.011	0.286±0.018	5.584±0.042	0.095±0.014	0.235±0.020
	Ours	**3.972**±0.052	0.428±0.015	0.534±0.011	4.747±0.040	0.333±0.011	0.481±0.008
Probabilistic	Clima (median)	5.382	-0.005	-0.035	6.025	-0.068	-0.074
	DRUM (mean)	4.985	0.150	0.151	5.519	0.098	0.135
	ED-LSTM (mean)	4.782	0.173	0.243	5.367	0.135	0.233
	ME-LSTM (mean)	4.682	0.209	0.307	5.436	0.121	0.264
	Ours (mean)	3.991	**0.452**	**0.575**	**4.655**	**0.391**	**0.547**

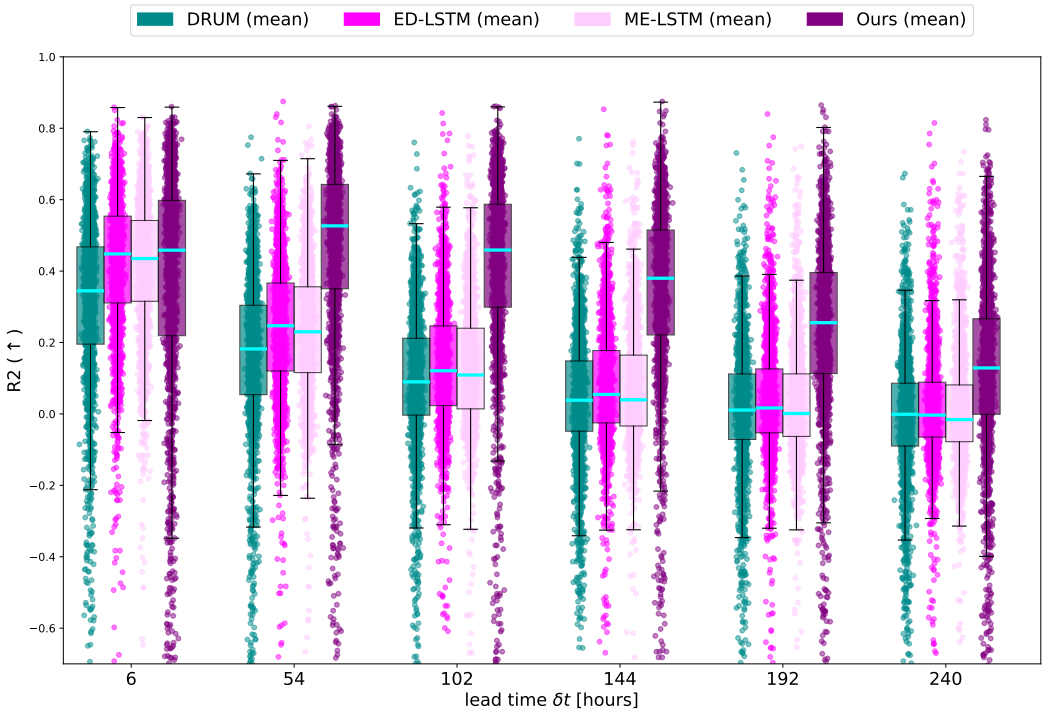


Figure 54: R2 of the river discharge forecasting with different lead time on reanalysis CAMELS split (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

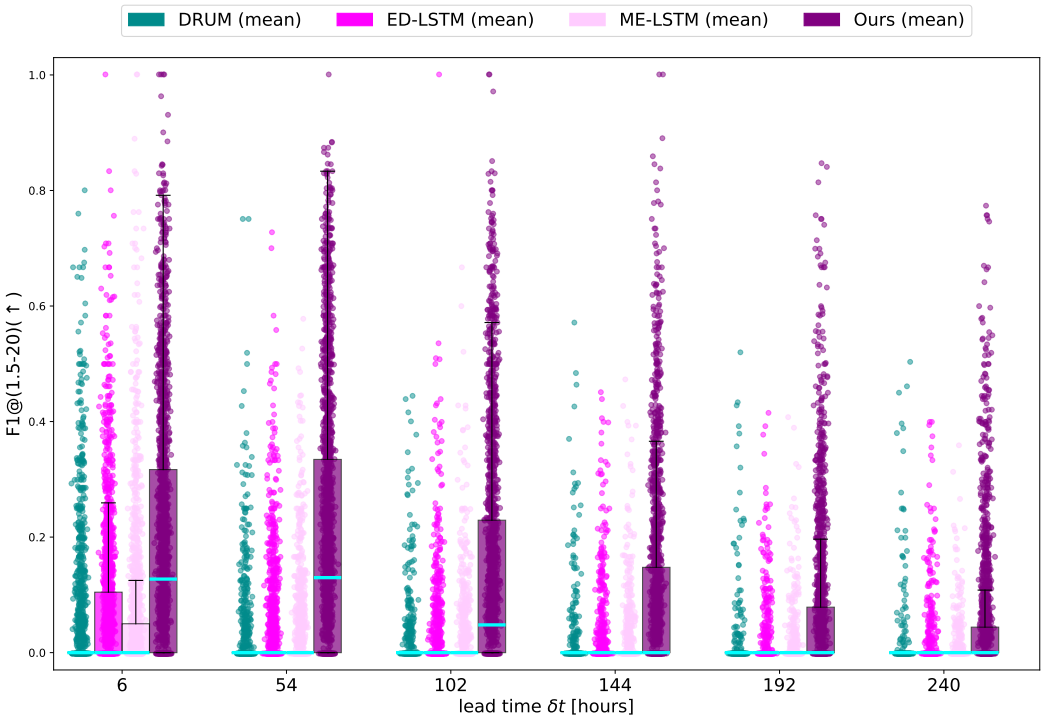


Figure 55: Mean F1@(1.5-20) of the river discharge forecasting with different lead time on reanalysis CAMELS split (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

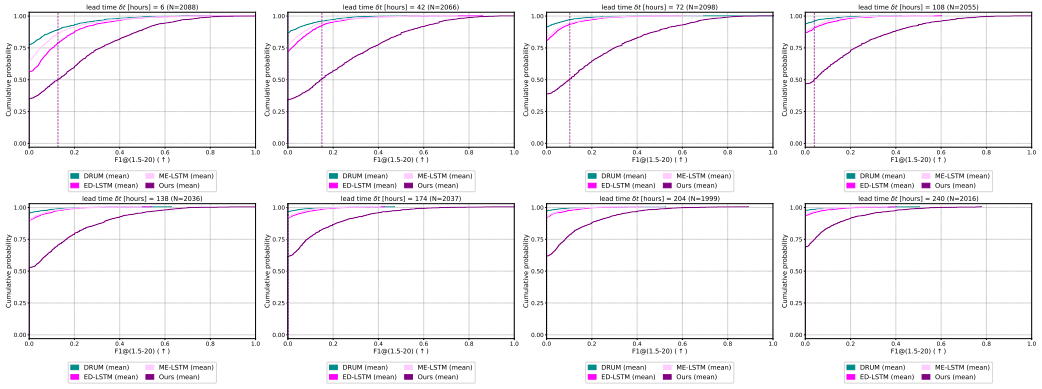


Figure 56: Cumulative distribution function (CDF) of station-level F1@(1.5-20) on the CAMELS reanalysis split (N=2206), across lead times. Dashed vertical lines shows each model's media.

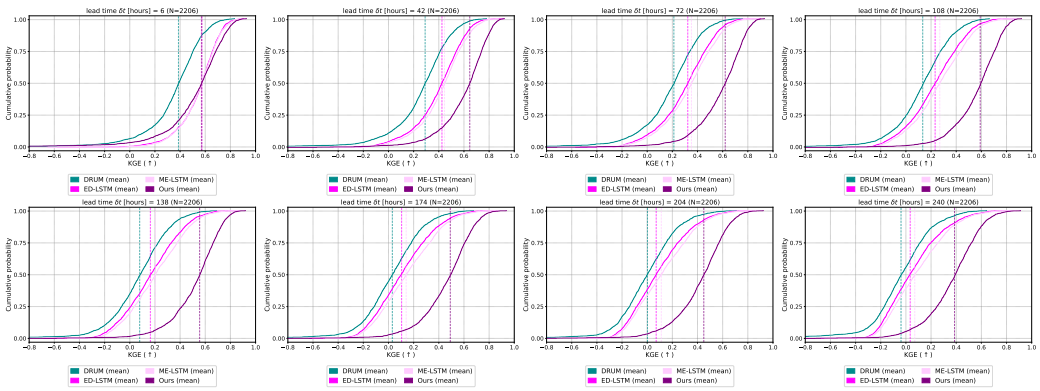


Figure 57: Cumulative distribution function (CDF) of station-level KGE on the CAMELS reanalysis split (N=2206), across lead times. Dashed vertical lines shows each model's media.

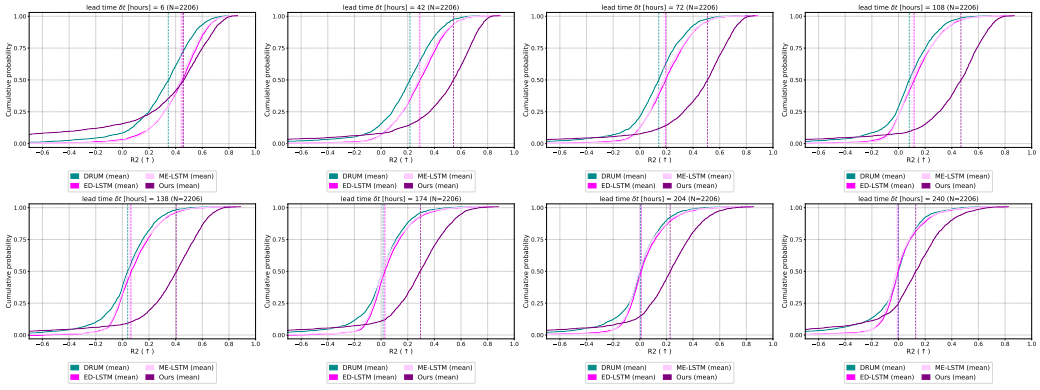


Figure 58: Cumulative distribution function (CDF) of station-level R2 on the CAMELS reanalysis split (N=2206), across lead times. Dashed vertical lines shows each model's media.

J.1.2 Experiments on EFAS Reanalysis for GRDC split

Table 25: Results on EFAS Reanalysis (GRDC split). (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024-2025)		
	Model	RMSE (↓)	R2 (↑)	KGE (↑)	RMSE (↓)	R2 (↑)	KGE (↑)
Deterministic	Persistence	10.417	0.096	0.542	12.056	-0.076	0.468
	ED-LSTM	9.739±0.065	0.239±0.009	0.318±0.009	10.975±0.005	0.151±0.004	0.269±0.013
	ME-LSTM	10.242±0.067	0.186±0.010	0.321±0.020	11.515±0.062	0.085±0.013	0.249±0.021
	Ours	8.494±0.138	0.420±0.019	0.531±0.014	9.851±0.074	0.314±0.010	0.474±0.006
Probabilistic	Clima (median)	11.014	0.002	-0.019	12.169	-0.062	-0.058
	DRUM (mean)	10.461	0.148	0.169	11.495	0.095	0.148
	ED-LSTM (mean)	10.092	0.178	0.255	11.218	0.125	0.237
	ME-LSTM (mean)	9.882	0.211	0.331	11.098	0.114	0.278
	Ours (mean)	**8.375**	**0.441**	**0.567**	**9.569**	**0.360**	**0.530**

Table 26: Results for probabilistic models on EFAS Reanalysis (GRDC split).

Results on the Validation set (2022-2023)					
	Clima	DRUM	ED-LSTM	ME-LSTM	Ours
CRPS (↓)	4.3328	4.5104	4.1379	3.9776	**3.1837**
Results on the Test set (2024-2025)					
	Clima	DRUM	ED-LSTM	ME-LSTM	Ours
CRPS (↓)	4.6667	4.9376	4.6384	4.6492	**3.7604**

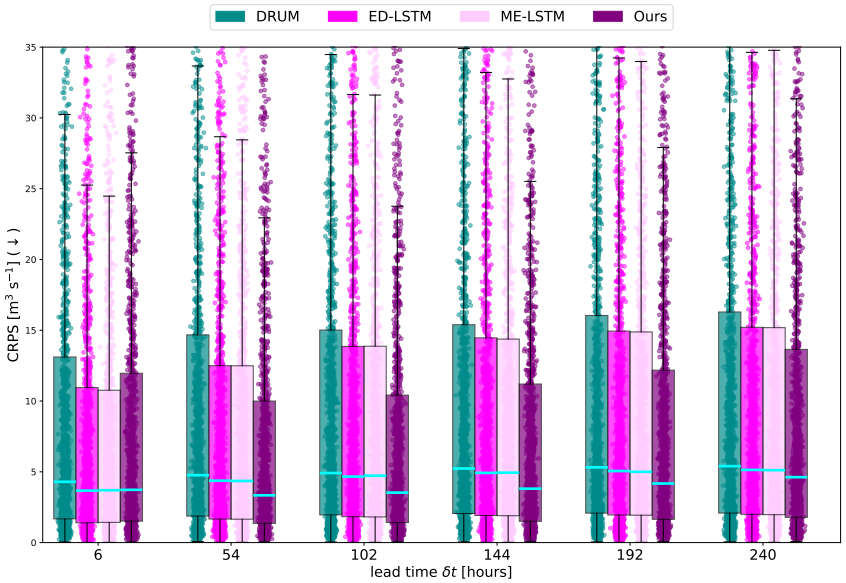


Figure 59: CRPS of the river discharge forecasting with different lead time on reanalysis GRDC split (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

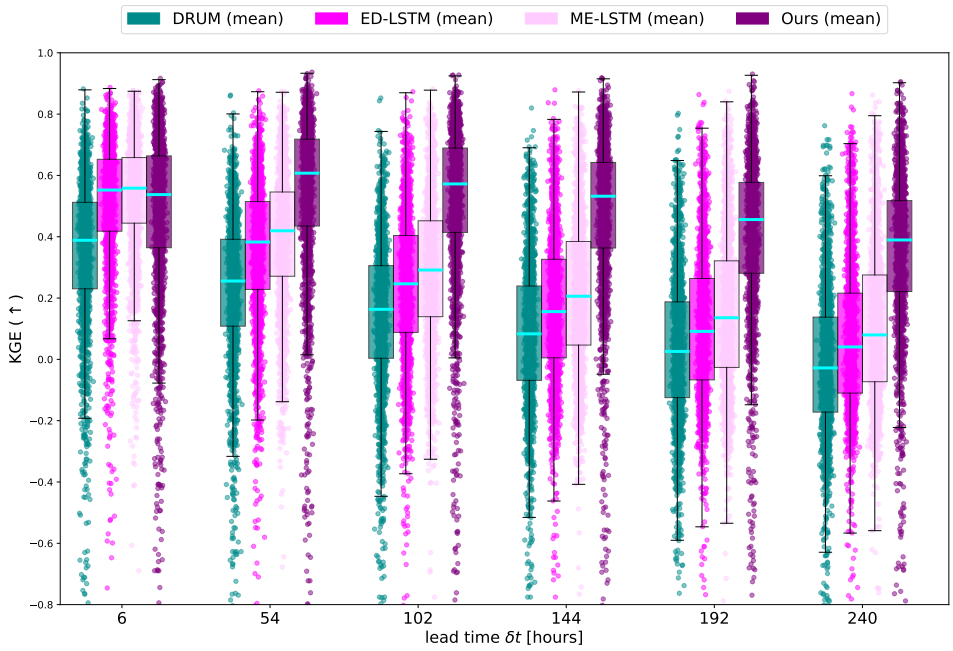


Figure 60: KGE of the river discharge forecasting with different lead time on reanalysis GRDC split (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

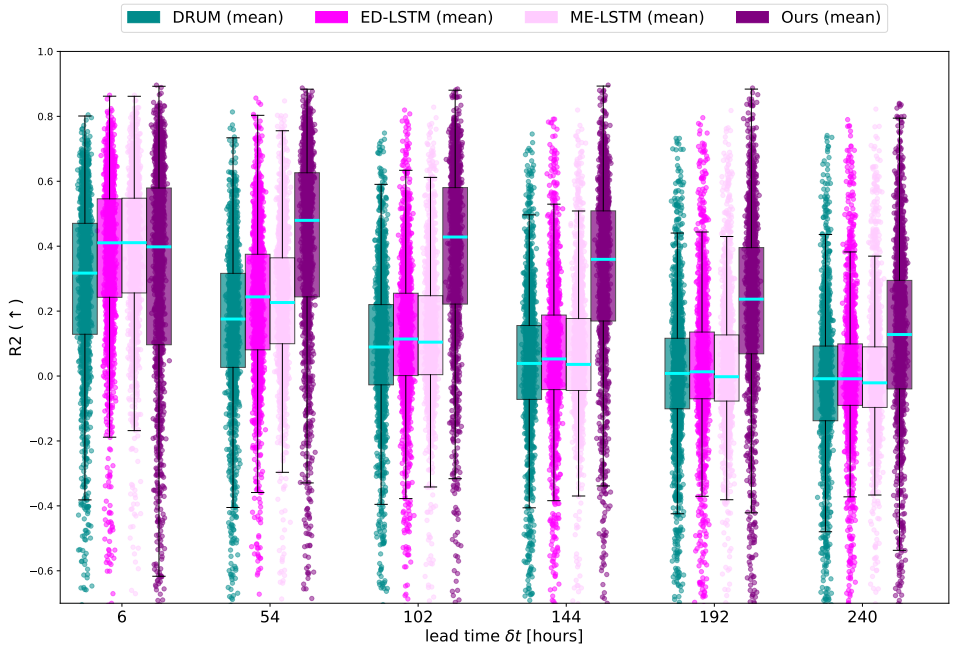


Figure 61: R2 of the river discharge forecasting with different lead time on reanalysis GRDC split (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

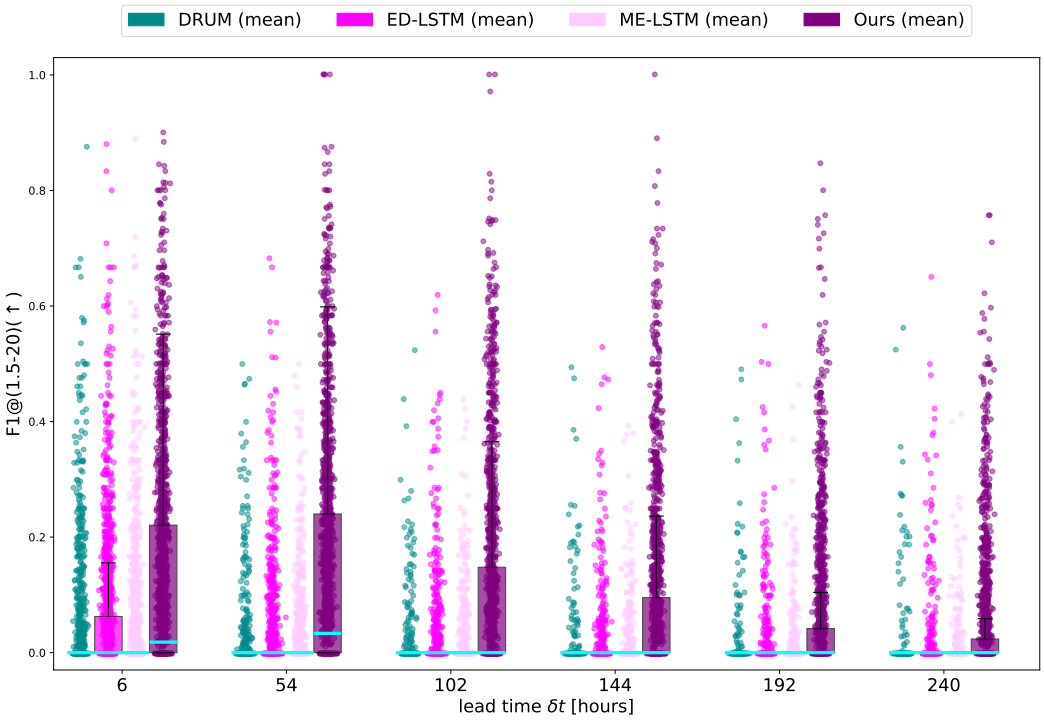


Figure 62: Mean F1@(1.5-20) of the river discharge forecasting with different lead time on reanalysis GDRC split (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

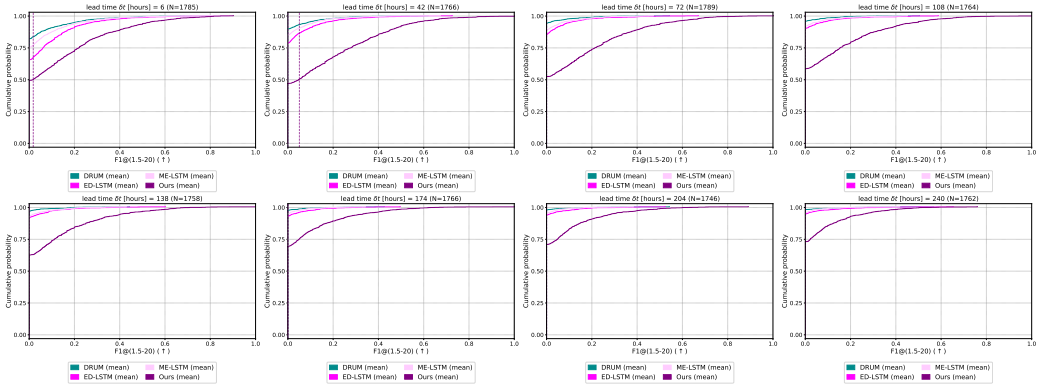


Figure 63: Cumulative distribution function (CDF) of station-level F1@(1.5-20) on the GRDC reanalysis split (N=1971), across lead times. Dashed vertical lines shows each model's media.

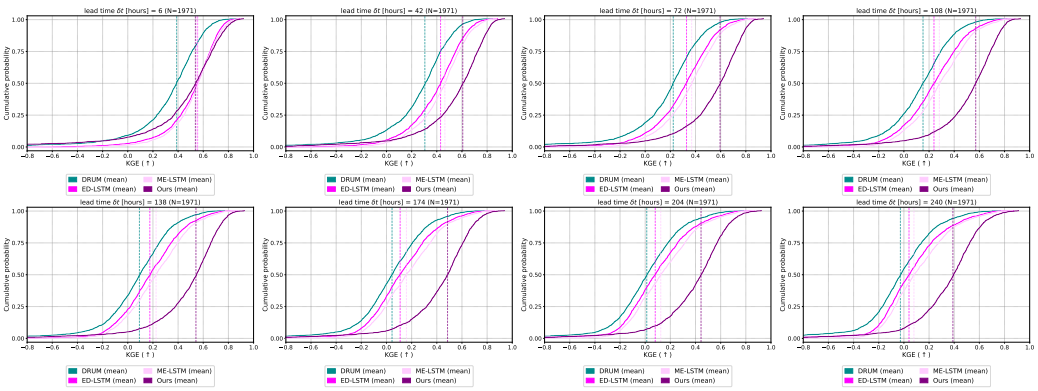


Figure 64: Cumulative distribution function (CDF) of station-level KGE on the GRDC reanalysis split (N=1971), across lead times. Dashed vertical lines shows each model's media.

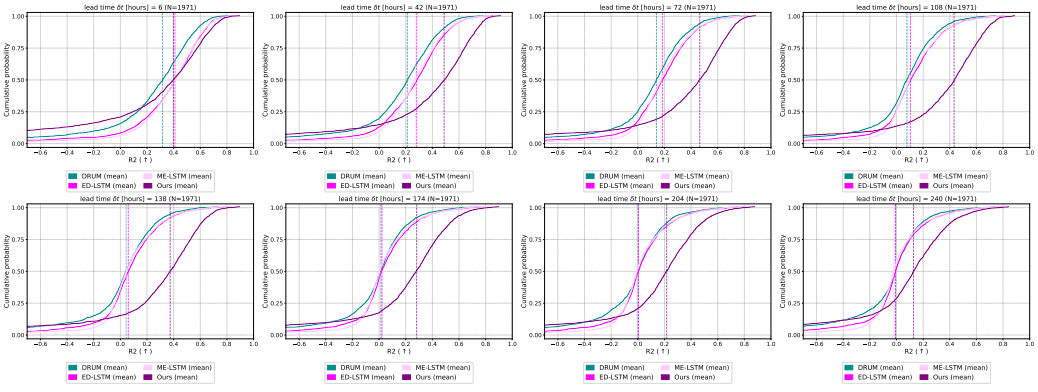


Figure 65: Cumulative distribution function (CDF) of station-level R2 on the GRDC reanalysis split (N=1971), across lead times. Dashed vertical lines shows each model's media.

J.2 Experiments on GRDC Observations

Table 27: Results on GRDC observations. (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024-2025)		
	Model	RMSE (↓)	R2 (↑)	KGE (↑)	RMSE (↓)	R2 (↑)	KGE (↑)
Deterministic	ED-LSTM	6.0542±0.0082	0.2705±0.0012	0.3491±0.0214	11.3616±0.1632	0.2366±0.0164	0.4501±0.0097
	ME-LSTM	5.8810±0.0206	0.2993±0.0044	0.3550±0.0083	11,4754±0.1683	0.2384±0.0246	0.4330±0.0138
	Ours	4.6795±0.0681	0.5603±0.0095	0.5890±0.0038	9.0163±0.0952	0.4576±0.0066	0.5839±0.0076
Probabilistic	DRUM (mean)	7.0236	0.1514	0.2177	12.6358	0.1153	0.3059
	ED-LSTM (mean)	6.4217	0.2004	0.2780	11.5629	0.2308	0.4266
	ME-LSTM (mean)	6.0267	0.2682	0.3921	11.6064	0.2084	0.4478
	Ours (mean)	**4.3823**	**0.6343**	**0.7021**	**4.7963**	**0.6174**	**0.6893**

Table 28: Results for probabilistic models on GRDC observations.

Results on the Validation set (2022-2023)				
	DRUM	ED-LSTM	ME-LSTM	Ours
CRPS (↓)	4.5375	3.9024	3.4928	**2.4592**
Results on the Test set (2024-2025)				
	DRUM	ED-LSTM	ME-LSTM	Ours
CRPS (↓)	5.6934	5.0091	5.0361	**3.5194**

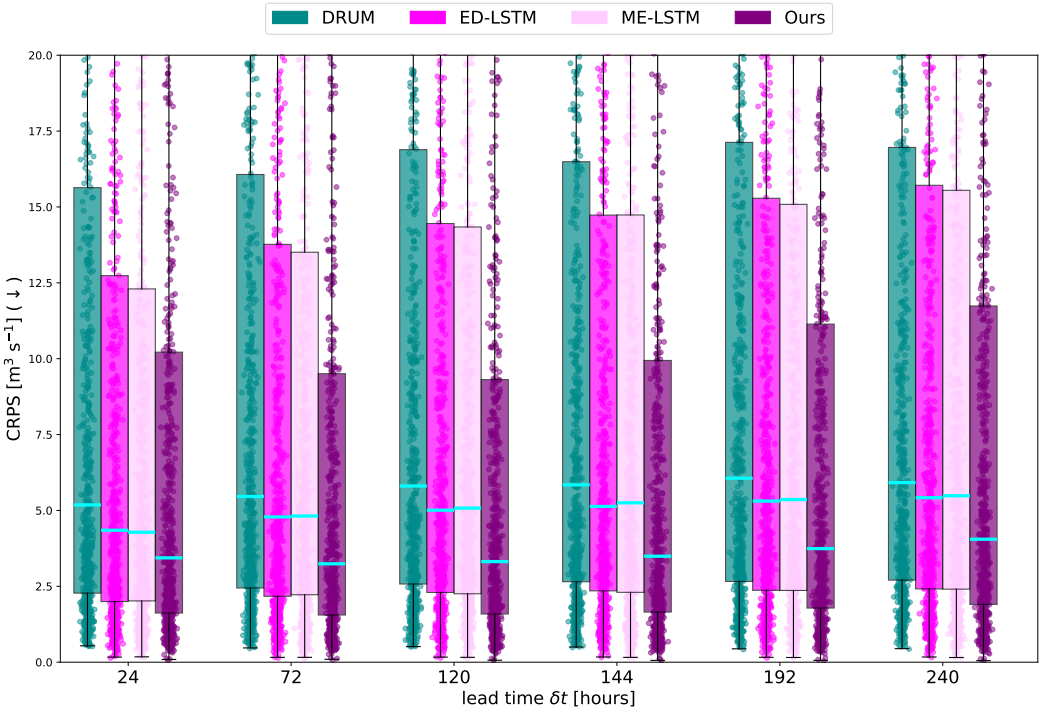


Figure 66: CRPS of the river discharge forecasting with different lead time on GRDC observations (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

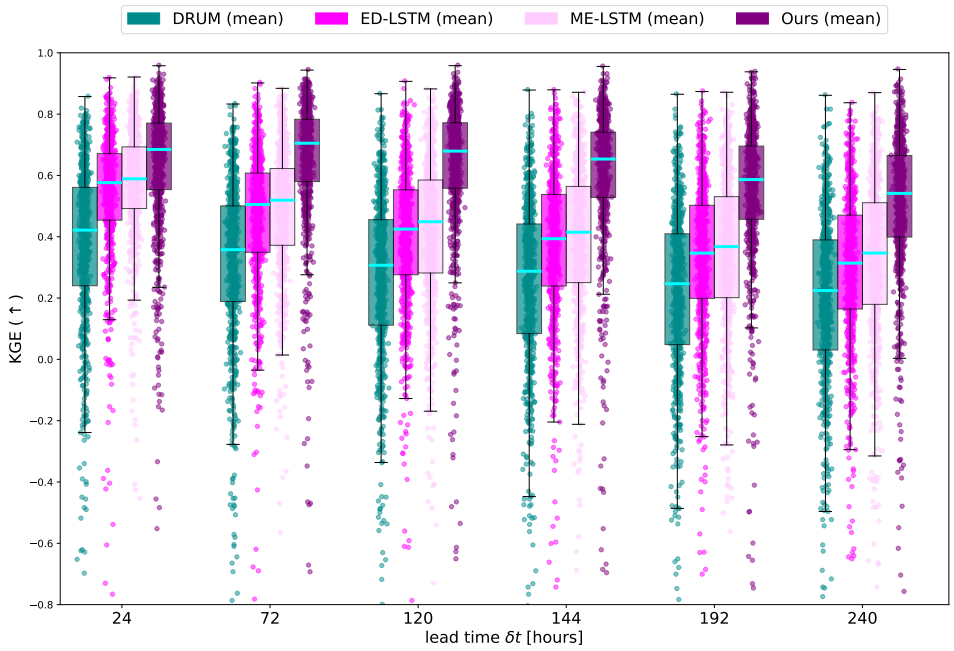


Figure 67: KGE of the river discharge forecasting with different lead time on GRDC observations (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

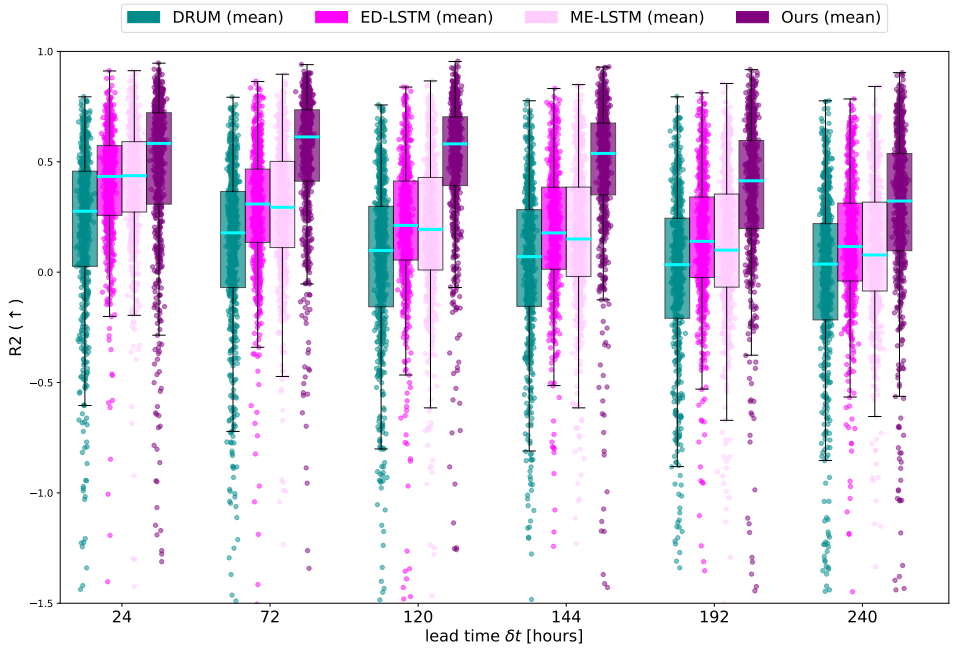


Figure 68: R2 of the river discharge forecasting with different lead time on GRDC observations (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

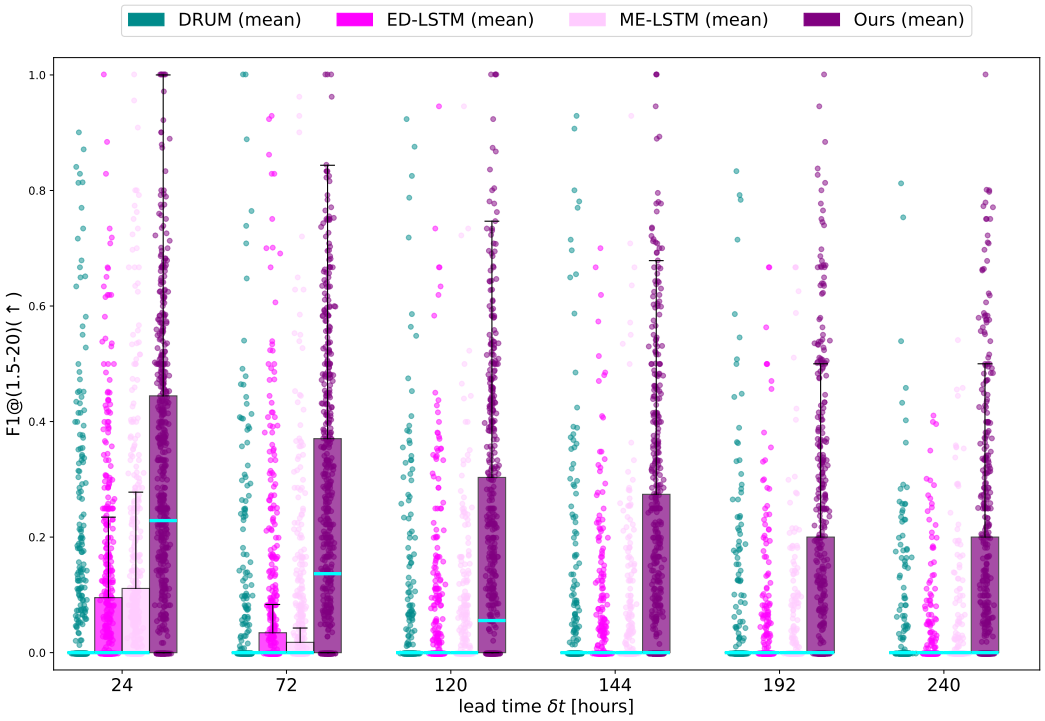


Figure 69: mean F1@(1.5-20) of the river discharge forecasting with different lead time on GRDC observations (test set 2024-2025, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

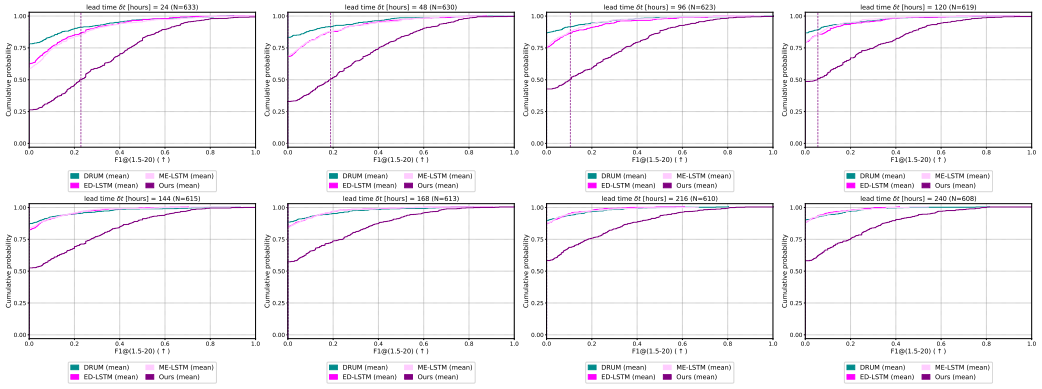


Figure 70: Cumulative distribution function (CDF) of station-level F1@(1.5-20) on GRDC observations (N=758), across lead times. Dashed vertical lines shows each model's media.

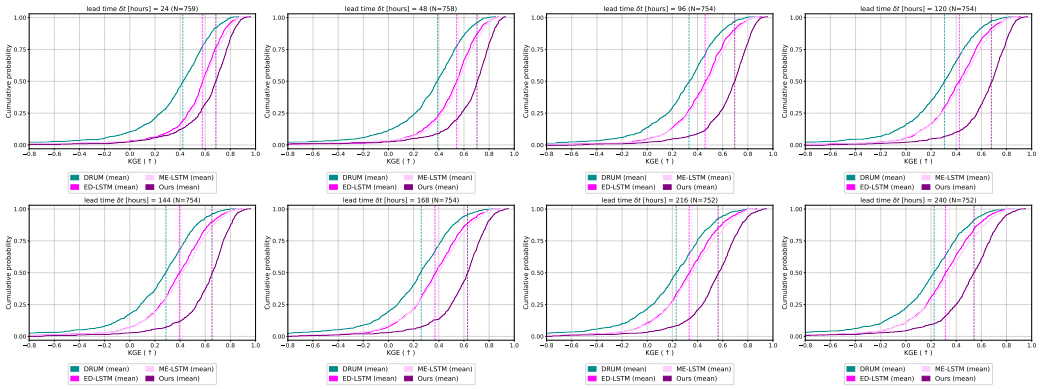


Figure 71: Cumulative distribution function (CDF) of station-level KGE on GRDC observations (N=758), across lead times. Dashed vertical lines shows each model's media.

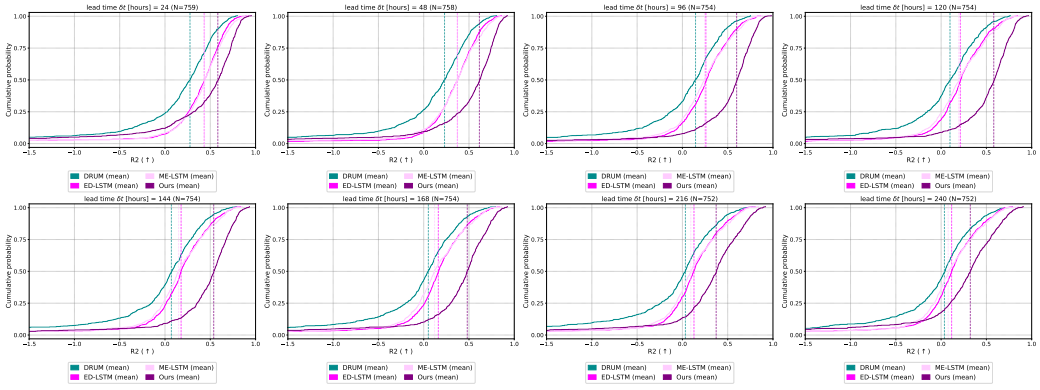


Figure 72: Cumulative distribution function (CDF) of station-level R2 on GRDC observations (N=758), across lead times. Dashed vertical lines shows each model's media.

J.3 Experiments on CAMELS Observations

Table 29: Results on CAMELS observations. (±) denotes the standard deviation for 3 runs.

		Validation (2022-2023)			Test (2024)		
	Model	RMSE (↓)	R2 (↑)	KGE (↑)	RMSE (↓)	R2 (↑)	KGE (↑)
Deterministic	ED-LSTM	6.0542±0.0082	0.2705±0.0012	0.3401±0.0214	6.8658±0.1326	0.0544±0.0072	0.0906±0.0136
	ME-LSTM	5.8810±0.0206	0.2993±0.0044	0.3550±0.0083	6.9621±0.0874	0.0500±0.0100	0.1186±0.0123
	Ours	4.6795±0.0681	0.5603±0.0095	0.5890±0.0038	5.8406±0.1311	0.3485±0.0202	0.4162±0.0284
Probabilistic	DRUM (mean)	7.0236	0.1514	0.2177	7.3433	0.0115	0.0576
	ED-LSTM (mean)	6.4217	0.2004	0.2780	7.0896	0.0452	0.0831
	ME-LSTM (mean)	6.0267	0.2682	0.3921	6.9401	0.0305	0.1046
	Ours (mean)	**4.3823**	**0.6343**	**0.7021**	**5.5031**	**0.4222**	**0.5143**

Table 30: Results for probabilistic models on CAMELS observations.

Results on the Validation set (2022-2023)				
	DRUM	ED-LSTM	ME-LSTM	Ours
CRPS (↓)	2.7239	2.3110	2.1237	**1.4840**
Results on the Test set (2024)				
	DRUM	ED-LSTM	ME-LSTM	Ours
CRPS (↓)	2.3685	2.2683	2.2683	**1.5678**

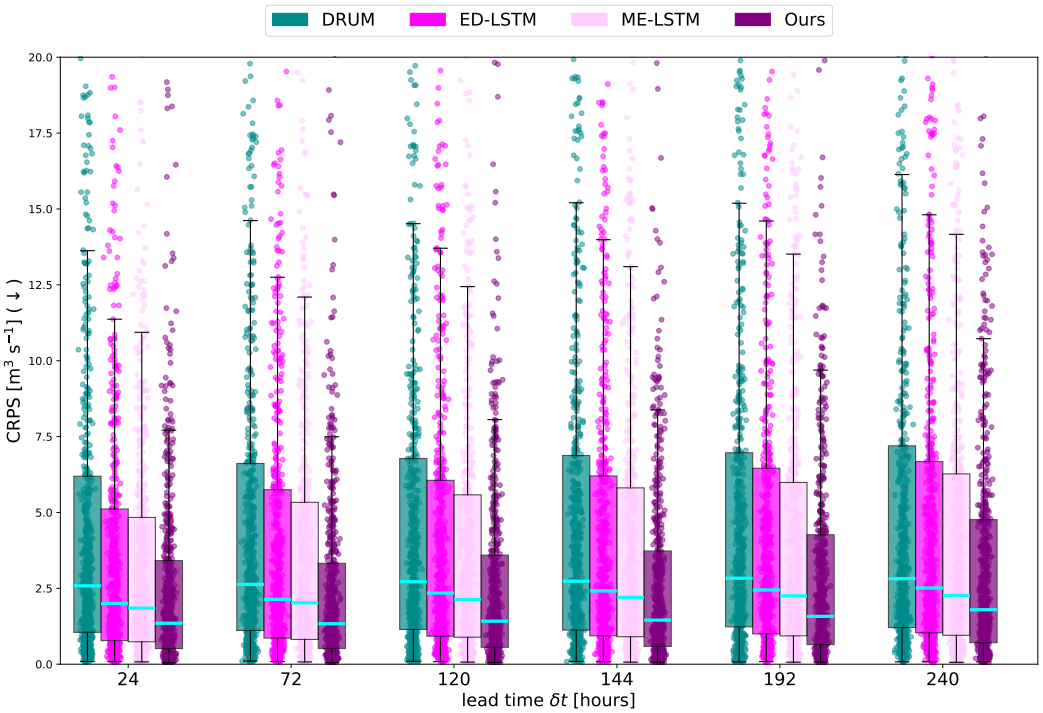


Figure 73: CRPS of the river discharge forecasting with different lead time on CAMELS observations (val set 2022-2023, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

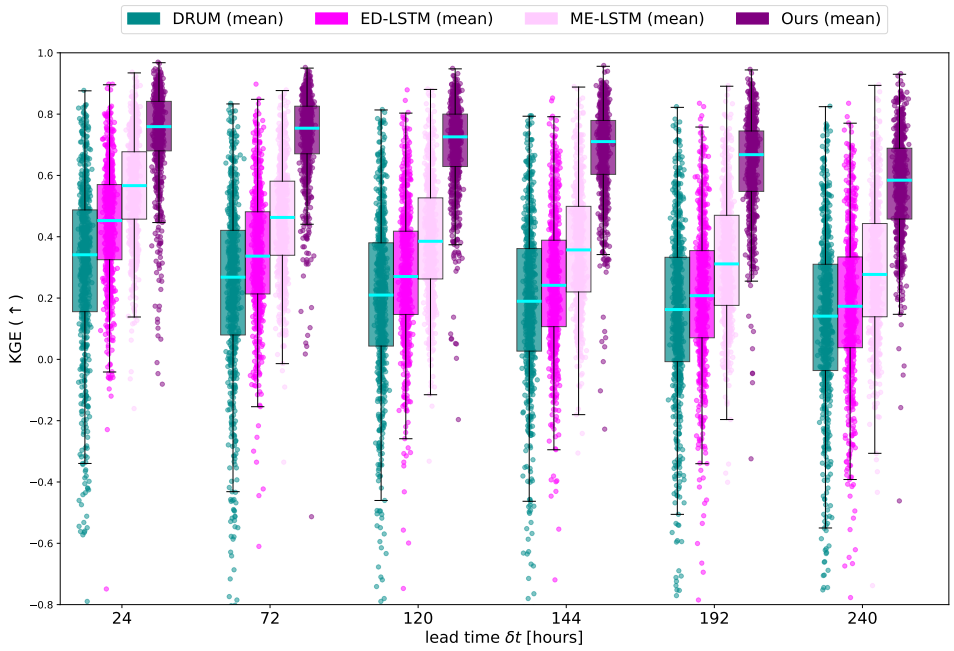


Figure 74: KGE of the river discharge forecasting with different lead time on CAMELS observations (val set 2022-2023, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

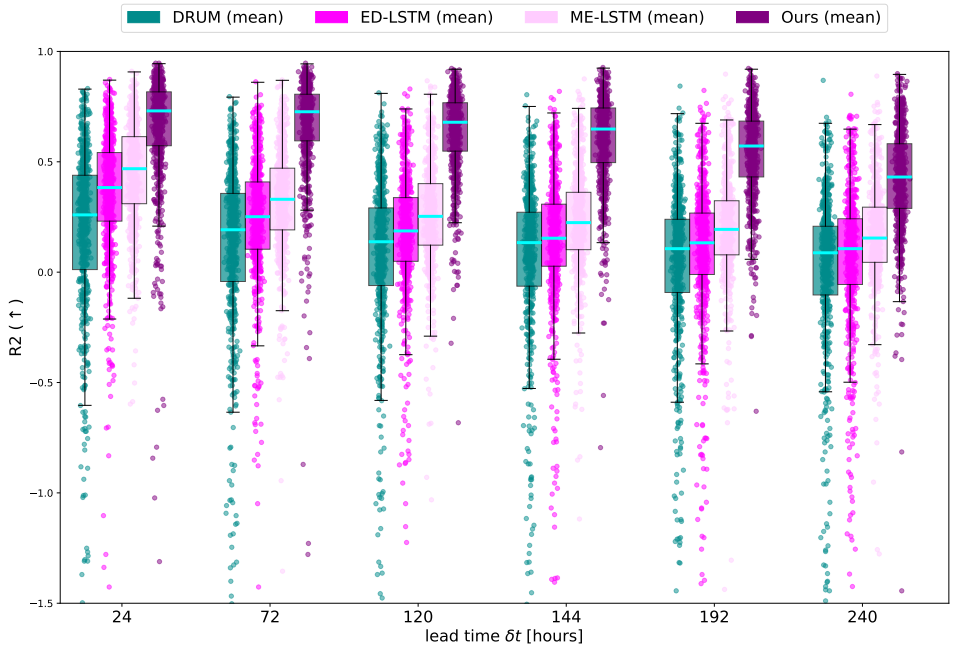


Figure 75: R2 of the river discharge forecasting with different lead time on CAMELS observations (val set 2022-2023, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

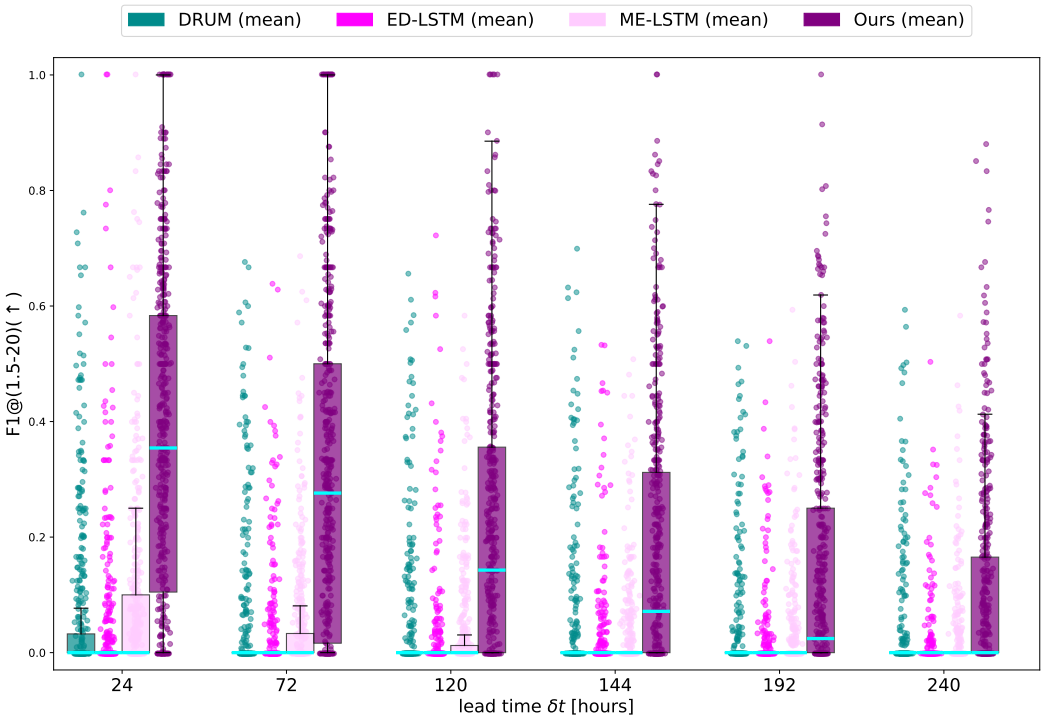


Figure 76: Mean F1@(1.5-20) of the river discharge forecasting with different lead time on CAMELS observations (val set 2022-2023, temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

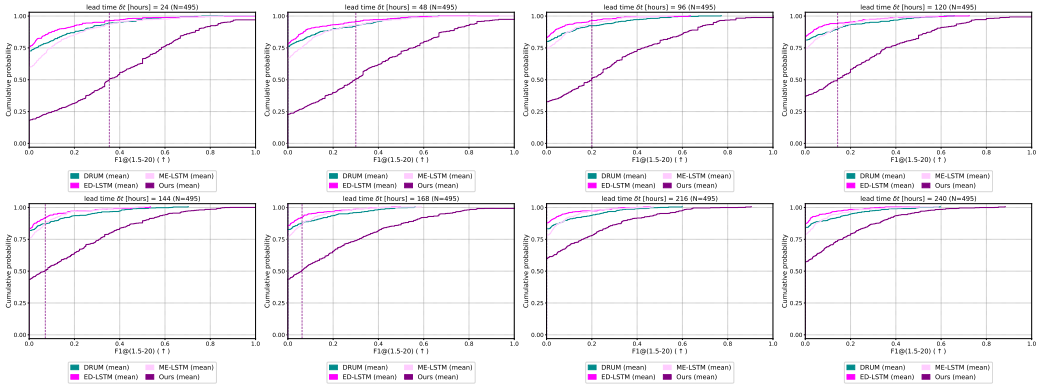


Figure 77: Cumulative distribution function (CDF) of station-level F1@(1.5-20) on CAMELS observations (N=715), across lead times. Dashed vertical lines shows each model's media.

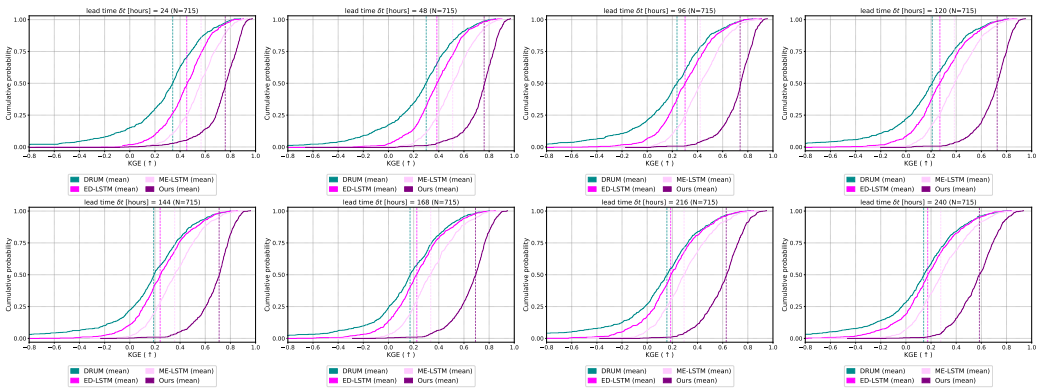


Figure 78: Cumulative distribution function (CDF) of station-level KGE on CAMELS observations (N=715), across lead times. Dashed vertical lines shows each model's media.

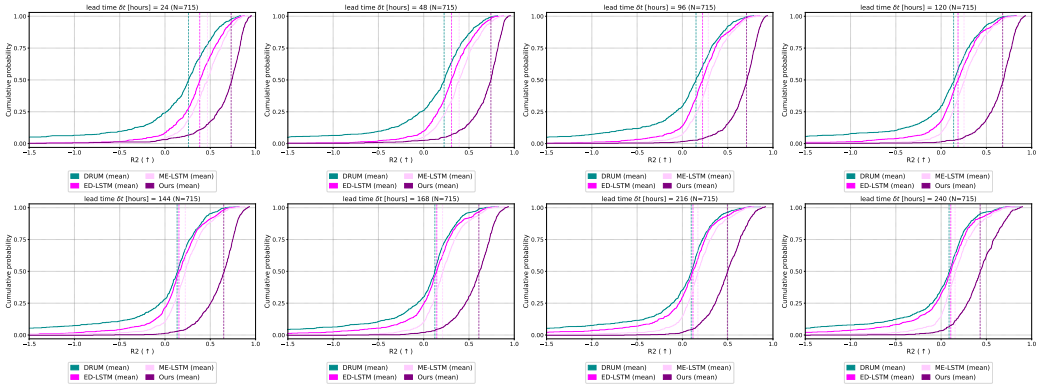


Figure 79: Cumulative distribution function (CDF) of station-level R2 on CAMELS observations (N=715), across lead times. Dashed vertical lines shows each model's media.

K IMPLEMENTATION DETAILS

To make the model compatible with an operational setting, we consider the real-time latency for each source of data. Fig. 80 gives an overview of the input initial conditions and weather forecast. We do not use ERA5-LAND for validating or testing. Instead IFS-HRES data is always used. The training, validation, and testing splits are shown in Fig. 81 along with the timeframe and coverage of input dynamic datasets and target streamflow data. During training, ERA5-Land is used to replace dataset if they are unavailable, e.g., to replace Analysis and IFS-HRES before 2010 and IMERG before 1998. All input data are normalized with z-score pre-computed from the training set except CPC, IMERG, and "rg" variable from EMO which are normalized with the log1p transformation. Accumulated variables are accumulated every 6-hours to match the output temporal resolution. For training we used bfloat16 mixed floating point precision. Epistemic uncertainty is additionally incorporated as deep ensemble [Lakshminarayanan et al., 2017] where we train 2 randomly initialized models.

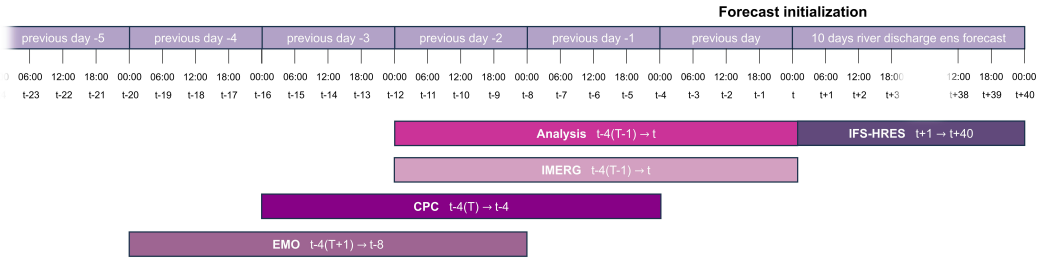


Figure 80: Overview of the input dynamic data to the model. We use 4 daily time steps as input for the initial condition ($T = 4$). For example, CPC data has 1 day latency ($t - 4$) so we take 4 data samples at $t - 4$, $t - 8$, $t - 12$, and $t - 16$, where t is the 6-hourly time step. Note that we ignore here the 4 hours latency for IMERG since we use the daily mean. In an operational setting, step t daily averaged for IMERG can be obtained using the mean over 00:00 UTC→∼20:00 UTC.

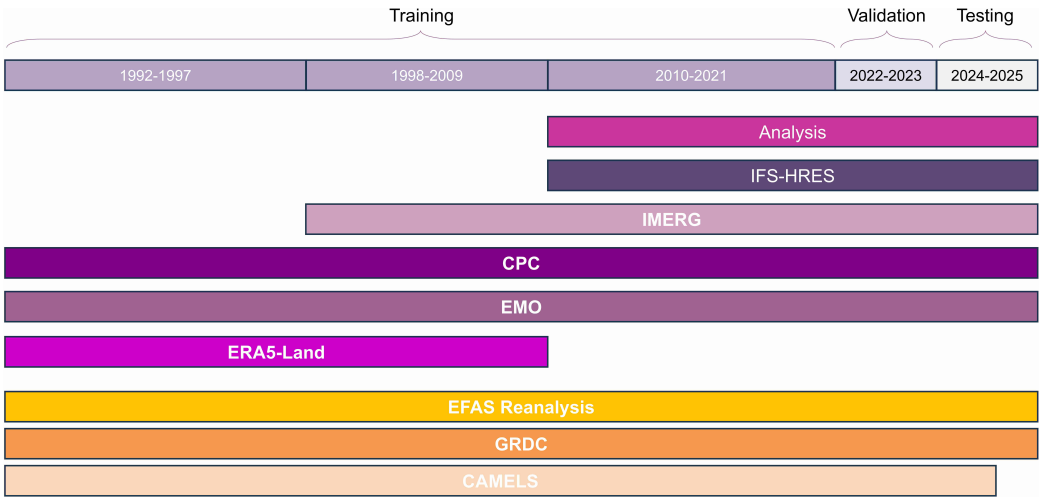


Figure 81: Timeframe and coverage of input dynamic datasets and target streamflow data. Unavailable sources are replaced with ERA5-Land.

We optimize the model gradually on maximum rollout steps of 40 using the following function:

$$n_{rollout} = 1 + \left\lfloor \left(\min\left(1, \frac{e}{0.75E}\right)\right) \cdot \left(S - 0.0001\right) \right\rfloor , \tag{10}$$

where e be the current epoch, $n_{rollout}$ is the number of rollout steps to use at epoch e, E is the total epochs, and S is maximum number of rollout steps.

In Table 31, we highlight the main hyperparameters used in training and finetunning the model.

Table 31: Implementation details.

Configuration	Pre-training (EFAS Reanalysis)	Fine-tuning (GRDC/CAMELS)
Optimizer	AdamW	AdamW
Learning rate	0.0003	0.0001
Minimum learning rate	0.0001	0.0001
Batch size (B)	1	1
Learning rate scheduler	Cosine annealing	Cosine annealing
Weight decay	0.001	0.001
Training epochs	20	20
Starting epoch for fine-tuning on IFS-HRES	11	8
Warmup epochs	0	1
Gradient clip	1	1
Input encoder length (T)	4	4
Lead time (L)	40	40
Number of input points P	200,000	200,000*
Embedding dimension for Analysis	96	96
Embedding dimension for EMO	64	64
Embedding dimension for CPC	16	16
Embedding dimension for IMERG	16	16
Number of encoder layers	3	3
Hidden dimension in encoder block (K)	192	192
Depth of encoder layers	[2, 2, 2]	[2, 2, 2]
Curves in encoder layers	{Sweep_H, Sweep_V, Gilbert, Gilbert trans}	{Sweep_H, Sweep_V, Gilbert, Gilbert trans}
Grouping size in encoder block	[(4, 200000), (2, 200000), (1, 200000)]	[(4, 200000), (2, 200000), (1, 200000)]
Dropout in encoder block	0.1	0.1
D_state in encoder block	8	8
D_conv in encoder block	3	3
Head dimension in encoder block	32	32
Chunk size in encoder block	256	256
MLP ratio in encoder block	1	1
Hidden dimension in forecast block (K)	192	192
Embedding dimension for IFS-HRES (K_{HRES})	64	64
Number of forecast layers	1	1
Depth of forecast layers	[2]	[2]
Curves in forecast layers	{Sweep_H, Sweep_V}	{Sweep_H, Sweep_V}
Grouping size in forecast block	[(1, 200000)]	[(1, 200000)]
Dropout in forecast block	0.1	0.1
D_state in forecast block	8	8
D_conv in forecast block	3	3
Head dimension in forecast block	32	32
Chunk size in forecast block	256	256
Chunk size for routing module	64	64
MLP ratio in encoder block	2	2
Hidden dimension in forecasting head	64	64
Dropout in head	0.1	0.1
Dimension for noise vector	32	32
Standard dviation for the noise vector	0.5	0.5

*For GRDC/CAMELS fine-tuning, we only compute the loss where observations are available.

During training and inference, fitting all ~ 5.2 million input points of the full European EFAS domain into memory is impractical. Following Shams Eddin et al. [2025], we order the points in the domain along a Gilbert space-filling curve and partition the curve it into sub-curves of $200,000$ points each. A simplified version of the splits is shown in Fig. 82.

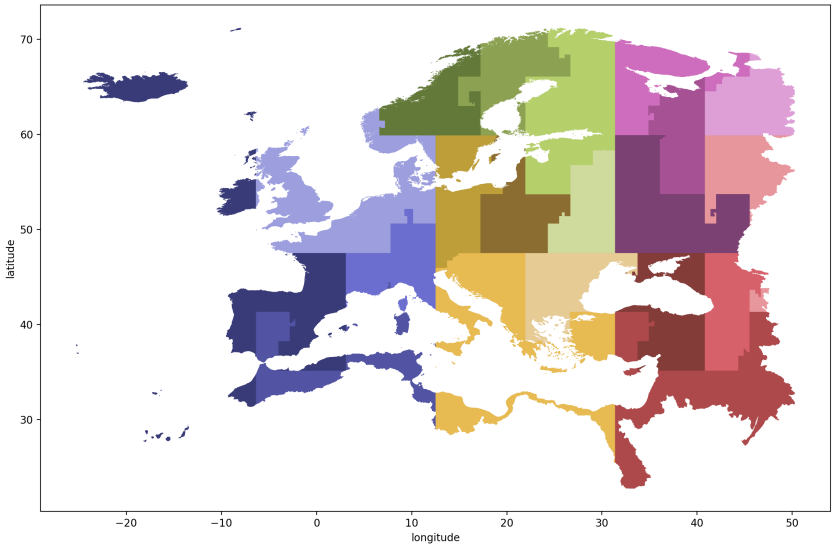


Figure 82: A simplified view of splitting along Gilbert space-filling curve over the EFAS domain. Each split has $200,000$ points. The model runs over these splits and the final results are stitch together to form the prediction over EFAS domain with 5.2 million points.

K.1 Computational time

Training the models on the Mid European domain was done on clusters with NVIDIA A100 $\sim$64 GB GPUs. The final probabilistic model trained over Europe on a cluster with NVIDIA H200 and NVIDIA B200 GPUs. Pre-training on EFAS reanalysis took about 27 hours on 32 GPUs (8 nodes). CRPS was computed for each node with 4 GPUs and the total loss was averaged across nodes. Fine-tuning on observational GRDC/CAMELS data took about 19.5 hours on 32 GPUs. In Table 32, we report the inference time (seconds) w.r.t. the number of input points for our model. We use one NVIDIA A40 GPU for all runs.

Table 32: Inference time in seconds.

10K	20K	40K	80K	160K	300K	600K
0.0154	0.301	0.591	1.172	2.347	4.395	8.913

K.2 Serialization

Serialization fixes the spatio-temporal scanning path over the points fed into each Mamba block. Following the space-filling curve concept [Peano & Peano, 1990] in RiverMamba, we define a bijective map $\xi : \mathbb{Z}^3 \to \mathbb{N}$ that assigns every point a unique 64-bit integer index according to its order along a curve. Sorting points by this index gives the *serialization*, while the inverse $\xi^{-1} : \mathbb{N} \to \mathbb{Z}^3$ recovers each point's original position, i.e. the *deserialization*. For simplicity, curves are defined on the 2D PlateCarree projection of the Earth. We consider the, Generalized Hilbert (Gilbert) curve [Hilbert & Hilbert, 1935] together with Sweep and Zigzag curves in both vertical and horizontal directions. We therefore give each encoder block its own curve, cycling through horizontal Sweep, vertical Sweep, Gilbert, and transposed Gilbert across successive blocks, with a block's spatial curve linked over time by joining its last point at t to its first point at $t{+}1$ (Fig. 83). Varying the curve per block enables the model to scan the sampled points from diverse spatial perspectives, capturing different contextual features.

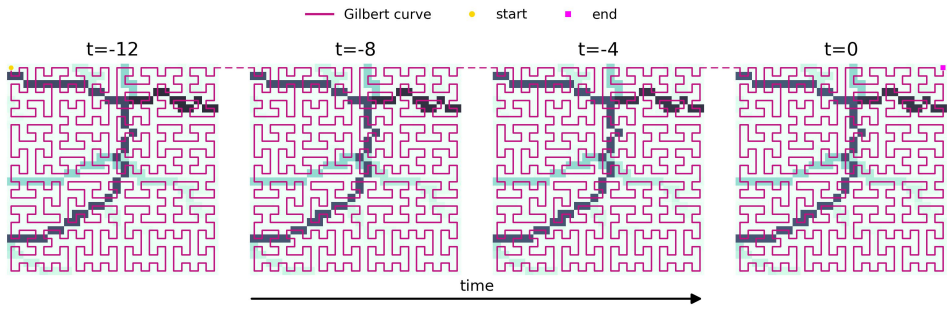


Figure 83: A spatio-temporal scans along 4 frames in the encoder block using Gilbert serialization.

K.3 Location-aware Adaptive Normalization (LOAN)

In order to condition the model on static attributes $\mathbf{X}_{AlphaE}$, the location-aware adaptive normalization layer (LOAN) [Shams Eddin et al., 2023] modulates the features $\mathbf{X}$ systematically within the encoder block:

$$\text{LOAN}(\mathbf{X}) = \left(\frac{\mathbf{X}}{\text{RMS}(\mathbf{X})}\right) + \text{GELU}(\text{Linear}(\mathbf{X}_{AlphaE}))\,, \tag{11}$$

where $\text{RMS}(\mathbf{X}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} x_{(k)}^2}$. A linear layer projects $\mathbf{X}_{AlphaE} \in \mathbb{R}^{B,1,P,V_s}$, with $V_s = 64$ being the number of embedding channels, to $\mathbb{R}^{B,1,P,K}$, which is then broadcast along the temporal dimension to match $\mathbf{X} \in \mathbb{R}^{B,T,P,K}$, the input to the LOAN layer. $\text{RMS}(\mathbf{X})$ is likewise broadcast across all K channels. For the first LOAN application within the encoder block only, we skip the RMS normalization (i.e. $\text{RMS}(\mathbf{X}) \to 1$ in Eq. 11) and condition directly on the normalized input features $\mathbf{X}$.

L Evaluation Metrics

We evaluate the model performance across five standard metrics in hydrological modeling and flood forecasting. This includes Root Mean Square Error (RMSE), the Coefficient of Determination (R^2), Kling–Gupta Efficiency (KGE), the F1 score, and the Continuous Ranked Probability Score (CRPS). Individual metric calculation between observed river discharge (X^{obs}) and predicted river discharge (X^{pred}) for points P are given below:

Root Mean Square Error (RMSE). computes the square root of the mean squared deviations between predicted and observed values:

$$\text{RMSE} = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(X_p^{\text{obs}} - X_p^{\text{pred}})^2}\,. \tag{12}$$

Coefficient of Determination (R2). Measures a model's explanatory power relative to the mean of observed data. In hydrological applications, R2 has the same meaning as Nash–Sutcliffe Efficiency (NSE). Values close to 1 reflect near-perfect agreement, whereas negative values indicate that the model performs worse than the mean:

$$\text{R2} = 1 - \frac{\sum_{p=1}^{P}(X_p^{\text{obs}} - X_p^{\text{pred}})^2}{\sum_{p=1}^{P}(X_p^{\text{obs}} - \bar{X}^{\text{obs}})^2}\,. \tag{13}$$

Kling–Gupta Efficiency (KGE). Is a combination of correlation, bias, and variability. Values close to 1 indicate high predictive accuracy, while scores below 0 indicate performance inferior to the observed mean:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}\,, \tag{14}$$

$$\beta = \frac{\bar{X}^{\mathrm{pred}}}{\bar{X}^{\mathrm{obs}}}, \quad \gamma = \frac{\mathrm{CV}^{\mathrm{pred}}}{\mathrm{CV}^{\mathrm{obs}}}, \quad r = \text{Pearson correlation coefficient}\,, \tag{15}$$

where r denotes the Pearson correlation between observed and simulated discharge series, β represents the bias ratio, and γ is the variability ratio. The respective coefficients of variation are given by $\mathrm{CV}^{\mathrm{obs}} = \sigma^{\mathrm{obs}}/\bar{X}^{\mathrm{obs}}$ and $\mathrm{CV}^{\mathrm{pred}} = \sigma^{\mathrm{pred}}/\bar{X}^{\mathrm{pred}}$, with σ^{obs} and σ^{pred} are the standard deviations of observed and predicted streamflow.

Continuous Ranked Probabilistic Score (CRPS). Is used for probabilistic forecast. It measures the deviation for each ensemble member from the observations using mean absolute error in the first term and measures the spread of the ensemble members in the second term:

$$\mathrm{fCRPS} = \frac{1}{P}\sum_{p=1}^{P}\left[\frac{1}{M}\sum_{m=1}^{M}\left|X_p^{\mathrm{pred}} - X_p^{\mathrm{obs}}\right| - \frac{1}{2M(M-1)}\sum_{m,m'=1}^{M}\left|X_p^{\mathrm{pred}} - X_p^{\mathrm{obs}}\right|\right], \tag{16}$$

where M is the number of ensemble members. During evaluation, we set $M = 12$ ensemble members for all probabilistic models.

F1 Score. Represents the harmonic mean of precision and recall. It can serve as a robust benchmark for imbalanced prediction tasks such as detecting less frequent flood events:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}\,. \tag{17}$$

These metrics are calculated independently for each point's time series before being aggregated to determine the final score. For RMSE, R2, KGE, and CRPS, we always report the median for all points, while for F1 score, we report the mean over all points. all scores are first aggregated per lead time forecast e.g., from $t = 0$ to $t = 40$ and then averaged to get the final score for all lead times. Categorical metrics, such as the F1 score, evaluate event detection performance across varied return intervals. For instance, when streamflow exceeds a 2-year flood threshold on a given day, an accompanying forecast above that threshold is marked as a true positive. Unless noted otherwise, all reported F1 scores reflect averages taken across return periods 1.5 to 20 years.

M FLOODS RETURN PERIODS

The return period (or recurrence interval) estimates how often a hydrological flood event of a given magnitude occurs on average. Here we apply it to flood frequency. A 2-year flood, for instance, has a 50% annual exceedance probability (AEP). Return period (RP) is simply the inverse of AEP: $RP = \frac{1}{AEP}$.

These return periods are commonly used to set discharge thresholds for flood warnings (e.g., triggering an alert once discharge exceeds the 2-year level). Depending on context, a high return period (e.g., 20 years) does not necessarily mean flooding will occur but it reflects the statistical rarity in streamflow magnitude. We therefore treat return period as a proxy for hydrological extremity, which we call flood severity.

Following the EFAS approach, we define thresholds for return periods of 1.5, 2, 5, 10, 20, 50, 100, 200, and 500 years. For EFAS reanalysis, return levels are estimated by fitting a Gumbel distribution (L-moments method) to annual maxima. We use the pre-computed return period data from the Copernicus Emergency Management Service (https://confluence.ecmwf.int/display/CEMS/Auxiliary+Data). Figs. 87 and 89 show the flood thresholds. For the GRDC and CAMELS observational datasets, return periods are computed per station using each station's full available record. We did not compute return periods for the ML reforecasts, as this would require a long reforecast climatology to fit an extreme value distribution.

N BASELINES

N.1 CLIMATOLOGY

Following Harrigan et al. [2023] and Shams Eddin et al. [2025], we compute the climatological baseline using the long-term river discharge record for the years 1992 to 2025. For each day of the year, climatology was computed over a 31-day centered moving window, from which we derived 11 fixed percentiles at 10% intervals. Climatology can be used as a probabilistic baseline with 11 ensemble members (11 percentiles).

N.2 PERSISTENCE

Persistence is defined as the daily river discharge of EFAS reanalysis from the day preceding the day on which the forecast was issued, i.e., for a forecast starting at 00:00 UTC t, the persistence is defined as the river discharge one day before at 00:00 UTC $t - 4$. The prediction remains constant across all lead-time steps. Note that this baseline has an advantage over the deep learning models because it has access to the values of the river discharge, while deep learning models need to estimate the magnitude of discharge from initial conditions and weather forecast.

N.3 EFAS FORECAST

Operational forecast from EFAS v5.0 was obtained from the ECMWF Early Warning Data Store (EWDS) https://doi.org/10.24381/cds.9f696a7a. These operational real-time data are provided by the Copernicus Emergency Management Service (CEMS), which is managed and developed by the European Commission's Joint Research Centre. EFAS forecasts are generated by forcing the LISFLOOD hydrological model [Knijff et al., 2010] with meteorological data from the ECMWF ensemble forecast (ENS). The deterministic EFAS (control run) does not use perturbation, while the probabilistic version has 51 ensemble members via perturbation. We used the high-resolution EFAS v5.0 forecast from 29 May 2024 to the end of 2025 for the deterministic control run and data from 01 June 2024 to 31 August 2024 for the probabilistic forecast. We compare to this baseline in Sec. 4.2 and Sec. F.

N.4 RIVERMAMBA

We adapted RiverMamba from Shams Eddin et al. [2025]. The baseline relies on Mamba [Gu & Dao, 2024] as a State Space Model (SSM) to effectively capture spatio-temporal relations in large-scale river networks through scanning on space-filling curves. The RiverMamba baseline has no explicit routing mechanism inside the model. It relies mainly on the spatio-temporal modeling of SSM without considering local drain direction. We followed Shams Eddin et al. [2025] and the official implementation from https://github.com/HakamShams/RiverMamba_code to define the main hyper-parameters. RiverMamba was only trained over Mid European domain for 16 lead time. Since RiverMamba has to build a dedicated forecasting layer for each step, we used hidden dimension of 128 for the encoder and forecasting layers to reduce overfitting and computations. For optimization, we pretrained for 20 epochs on reanalysis and fine-tune for 10 epochs on observations. We compare to this baseline in Sec. 4.1 and E.

N.5 DRUM

This is a probabilistic and diffusion-based baseline adapted from Ou et al. [2025]. DRUM trains a conditional and an unconditional noise-prediction network jointly for classifier-free guidance and samples forecasts via DDIM. We used the official repository from https://github.com/ozg2021/DRUM/tree/main to train the model. The model is trained on single-step nowcasting and 40 forecast steps at 6-hourly resolution using denoising jointly in one forward pass. We used 1200 diffusion steps (linear β schedule, $\beta_1 = 10^{-4}$, $\beta_T = 0.02$), 120 step DDIM, hidden size of 256, and time embedding dimension of 128. For optimization, we used AdamW with 10^{-4} learning rate and 10^{-2} weight decay for 200 epochs on EFAS reanalysis and for 50 epochs for fine-tuning on observation.

N.6 ED-LSTM (ENCODER-DECODER LSTM

This is a probabilistic baseline from Google's operational flood forecasting system [Nearing et al., 2024]. We followed the implementation in NeuralHydrology (https://github.com/neuralhydrology/neuralhydrology). In ED-LSTM, a bidirectional hindcast LSTM encodes the meteorological forcing initial conditions, and its final hidden and cell states pass through a learnable handoff network to initialize a unidirectional forecast LSTM that decodes the forecast horizon. The probabilistic forecast is generated via a Mixed Asymmetric Laplace (CMAL) with $K{=}3$ components. The model is trained on a single-step nowcasting and 40 forecast steps at 6-hourly resolution. We used a hidden size of 256 for both LSTMs and a state-handoff network with hidden layers $[512, 256]$. For optimization, we used Adam with 10^{-3} learning rate for 35 epochs on EFAS reanalysis and 20 epochs for fine-tuning on observations.

N.7 ED-LSTM (MEAN EMBEDDING LSTM)

This is the new probabilistic baseline from Google's operational flood forecasting system adapted from Cohen et al. [2026]. We followed the implementation in NeuralHydrology (https://github.com/neuralhydrology/neuralhydrology). The probabilistic forecast is generated via a Mixed Asymmetric Laplace (CMAL) with $K{=}3$ components. The model is trained on a single-step nowcasting and 40 forecast steps at 6-hourly resolution. We used a hidden size of 256 for both LSTMs, output dropout of 0.4, and an initial forget-gate bias of 3. For optimization, we used Adam with 5×10^{-4} learning rate for 35 epochs on EFAS reanalysis and 20 epochs for fine-tuning on observation.

O DATASET

O.1 EFAS REANALYSIS RIVER DISCHARGE DATA

The European Flood Awareness System (EFAS) Dottori et al. [2017] provides consistent representation of hydrological variables across Europe with sub-daily temporal resolution and high spatial resolution of $0.017°$ (1-arcminute ~ 1.5 km at EFAS latitudes) compared to $0.05°$ in the Global Flood Awareness System (GloFAS) [Harrigan et al., 2020] (Fig. 84). The system is fully operational and developed by the European Commission's Joint Research Centre (JRC) while operated by ECMWF under the Copernicus Emergency Management Service (CEMS). It delivers river discharge reanalysis from 1992 to present covering the European domain (72° N-24° S, 35° W-74° E) on the regular latitude-longitude grid. The reanalysis is generated using the LISFLOOD hydrological and kinematic wave equations for river routing [Knijff et al., 2010] to simulate realistic river discharge (dis06, in m^3 s^{-1}) across Europe. The sub-daily EFAS reanalysis discharge data represents the mean value of river discharge for the previous 6 hours in UTC as a reference time.

Fig. 84 shows the full domain of EFAS which we used for the main experiments and Fig. 85 shows the domain we defined over Mid Europe for the ablations and other experiments on Mid Europe. The full domain has an image size of 2970×4530 and ~ 5.2 million points over land, while the domain over Mid Europe has 540×600 and around $\sim 281K$ points over land. In our work, we used the latest version v5.0 of EFAS and we only used the river discharge "dis06" as a target variable for the forecast (no data from EFAS is used as initial condition or input to the model in our work). The dataset is publicly available on Climate Data Store and Early Warning Data Store (EWDS) https://doi.org/10.24381/cds.e3458969.

O.2 OBSERVATIONAL RIVER DISCHARGE DATA

No streamflow data were used as inputs to our model. They were only used as targets for fine-tuning and evaluation. Note that we fine-tuned on each observational dataset individually. In the following, we describe GRDC and CAMELS observational datasets that we used in our study.

O.2.1 GRDC

Observed river discharge is obtained from the Global Runoff Data Centre (GRDC) which provides quality controlled discharge measurements from stations worldwide. The GRDC Data is given as

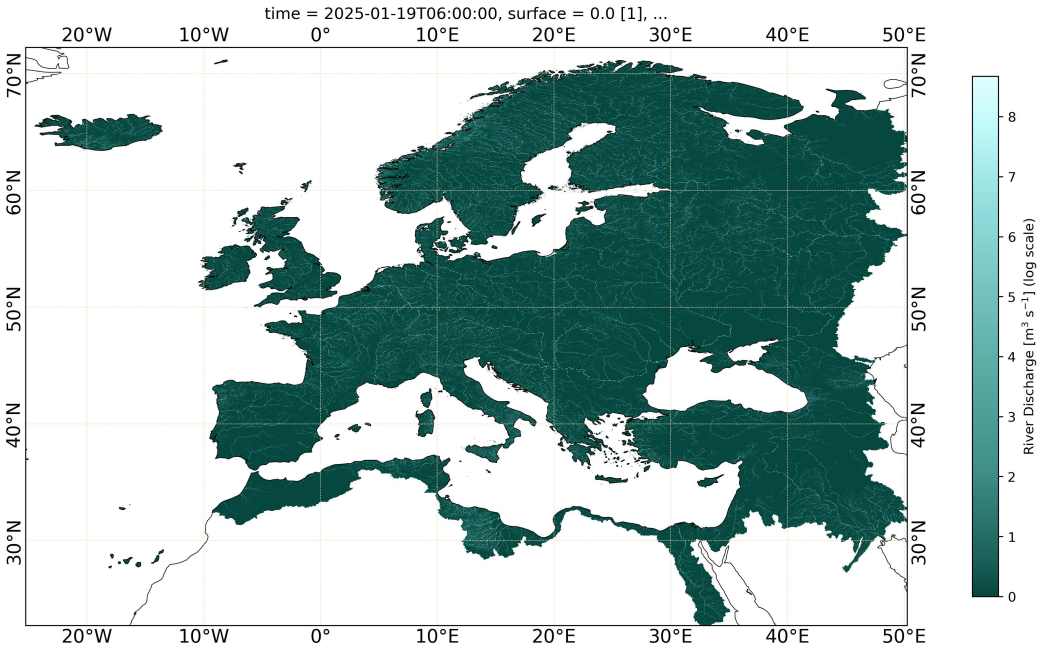


Figure 84: Overview of the EFAS domain.

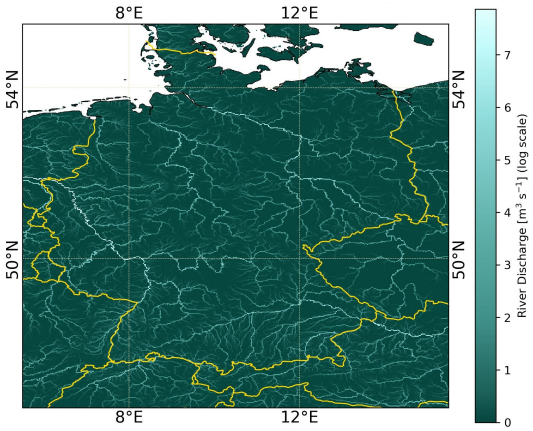


Figure 85: Overview of the Mid European domain as a subset of EFAS domain. We used this domain for the ablation studies.

daily time series and it is publicly available at https://grdc.bafg.de/. Every GRDC daily value is timestamped at 00:00 at the start of the day (left-labeled). Following Shams Eddin et al. [2025] and Dottori et al. [2017], we discarded catchments with a drainage area under 150 km^2, leaving 2,158 GRDC stations after matching them to the EFAS drainage network (Sec. O.1). This avoids large mismatches between the drainage area reported by GRDC and the one provided by the LISFLOOD EFAS static layers. Each station was geo-located by projecting it onto the EFAS grid, evaluating its 9 nearest grid cells, and keeping the cell with the highest KGE. Stations whose matched cell still differed from the reported drainage area by more than 20% and scored below 0.4 KGE were then dropped. This left 1,971 stations for evaluation across the EFAS domain. Over this final set, the median drainage-area mismatch fell to 2.4% (interquartile range 0.75%–6.05%). Discharge values are reported daily in m^3 s^{-1} and were converted from local time to 00:00 UTC via linear interpolation. For evaluation, the originally left-labeled GRDC series were converted to right-labeled series so they align temporally with the right-labeled EFAS simulation output. F1 scores for the GRDC evaluation are computed over detection windows synchronized between model predictions and observations.

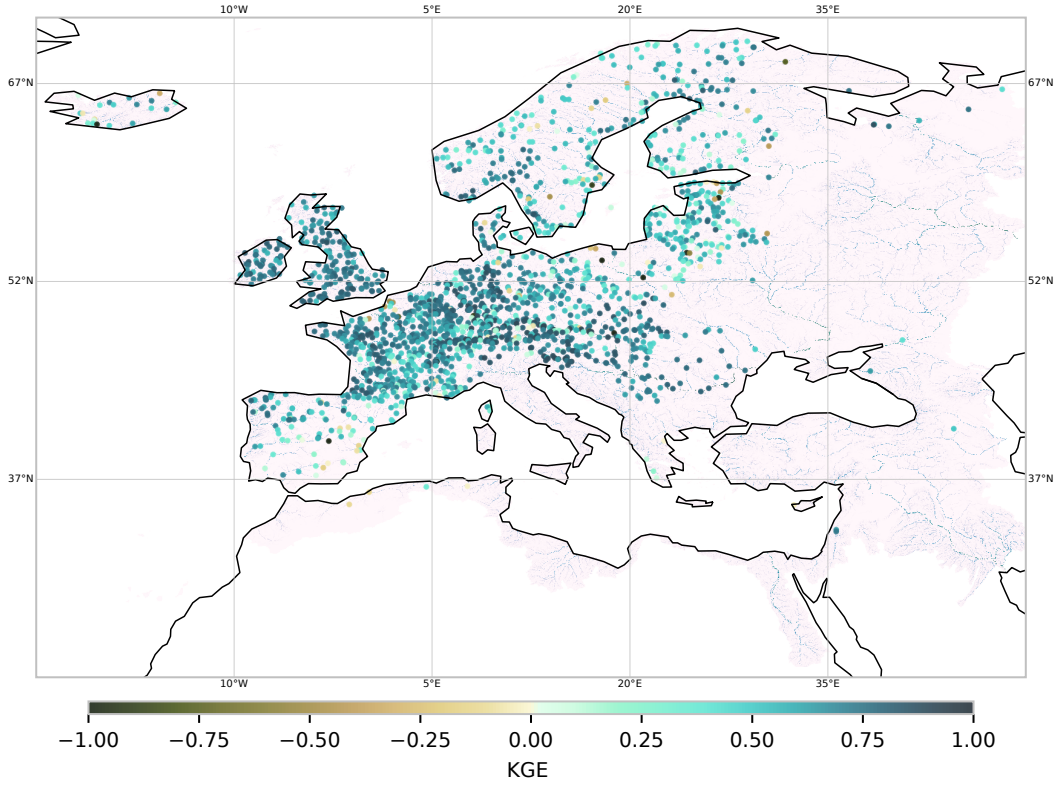


Figure 86: Locations of the 1,971 selected GRDC stations used to train and evaluate the model. Color encodes the KGE of the EFAS reanalysis discharge against the GRDC observations at each station.

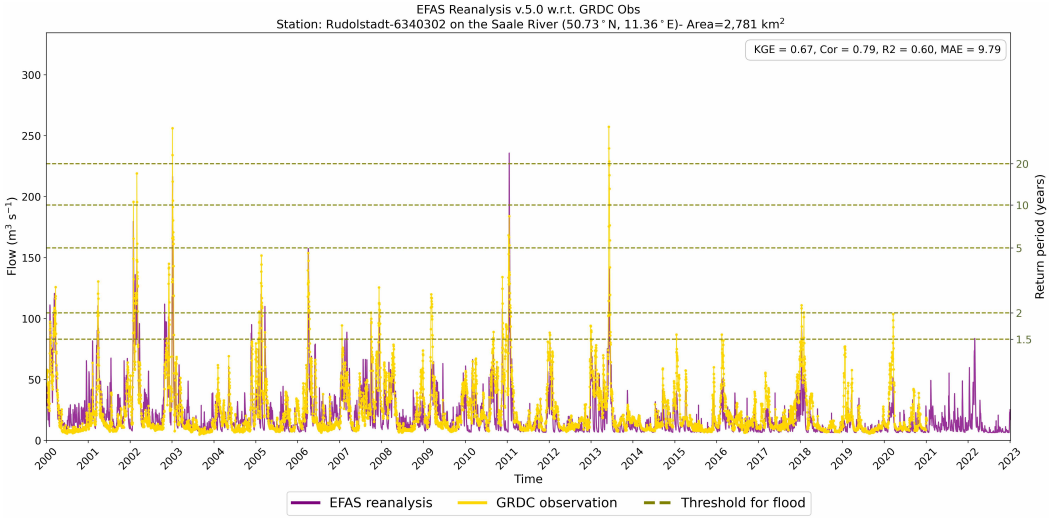


Figure 87: Hydrograph for EFAS reanalysis (purple line) from 1 January 2000 to 31 December 2022 and observations (gold line), for a gauging station on the Saale River (Rudolstadt, GRDC-6340302). The top-right box displays summary statistics from the reanalysis's evaluation against the observations.

Fig. 86 gives an overview of the KGE values for EFAS reanalysis data against the GRDC observation at the locations of all 1,971 selected stations. In general, there is a good agreement between EFAS and GRDC data over Europe, with a median KGE at 0.65 and an interquartile range of 0.50 to 0.75. Fig. 87 shows an exemplar hydrograph for a gauged station.

It should be noted that, unlike the GRDC observations, the EFAS reanalysis represents naturalized flow and does not account for anthropogenic influences like dams, reservoirs, irrigation withdrawals, and other water-management practices.

O.2.2 CAMELS

In addition to GRDC, we processed the CAMELS (Catchment Attributes and MEteorology for Large-sample Studies) as a separate dataset. CAMELS dataset comes from different sources from various national agencies and research projects. All CAMELS datasets have daily data and only few have hourly data e.g., CAMELS-DE [Loritz et al., 2024]. To make the dataset compatible, we use the daily river discharge from all CAMELS datasets.

We combine 10 national CAMELS datasets, following the same matching procedure used for GRDC (Sec. O.2.1). Each station was discarded if its drainage area was below 150 km^2. For matching we projected the points onto the EFAS grid and evaluated its 9 nearest grid cells, keeping the cell with the highest KGE, and then dropped if its matched cell still differed from the reported drainage area by more than 20% and scored below 0.4 KGE. We excluded the area matching procedure for CAMELS-ES because reanalysis has very bad quality there where EFAS reanalysis shows poor agreement with observations in Spain. This left 2,215 CAMELS stations across the ten datasets which are listed in Table 33, with a median drainage-area mismatch of 3.6% (interquartile range 1.2%–10.6%).

Table 33: CAMELS datasets used in this work. Shown are the source of data, observation period, and the number of stations retained after matching to the EFAS drainage network.

Dataset	Source	Country	Start	End	# Stations
CAMELS-CH	Höge et al. [2023]	Switzerland	1981	2020	136
CAMELS-CZ	Institute et al. [2025]	Czechia	1978	2024	173
CAMELS-DE	Loritz et al. [2024]	Germany	1951	2020	605
CAMELS-DK	Liu et al. [2025]	Denmark	1989	2019	54
CAMELS-ES	Casado-Rodríguez et al. [2026]	Spain	1991	2020	244
CAMELS-FI	Seppä et al. [2026]	Finland	1961	2023	234
CAMELS-FR	Delaigue et al. [2025]	France	1970	2021	368
CAMELS-GB	Coxon et al. [2026]	Great Britain	1970	2022	332
CAMELS-LUX	Nijzink et al. [2025]	Luxembourg	2004	2021	25
CAMELS-SE	Teutschbein [2024]	Sweden	1961	2020	44
Total*					2,206

*Row counts sum to 2,215; 9 grid cells are shared between neighboring countries' stations and are counted once in the combined mask used for training/evaluation.

Fig. 88 maps the KGE between EFAS reanalysis and CAMELS observations at all 2,215 stations. In general, there is a good agreement between EFAS and CAMELS data over Europe (except Spain), with a median KGE at 0.54 and an interquartile range of 0.18 to 0.69. Fig. 89 shows a representative hydrograph for one gauged station.

O.3 IFS-HRES DATA

We get operational weather forecast data from the ECMWF Archive Catalogue (https://www.ecmwf.int/en/forecasts/dataset/operational-archive). The forecast is produced by the deterministic forecast of the ECMWF Integrated Forecast System (IFS) High Resolution (HRES) atmospheric model. The forecast runs once per day at 00:00 UTC and up to 10 days lead time with 6-hourly temporal resolutions and 41 time steps (step 0 is analysis, see Sec. O.4). We processed data from 2007 to 2025. The data before 2007 were replaced by ERA5-Land data (Sec. O.5). The data has at 3 spatial resolutions, 25 KM from 2007-01-01 to 2010-01-26, 16 KM from 2010-01-27 to 2016-03-08, and 9 KM from 2016-03-09 to 2026-01-01 at $0.1^\circ \times 0.1^\circ$ resolution. To match the resolution of EFAS (1 arcminute), we precompute the nearest neighbored point from IFS-HRES to EFAS points for these 3 different resolutions. During training/inference, we used the precomputed nearest neighborhood mappings to get IFS-HRES forecast data at EFAS resolution. The variables

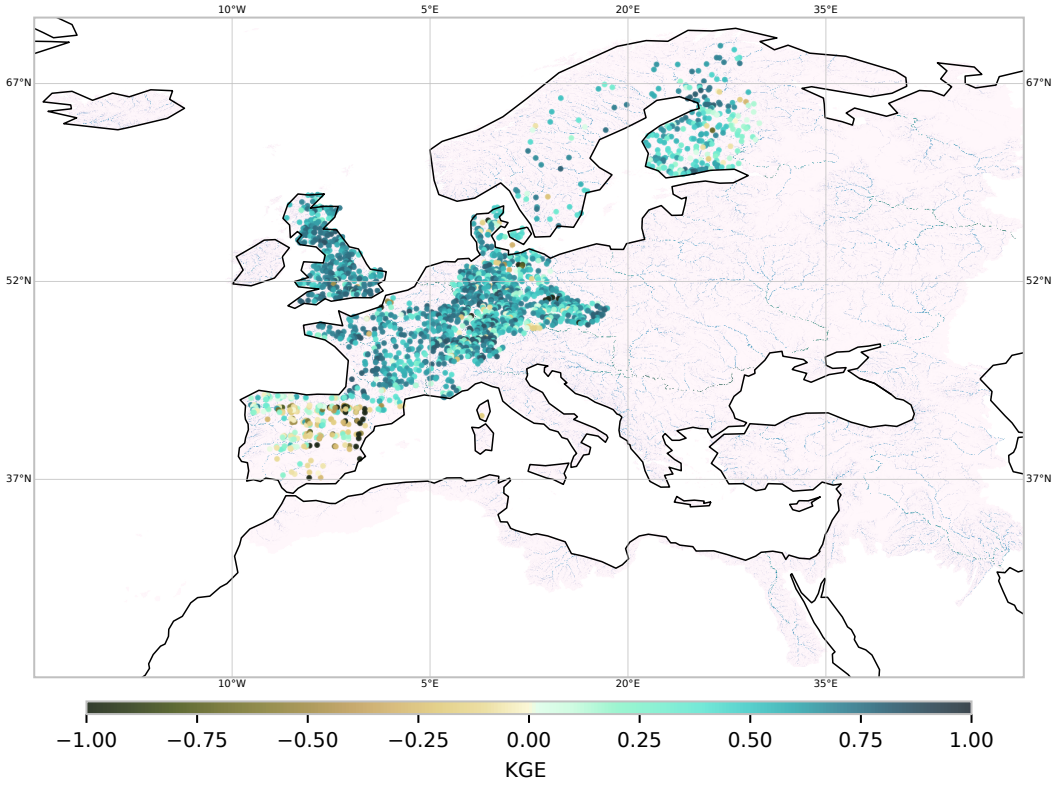


Figure 88: Locations of the 2,240 selected CAMELS stations used to train and evaluate the model. Color encodes the KGE of the EFAS reanalysis discharge against the CAMELS observations at each station.

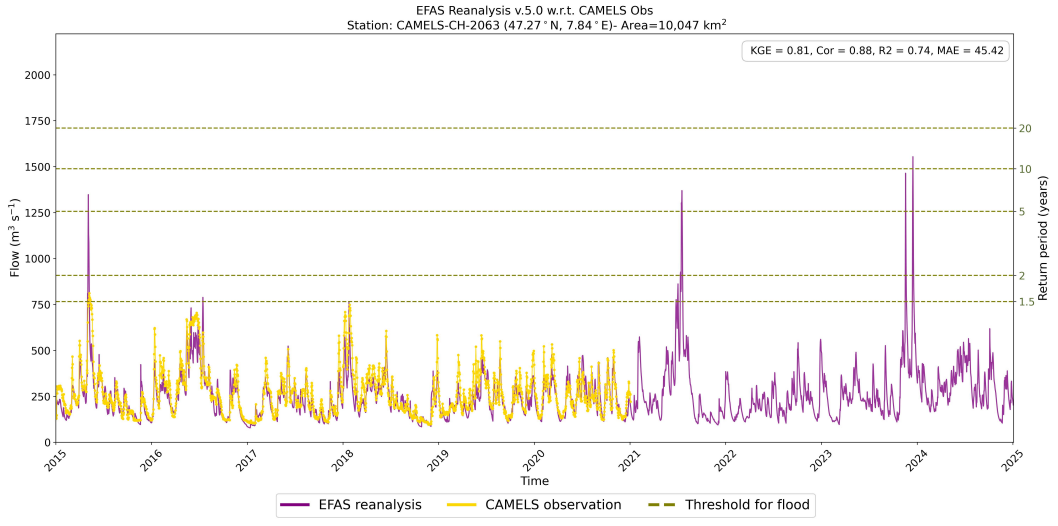


Figure 89: Hydrograph for EFAS reanalysis (purple line) from 1 January 2015 to 31 December 2024 and observations (gold line), for a gauging station in Switzerland (CAMELS-CH-2063). The top-right box displays summary statistics from the reanalysis's evaluation against the observations.

include 6 instantaneous variables and 8 accumulated variables. Details about the processed variables are available in Table 34.

O.4 ANALYSIS DATA

Near real-time operational analysis data are used as initial condition and obtained from step 0 of IFS-HRES (Sec. O.3). Analysis data only provide instantaneous variables, to have the accumulated variables for analysis, we use the forecast of the previous day at step 4. The processed variables are the same as IFS-HRES (Sec. O.3) and described in Table 34.

O.5 ERA5-LAND REANALYSIS DATA

ERA5-Land reanalysis provides a reconstruction of the historical land surface state [Muñoz Sabater et al., 2021]. We do not use ERA5-Land as initial condition. ERA5-Land is only used to substitute the forecast and analysis of IFS-HRES data (Sec. O.3) for the years 1992-2006 in the initital phase of the training. We processed the data from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (https://doi.org/10.24381/cds.e2161bac). ERA5-Land is provided at $0.1° \times 0.1°$ at regular latitude and longitude (Plate Carrée projection). To match the resolution of EFAS (1 arcminute), we precompute the nearest neighbored point from ERA5-Land to EFAS points and during training/inference, we used the precomputed nearest neighborhood mapping to get ERA5-Land data at EFAS resolution. The processed variables are the same as IFS-HRES (Sec. O.3) and described in Table 34. More details about ERA5-Land can be found in Muñoz Sabater et al. [2021].

Table 34: Details about the processed variables from the ECMWF Integrated Forecast System (IFS) High Resolution (HRES) atmospheric model (Sec. O.3) and ERA5-Land reanalysis (Sec. O.5). IFS-HRES accumulated state variables start at 00:00 UTC of the forecast time and ends at the end of the forecast. ERA5-Land accumulation starts at 00:00 UTC previous day and ends at 00:00 UTC current day.

Variable	Long name	Unit	Height	Surface parameters
d2m	2m dewpoint temperature	K	2m	instantaneous
e	total evaporation	m of water equivalent	surface	accumulated
ro	runoff	m	surface & subsurface	accumulated
sd	snow depth	m of water equivalent	surface	accumulated
sf	snowfall	m of water equivalent	surface	accumulated
smlt	snowmelt	m of water equivalent	surface	accumulated
sp	surface pressure	Pa	surface	instantaneous
ssr	surface net solar radiation	J/m^2	surface	accumulated
str	surface net thermal radiation	J/m^2	surface	accumulated
swvl1	volumetric soil water	m^3/m^3	soil layer (0 - 7 cm)	instantaneous
t2m	2m temperature	K	2m	instantaneous
tp	total precipitation	m	surface	accumulated
u10	10 metre U wind component	m/s	10m	instantaneous
v10	10 metre V wind component	m/s	10m	instantaneous

O.6 EMO DATA

European Meteorological Observations (EMO) are gridded high-resolution data obtained from historical and real-time ground weather stations across Europe. EMO data are used originally to provide the EFAS with near real-time initial condition. We also use EMO as initial condition for our model with 2 days latency to mimic the operational setting (EMO data is updated annually for public use but real-time EMO data can be obtained with a close collaboration with ECMWF). We processed the data daily at 1 arcminute spatial resolution for the years 1990 to 2025. The processed variables are described in Table 35. The data for can be obtained from https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/CEMS-EFAS/meteorological_forcings/EMO-1arcmin/.

O.7 IMERG DATA

We obtain observational precipitation estimates from NASA's Integrated Multi-satellitE Retrievals for GPM (IMERG V07) product [Huffman et al., 2020]. IMERG data are used only as initial condition. We use the near real-time early run data of processed mean daily precipitation at $0.1° \times 0.1°$

Table 35: Details about the processed variables from EMO data [Thiemig et al., 2022].

Variable	Long name	Unit	Height	Surface parameters
pd	water vapour pressure	hPa	-	instantaneous
pr	total precipitation	mm	surface	accumulated
rg	downward surface solar radiation	J/m^2	surface	accumulated
tn	minimum air temperature	C$^\circ$	2m	instantaneous
tx	maximum air temperature	C$^\circ$	2m	instantaneous
ws	wind speed	m/s	10m	instantaneous

spatial resolution. To match the resolution of EFAS data, we regridded IMERG data onto the EFAS domain using nearest point algorithm (nearest_s2d) implemented by Zhuang et al. [2020]. The daily mean rate in mm/day is derived by computing the mean precipitation rate in mm/hour for every grid cell of the respective day, and then multiplying the result by 24 to get mm/day. IMERG is available from the year 1998. Since we start the training from year 1992, we substitute IMERG data for the years 1992-1998 with the total precipitation (tp) of ERA5-Land (Sec. O.5) converted from m to accumulated mm/day. The IMERG data for operational forecast can be retrieved from https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html with ~ 4 hours latency.

O.8 CPC DATA

We obtained the observational gauged-based precipitation data from the National Oceanic and Atmospheric Administration (NOAA), Climate Prediction Center (CPC). The data are accumulated daily at $0.5^\circ \times 0.5^\circ$ spatial resolution. To match the resolution of EFAS data, we regridded CPC data onto the EFAS domain using nearest point algorithm (nearest_s2d) implemented by Zhuang et al. [2020]. CPC data are only used as initial condition. The CPC data for operational use can be retrieved from https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html with ~ 1 day latency. More details about the data can be found in Xie et al. [2007]; Chen et al. [2008]; Xie et al. [2010].

O.9 STATIC DATA

In addition to AlphaEarth/LISFLOOD static features, we also use the Cartesian coordinates for the points on the WGS-84 ellipsoid as positional encoding in the way as Shams Eddin et al. [2025].

O.9.1 ALPHAEARTH

AlphaEarth embeddings provide annual representation of the Earth surface at very high resolution (10×10 meter squares) with 64 bands (A01-A64). The embeddings are produced by Google DeepMind and obtained by training a foundation model on multi-source Earth observations [Brown et al., 2025]. To match the resolution of EFAS (1 arcminute), we first preprocessed the AlphaEarth data on Google Engine by taking the mean embeddings at 1 KM resolution (processed in small tiles). Then we generate a mosaic of these tiles and map it to EFAS grid via bilinear interpolation. Since the data only span the years 2017 to 2024 (at the date of processing), we average over the time dimension to get the static embeddings. Fig. 90 shows the embeddings for 3 years over Germany and Fig. 91 shows the processed data over EFAS domain. The raw data can be obtained from https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL?utm_source=deepmind.google&utm_medium=referral&utm_campaign=gdm&utm_content.

O.9.2 LISFLOOD

We used LISFLOOD static maps from Choulga et al. [2024]. The data are provided at the same resolution of EFAS domain (1 arcminute). For the experiments, we used 119 time-invariant variables covering 7 main categories and excluding reservoirs, lakes and some static water demand variables. The EFAS LISFLOOD static maps can be obtained from the

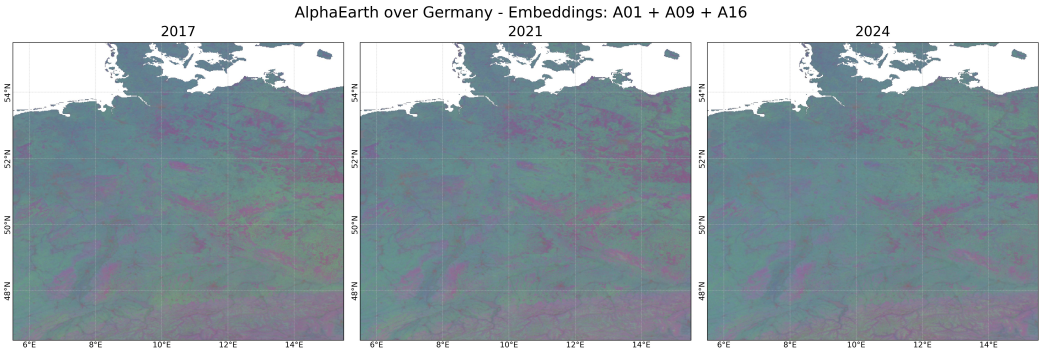


Figure 90: AlphaEarth Embedding over Germany for the years 2017, 2021, and 2024. Shown are Bands; A01, A09, and A16 as an RGB image.

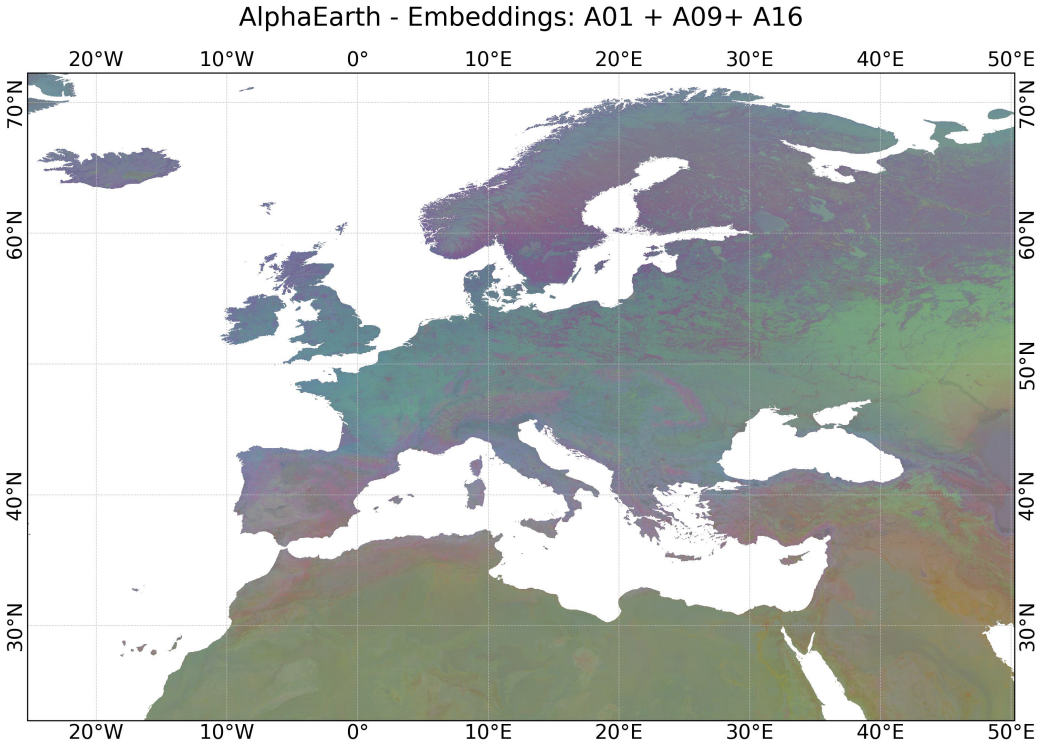


Figure 91: AlphaEarth Embedding over EFAS domain. Shown are Bands; A01, A09, and A16 as an RGB image.

Joint Research Centre Data Catalogue at https://data.jrc.ec.europa.eu/dataset/f572c443-7466-4adf-87aa-c0847a169f23.

P CODE AND DATA AVAILABILITY

We will release the datasets over mid Europe and the code upon publication. The source code will include instructions on how to access and use the data. GRDC data that we used in this study is available after signing a license agreement with the data provider at https://grdc.bafg.de/. For CAMELS data, see Sec. O.2.2 in the Appendix. Raw ECMWF-IFS HRES weather forecasts and analysis can be downloaded from ECMWF archive catalogue (MARS). The raw River discharge and related historical data from the European Flood Awareness System (EFAS) can be downloaded from the Early Warning Data Store

(EWDS) at https://doi.org/10.24381/cds.e3458969 [Dottori et al., 2017]. The raw LISFLOOD static and parameter maps can be downloaded from the European Commission, Joint Research Centre (JRC) at https://data.jrc.ec.europa.eu/dataset/f572c443-7466-4adf-87aa-c0847a169f23 [Choulga et al., 2024]. The raw EMO data can be downloaded from the European Commission, Joint Research Centre (JRC) at EMOhttp://data.europa.eu/89h/0bd84be4-cec8-4180-97a6-8b3adaac4d26. The raw CPC data can be obtained from the National Oceanic and Atmospheric Administration (NOAA), Climate Prediction Center (CPC) at https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html. IMERG data can be downloaded from NOAA at https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html. The raw ERA5-Land reanalysis data can be downloaded from NOAA at Copernicus Climate Change Service (C3S), Climate Data Store (CDS) at https://doi.org/10.24381/cds.e2161bac [Muñoz Sabater et al., 2021].

REFERENCES

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Jesús Casado-Rodríguez, Gonçalo Ramos-Gomes, and Peter Salamon. Simulación del caudal en españa utilizando redes neuronales long short-term memory. *Ingeniería del agua*, 30(1):63–78, 2026.

Mingyue Chen, Wei Shi, Pingping Xie, Viviane B. S. Silva, Vernon E. Kousky, R. Wayne Higgins, and John E. Janowiak. Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research: Atmospheres*, 113(D4):D04110, 2008. doi: 10.1029/2007JD009132.

M. Choulga, F. Moschini, C. Mazzetti, S. Grimaldi, J. Disperati, H. Beck, P. Salamon, and C. Prudhomme. Technical note: Surface fields for global environmental modelling. *Hydrology and Earth System Sciences*, 28(13):2991–3036, 2024. doi: 10.5194/hess-28-2991-2024. URL https://hess.copernicus.org/articles/28/2991/2024/.

D. Cohen, R. Amira, R. Aschner, Y. Carny, B. Feinstein, H. Fester, S. Fronman, M. Gauch, O. Gilon, R. Green, A. Hassidim, D. Klotz, F. Kratzert, D. Korenfeld, G. Loike, A. Markel, Y. Matias, R. Mayo, A. Metzger, B. Mosheyev, A. Niego, S. Rees, E. Reinstein, A. Sicherman, G. Shalev, O. Shefi, Y. Shildan, I. Zemach, O. Zlydenko, and G. Nearing. Extending medium-range global flood forecasts: The google global flood forecasting model version 2. *EGUsphere*, 2026:1–31, 2026. doi: 10.5194/egusphere-2026-2283. URL https://egusphere.copernicus.org/preprints/2026/egusphere-2026-2283/.

G. Coxon, Y. Zheng, R. Barbedo, H. Cooper, F. Fileni, H. J. Fowler, M. Fry, A. Green, T. Gribbin, H. Harfoot, E. Lewis, G. G. R. Neto, X. Qiu, S. Salwey, and D. E. Wendt. Camels-gb v2: hydrometeorological time series and landscape attributes for 671 catchments in great britain. *Earth System Science Data*, 18(6):4345–4371, 2026. doi: 10.5194/essd-18-4345-2026. URL https://essd.copernicus.org/articles/18/4345/2026/.

O. Delaigue, G. M. Guimarães, P. Brigode, B. Génot, C. Perrin, J.-M. Soubeyroux, B. Janet, N. Addor, and V. Andréassian. Camels-fr dataset: a large-sample hydroclimatic dataset for france to explore hydrological diversity and support model benchmarking. *Earth System Science Data*, 17(4):1461–1479, 2025. doi: 10.5194/essd-17-1461-2025. URL https://essd.copernicus.org/articles/17/1461/2025/.

F. Dottori, M. Kalas, P. Salamon, A. Bianchi, L. Alfieri, and L. Feyen. An operational procedure for rapid flood risk assessment in europe. *Natural Hazards and Earth System Sciences*, 17(7):1111–1126, 2017. doi: 10.5194/nhess-17-1111-2017. URL https://nhess.copernicus.org/articles/17/1111/2017/.

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

S. Harrigan, E. Zsoter, L. Alfieri, C. Prudhomme, P. Salamon, F. Wetterhall, C. Barnard, H. Cloke, and F. Pappenberger. Glofas-era5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*, 12(3):2043–2060, 2020. doi: 10.5194/essd-12-2043-2020. URL https://essd.copernicus.org/articles/12/2043/2020/.

S. Harrigan, E. Zsoter, H. Cloke, P. Salamon, and C. Prudhomme. Daily ensemble river discharge reforecasts and real-time forecasts from the operational global flood awareness system. *Hydrology and Earth System Sciences*, 27(1):1–19, 2023. doi: 10.5194/hess-27-1-2023. URL https://hess.copernicus.org/articles/27/1/2023/.

David Hilbert and David Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes: Nebst Einer Lebensgeschichte*, pp. 1–2, 1935. doi: https://doi.org/10.1007/BF01199431.

M. Höge, M. Kauzlaric, R. Siber, U. Schönenberger, P. Horton, J. Schwanbeck, M. G. Floriancic, D. Viviroli, S. Wilhelm, A. E. Sikorska-Senoner, N. Addor, M. Brunner, S. Pool, M. Zappa, and F. Fenicia. Camels-ch: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic switzerland. *Earth System Science Data*, 15(12):5755–5784, 2023. doi: 10.5194/essd-15-5755-2023. URL https://essd.copernicus.org/articles/15/5755/2023/.

George J. Huffman, David T. Bolvin, Dan Braithwaite, Kuo-Lin Hsu, Robert J. Joyce, Christopher Kidd, Eric J. Nelkin, Soroosh Sorooshian, Erich F. Stocker, Jackson Tan, David B. Wolff, and Pingping Xie. *Integrated Multi-satellite Retrievals for the Global Precipitation Measurement (GPM) Mission (IMERG)*, pp. 343–353. Springer International Publishing, Cham, 2020. ISBN 978-3-030-24568-9. doi: 10.1007/978-3-030-24568-9_19. URL https://doi.org/10.1007/978-3-030-24568-9_19.

Czech Hydrometeorological Institute, Jakub Krejčí, and Grey Nearing. [camels-cz] catchment attributes and meteorology for large-sample studies – czechia, November 2025. URL https://doi.org/10.5281/zenodo.17769325.

J. M. Van Der Knijff, J. Younis, and A. P. J. De Roo and. Lisflood: a gis-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, 24(2):189–212, 2010. doi: 10.1080/13658810802549154. URL https://doi.org/10.1080/13658810802549154.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

J. Liu, J. Koch, S. Stisen, L. Troldborg, A. L. Højberg, H. Thodsen, M. F. T. Hansen, and R. J. M. Schneider. Camels-dk: hydrometeorological time series and landscape attributes for 3330 danish catchments with streamflow observations from 304 gauged stations. *Earth System Science Data*, 17 (4):1551–1572, 2025. doi: 10.5194/essd-17-1551-2025. URL https://essd.copernicus.org/articles/17/1551/2025/.

R. Loritz, A. Dolich, E. Acuña Espinoza, P. Ebeling, B. Guse, J. Götte, S. K. Hassler, C. Hauffe, I. Heidbüchel, J. Kiesel, M. Mälicke, H. Müller-Thomy, M. Stölzle, and L. Tarasova. Camels-de: hydro-meteorological time series and attributes for 1582 catchments in germany. *Earth System Science Data*, 16(12):5625–5642, 2024. doi: 10.5194/essd-16-5625-2024. URL https://essd.copernicus.org/articles/16/5625/2024/.

J. Muñoz Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, B. Martens, D. G. Miralles, M. Piles, N. J. Rodríguez-Fernández, E. Zsoter, C. Buontempo, and J.-N. Thépaut. Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021. doi: 10.5194/essd-13-4349-2021. URL https://essd.copernicus.org/articles/13/4349/2021/.

Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, et al. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563, 2024. doi: https://doi.org/10.1038/s41586-024-07145-1.

J. Nijzink, R. Loritz, L. Gourdol, D. Zoccatelli, J. F. Iffly, and L. Pfister. Camels-lux: Highly resolved hydro-meteorological and atmospheric data for physiographically characterized catchments around luxembourg. *Earth System Science Data Discussions*, 2025:1–34, 2025. doi: 10.5194/essd-2024-482. URL https://essd.copernicus.org/preprints/essd-2024-482/.

Zhigang Ou, Congyi Nai, Baoxiang Pan, Yi Zheng, Chaopeng Shen, Peishi Jiang, Xingcai Liu, Qiuhong Tang, Wenqing Li, and Ming Pan. Probabilistic diffusion models advance extreme flood forecasting. *Geophysical Research Letters*, 52(15):e2025GL115705, 2025. doi: https://doi.org/10.1029/2025GL115705. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025GL115705. e2025GL115705 2025GL115705.

Giuseppe Peano and G Peano. *Sur une courbe, qui remplit toute une aire plane*. Springer, 1990. doi: https://doi.org/10.1007/BF01199438.

I. Seppä, C. Gonzales Inca, J. Uusikivi, and P. Alho. Camels-fi: hydrometeorological time series and landscape properties for 320 catchments in finland. *Earth System Science Data*, 18(7):4745–4769, 2026. doi: 10.5194/essd-18-4745-2026. URL https://essd.copernicus.org/articles/18/4745/2026/.

Mohamad Hakam Shams Eddin, Ribana Roscher, and Juergen Gall. Location-aware adaptive normalization: A deep learning approach for wildfire danger forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023. doi: 10.1109/TGRS.2023.3285401.

Mohamad Hakam Shams Eddin, Yikui Zhang, Stefan Kollet, and Jürgen Gall. River-mamba: A state space model for global river discharge and flood forecasting. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 133401–133423. Curran Associates, Inc., 2025. doi: 10.52202/085713-4446. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c1501f836c75dd1da19974a38c64ee1e-Paper-Conference.pdf.

Claudia Teutschbein. Camels-se: Long-term hydroclimatic observations (1961–2020) across 50 catchments in sweden as a resource for modelling, education, and collaboration. *Geoscience Data Journal*, 11(4):655–668, 2024. doi: https://doi.org/10.1002/gdj3.239. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/gdj3.239.

Vera Thiemig, Goncalo N. Gomes, Jon O. Skøien, Markus Ziese, Armin Rauthe-Schöch, Elke Rustemeier, Kira Rehfeldt, Jakub P. Walawender, Christine Kolbe, Damien Pichon, Christoph Schweim, and Peter Salamon. EMO-5: a high-resolution multi-variable gridded meteorological dataset for Europe. *Earth System Science Data*, 14(7):3249–3272, 2022. doi: 10.5194/essd-14-3249-2022.

Pingping Xie, Mingyue Chen, Song Yang, Akiyo Yatagai, Tadahiro Hayasaka, Yoshihiro Fukushima, and Changming Liu. A gauge-based analysis of daily precipitation over east asia. *Journal of Hydrometeorology*, 8(3):607–626, 2007.

Pingping Xie, M Chen, and W Shi. Cpc unified gauge-based analysis of global daily precipitation. In *Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc*, volume 2, 2010.

Jiawei Zhuang, raphael dussin, André Jüling, and Stephan Rasp. JiaweiZhuang/xESMF: v0.3.0 Adding ESMF.LocStream capabilities, March 2020. URL https://doi.org/10.5281/zenodo.3700105.