RiverMamba: A State Space Model for Global River Discharge and Flood Forecasting

Mohamad Hakam Shams Eddin **Juergen Gall**

Institute of Computer Science, University of Bonn
Lamarr Institute for Machine Learning and Artificial Intelligence
{shams, gall}@iai.uni-bonn.de

Yikui Zhang **Stefan Kollet**

Institute of Bio- and Geosciences Agrosphere (IBG-3), Research Centre Jülich
Centre for High-Performance Scientific Computing in Terrestrial Systems,
Geoverbund ABC/J, Jülich
{yik.zhang, s.kollet}@fz-juelich.de

Abstract

Recent deep learning approaches for river discharge forecasting have improved the accuracy and efficiency in flood forecasting, enabling more reliable early warning systems for risk management. Nevertheless, existing deep learning approaches in hydrology remain largely confined to local-scale applications and do not leverage the inherent spatial connections of bodies of water. Thus, there is a strong need for new deep learning methodologies that are capable of modeling spatio-temporal relations to improve river discharge and flood forecasting for scientific and operational applications. To address this, we present RiverMamba, a novel deep learning model that is pretrained with long-term reanalysis data and that can forecast global river discharge and floods on a 0.05° grid up to 7 days lead time, which is of high relevance in early warning. To achieve this, RiverMamba leverages efficient Mamba blocks that enable the model to capture spatio-temporal relations in very large river networks and enhance its forecast capability for longer lead times. The forecast blocks integrate ECMWF HRES meteorological forecasts, while accounting for their inaccuracies through spatio-temporal modeling. Our analysis demonstrates that RiverMamba provides reliable predictions of river discharge across various flood return periods, including extreme floods, and lead times, surpassing both AI- and physics-based models. The source code and datasets are publicly available at the project page https://hakamshams.github.io/RiverMamba.

1 Introduction

Riverine floods are one of the most destructive natural disasters, with their risk anticipated to rise in the future as a result of climate change and socioeconomic developments [1–5]. They arise from compound effects, including atmospheric conditions like heavy precipitation caused by circulation patterns and snowmelt succeeding high temperature, all shaped by the specific characteristics of the river drainage area [6]. The interaction of these elements influences flood timing, scale, and severity [6]. This complexity makes future flood risk assessment challenging, as a changing climate may alter these drivers in unpredictable ways [7]. Therefore, early prediction of flood risk, especially for extreme floods, is a key measure for effective flood risk mitigation [8, 9].

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

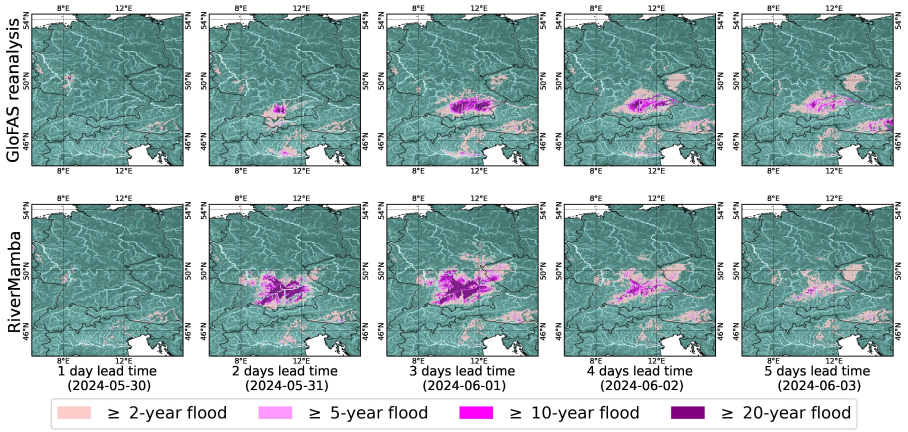

Figure 1: Example of a 5-day forecast of river discharge and flood events. In early June 2024, a significant flood affected Southern Germany. While the top row shows the floods obtained from the GloFAS reanalysis, the bottom row shows the river discharge forecast by our approach. The severity of floods is categorized by the statistical flood return period, i.e., occurring every 10 years.

To support national forecasting initiatives, current operational flood early warning systems can forecast river discharge in real-time and provide flood forecasts at different scales [10–12]. The discharge forecasts derived from these systems can be further processed using inundation models to create anticipated flooded areas [13, 14]. The Global Flood Awareness System (GloFAS) [15, 16], developed under the Copernicus Emergency Management Service (CEMS) and operated by the European Centre for Medium-Range Weather Forecasts (ECMWF), represents the cutting-edge physics-based model for real-time and worldwide hydrological forecasting. However, physics-based hydrological models are expensive to run and require extensive calibration to handle complex catchment characteristics.

AI-based early warning systems are thus considered as vital tools to enhance climate risk resilience [17, 18] and to enable flood forecasting without requiring full physical process understanding [19, 20]. While deep-learning approaches for weather forecasting [21–23] have been investigated in recent years, very little work has been done for forecasting river discharge at large spatial regions since it is very challenging. It requires the combination of sparse gauged river observations with high-resolution land surface, re-analysis, and weather forecast data. Furthermore, floods occur rarely and the goal is to forecast floods of different severity as shown in Fig. 1. Recently, an LSTM-based model has been proposed [24]. While it achieves promising results, it forecasts floods only locally at sparse river basins and does not consider routing. Modeling spatio-temporal relations, however, is very important and required to generate consistent dense maps as in Fig. 1, since river discharge at points near connected bodies of water is highly correlated.

In this work, we propose the first deep learning approach for global river discharge and flood forecasting that is not only capable of forecasting at sparse gauged observation points, but also of forecasting accurate, high-resolution ($0.05°$) global river discharge maps. In order to deal with the sparseness of gauged river points and the computational complexity of modeling spatio-temporal relations at the global scale, our proposed RiverMamba leverages Mamba blocks, which are bidirectional state space models [25–28], and spatio-temporal forecast blocks. Using a specialized procedure to convert sampled points into 1D sequences, RiverMamba maintains a very large spatio-temporal receptive field, connecting the routing of the river channel networks and the teleconnection of meteorological data across space and time. RiverMamba has thus the possibility to consider a spatio-temporal context that covers very large river networks like the Amazon River. The forecast layers are further forced by high-resolution meteorological data (HRES) to generate medium-range river discharge forecasts up to 7 days lead time. To address uncertainty in the meteorological forcing, we built the forecast layers so that they can, for each catchment point, incorporate information about meteorological forcing from the neighboring points and throughout the temporal dimension. Thus, RiverMamba ensures a consistent forecast through space and time. Our contributions can be summarized as follows:

- We introduce a novel Mamba-based approach, called RiverMamba, for global river discharge and flood forecasting. It is the first deep learning approach that is capable of providing maps of global river discharge forecasting at $0.05°$, and it introduces a novel methodology to

hydrology. It is able to integrate sparse gauged observations, river attributes, high-resolution reanalysis data, and weather forecast. The efficient structure allows to model spatio-temporal relations covering entire river networks.

- We evaluate RiverMamba on both long-term real-world reanalysis and observational data where it outperforms state-of-the-art AI- and physics-based operational systems for global flood forecasting.

2 Related works

Flood forecasting. Floods can be categorized into three common types. The first type is the *fluvial* or *riverine flood* [29]. It occurs when the water level in a stream rises and overflows onto the adjacent land. The second type is the *coastal flood*, also known as storm surges [30]. The third type is the *pluvial flood*, often referred to as *flash flood* [31–33] that can occur with extreme rainfalls. Machine learning (ML) has become an essential element for the development of hydrological simulation and flood models [34, 35]. Each type of flood has unique drivers and impacts. Consequently, ML methods require different strategies to forecast them. Related tasks to flood forecasting are urban flood modeling [36–39], flood inundation [40–42], and flood extension and susceptibility mapping [43–46]. In this work, we are interested in forecasting riverine floods (fluvial) based on river discharge.

River discharge forecasting. River discharge can be used to detect fluvial flood signals when the magnitude of the flow exceeds certain thresholds. Current deep learning methods for forecasting river discharge are primarily based on locally lumped models [47, 48], hypothesizing that a single model can generalize across many catchments without considering the spatial-temporal information over grids [49]. The dominating backbone is the LSTM model [50] which is used in most recent studies such as EA-LSTM [51, 52], ED-LSTM [53, 54], Hydra-LSTM [55], MC-LSTM [56], MF-LSTM [57], and DRUM [58]. These models learn features specific to individual rivers or entities and lack spatial and topological information. However, river networks have spatio-temporal causal relations [59]. Only a few studies deviate from this conventional modeling and propose to model the network topology with Graph Neural Networks [60–62]. They are still limited to small scales and the graph models fail in most cases to capture topological information [60]. Others applied an LSTM model on a coarse grid to estimate runoff and then coupled it with a river routing model to produce daily discharge at coarse resolution [63]. In [64], LSTM resolves local runoff spatially on a regular grid in central Europe. Then, routing the runoff along the entire river networks is implemented as 1D-convolutions and fully connected layers. The impact of defining routing explicitly with physics-informed neural networks has also showed an advantage in recent studies, especially, in improving streamflow in large continental river networks compared to models that do not consider routing [65, 66]. In a hybrid modeling framework, physical equations including river routing are parametrized using 3D-convolutions and fully connected layers for distributed hydrological modeling [67]. The most relevant work is the Encoder-Decoder LSTM [51] developed for the Google global operational forecasting system [24], which is a locally lumped model. In this work, distinct from previous works, we propose an approach that is capable of modeling a large spatio-temporal context and forecasting medium-range river discharge at grid-scale.

State space model (SSMs) and the Mamba family. Linear SSMs [25] and structured SSMs like S4 [26] and S5 [27] were primarily introduced for long-sequence modeling in NLP. Recently, Mamba [28] introduced the selective scan mechanism, enabling efficient training and linear-time inference. Built upon Mamba, VMamba [68] and Vim [69] in the vision domain were introduced as appealing alternatives to the quadratic complexity of vision transformers [70] while improving scaling efficiency on long token sequences. A series of works have adapted Mamba to tasks like image generation [71, 72], image classification [73, 74], video understanding [75, 76], motion generation [77], dense action anticipation [78], and point cloud processing [79, 80]. In this work, we propose a Mamba-based approach for global river discharge and flood forecasting.

3 RiverMamba

In this work, we present the first deep learning approach that not only forecasts flood events at sparse gauged river observations, but that is capable of forecasting accurate, high-resolution (i.e., at $0.05°$) maps of river discharge up to few days at global scale, as shown in Figs. 1 and 2. These maps are essential to forecast flood events of various severity like a flood that re-occurs statistically within a

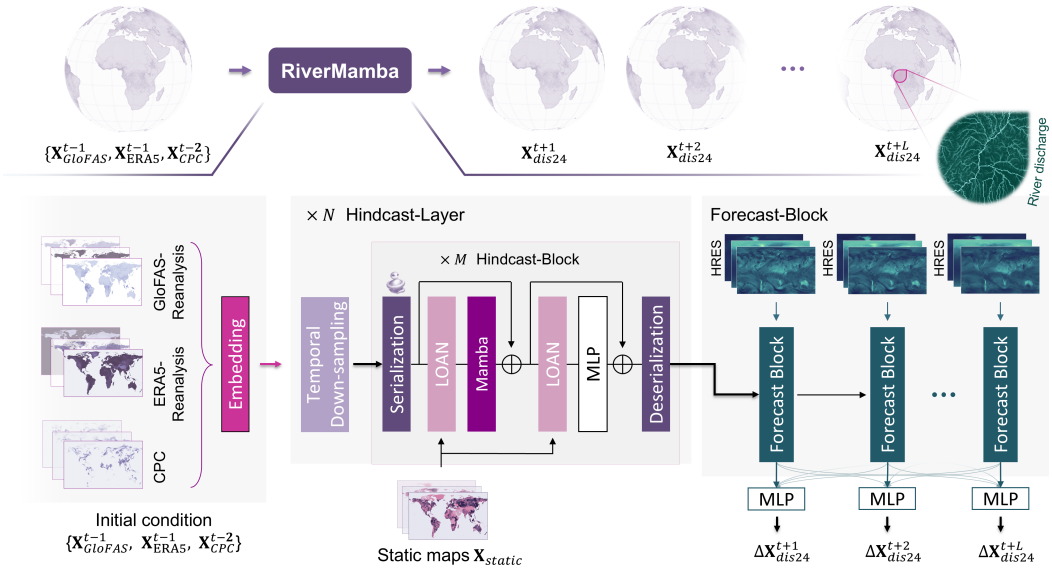

Figure 2: An overview of the proposed RiverMamba model for river discharge forecasting. The model forecasts at time t, high-resolution river discharge maps $\mathbf{X}_{dis24}^{t+1:t+L}$ from initial conditions ($\mathbf{X}_{ERA5}^{t-T:t-1}$, $\mathbf{X}_{GloFAS}^{t-T:t-1}$, $\mathbf{X}_{CPC}^{t-T-1:t-2}$), static river attributes ($\mathbf{X}_{static}$), and meteorological forecasts ($\mathbf{X}_{HRES}^{t+1:t+L}$).

1.5-year return period or a 'flood of the century'. This is very challenging since it requires a model that models spatial-temporal relations in an efficient way and integrates different sources of data (Fig. 2).

As input, we use the initial condition of the forecasts from ERA5-Land reanalysis [81], denoted by $\mathbf{X}_{ERA5}^{t-T:t-1} = \{\mathbf{X}_{ERA5}^{t-T}, \ldots, \mathbf{X}_{ERA5}^{t-2}, \mathbf{X}_{ERA5}^{t-1}\}$, the initial condition $\mathbf{X}_{GloFAS}^{t-T:t-1}$ from the GloFAS reanaylsis data [82], and the initial condition $\mathbf{X}_{CPC}^{t-T-1:t-2}$ from the operational global unified gauge-based analysis of daily precipitation [83–85]. We also include data from weather forecasts, where we use the high-resolution meteorological forcing forecasts $\mathbf{X}_{HRES}^{t+1:t+L}$ from the ECMWF Integrated Forecast System (IFS), where L is the lead time for the forecast. We generate the river discharge forecast at t, using 00:00 UTC as reference time, for $t{+}1$ until $t{+}L$. This means that we do not address nowcasting but only forecasting as it is more relevant. We also do not include any nowcasts ($\mathbf{X}_{HRES}^{t}$) as input. The rationale behind this is to ensure broader applicability, since many weather forecast systems especially ML models provide forecasts at $t > 0$. However, adding nowcasts to the model is straightforward if they are available. To make the setup as realistic as possible, we do not include any data after 00:00 UTC and we consider $\mathbf{X}_{GloFAS}$ and $\mathbf{X}_{ERA5}$ at day $t-1$ and $\mathbf{X}_{CPC}$ at day $t-2$. Additionally, we include river attributes $\mathbf{X}_{static}$ like catchment morphology from LISFLOOD [86]. The input variables are described in details in the suppl. material. Given these inputs, RiverMamba forecasts changes of the daily mean river discharge $\Delta\mathbf{X}_{dis24}^{t+1:t+L}$ relative to the daily mean river discharge at $t{-}1$, i.e., $\mathbf{X}_{dis24}^{t-1}$. The forecast daily mean river discharge is thus given by $\mathbf{X}_{dis24}^{t+l} = \mathbf{X}_{dis24}^{t-1} + \Delta\mathbf{X}_{dis24}^{t+l}$.

An overview of RiverMamba is shown in Fig. 2. For training, we sample P points that are on the land surface and near water bodies. The details are described in the suppl. material. For each point p, we obtain a temporal sequence of embedding vectors $\mathbf{X}_{embed}^{t-T:t-1}(p)$:

$$\mathbf{X}_{embed}^{t}(p) = \mathrm{LN}\Big(\mathrm{Tanh}\Big(\mathrm{Concat}\big(\mathrm{Linear}(\mathbf{X}_{ERA5}^{t}(p)), \mathrm{Linear}(\mathbf{X}_{GloFAS}^{t}(p)), \mathrm{Linear}(\mathbf{X}_{CPC}^{t-1}(p))\big)\Big)\Big), \tag{1}$$

where LN is the layer norm and Linear is the projection layer. The dimensions of the input are $\mathbf{X}_{ERA5} \in \mathbb{R}^{B\times T\times P\times V_e}$, $\mathbf{X}_{GloFAS} \in \mathbb{R}^{B\times T\times P\times V_g}$, and $\mathbf{X}_{CPC} \in \mathbb{R}^{B\times T\times P\times 1}$, where B is the batch size, V_e is the number of variables from ERA5, and V_g is the number of variables from GloFAS. The embedding $\mathbf{X}_{embed} \in \mathbb{R}^{B\times T\times P\times K}$, where $K = 192$ is the dimensionality of the embedding, is then the input to the encoder defined by the hindcast layers.

Hindcast layer. The hindcast layers model spatio-temporal relations and aggregate the observations over time. Except for the 1^{st} layer which processes the full temporal resolution, the temporal resolution is down-sampled by a factor of 2 with a linear layer at the beginning of each hindcast layer, such that the output of the last hindcast layer $\mathbf{X}_{hindcast} \in \mathbb{R}^{B \times 1 \times P \times K}$ has a temporal resolution of $T = 1$. In our implementation, we chose $T = 4$ as for the GloFAS operational system and a temporal down-sampling of 2. Consequently, we defined 3 layers to encode the input.

The hindcast layers further integrate the static river attributes $\mathbf{X}_{static}$ that contain additional information like catchment morphology, which is relevant for flood forecasting. While we analyze the impact of the different inputs, in particular the river attributes, in the suppl. material, another key aspect of the hindcast blocks is the specialized serialization of the spatio-temporal points and the Mamba blocks [28, 68, 69]. The serialization defines the way the sampled points are connected, and the Mamba block efficiently updates the features of each point based on the spatio-temporal structure. This is a very important design choice since transformer blocks are computationally infeasible for global flood forecasting, whereas [24] does not consider spatial relations at all. In the suppl. material, we also show that an alternative using Flash-Attention [87, 88] is inferior in terms of inference time and accuracy compared to our approach.

The output of the last hindcast layer is then processed along with the HRES meteorological forcing by forecast blocks, and MLP-based regression heads predict for each lead time l the difference of daily mean river discharge $\Delta\mathbf{X}_{dis24}^{t+l}$ with respect to the daily mean river discharge at $t-1$. In the following, we describe the components of RiverMamba in details.

Hindcast block. As shown in Fig. 2, the hindcast block has three main components: serialization and deserialization, location-aware adaptive normalization layers (LOAN) to integrate static river attributes, and the Mamba block.

Serialization. The serialization defines the spatio-temporal scanning path over all sampled points for the following Mamba block. For this, we propose space-filling curves that sequentially traverse through all points. The concept was introduced in [89] and the space-filling can be defined as a bijective function $\Phi : \mathbb{Z}^3 \rightarrow \mathbb{N}$, where every point in the discrete space corresponds to a unique index within the sequence. We call this mapping the serialized encoding. The serialized decoding is done as $\Phi^{-1} : \mathbb{N} \rightarrow \mathbb{Z}^3$, where every index is mapped back into its corresponding position. We call this deserialization. We investigated three curves: the Generalized Hilbert (Gilbert) curve, which is a generalized version of the Hilbert curve [90], as well as the Sweep and Zigzag curves in vertical and horizontal directions. Examples of space-filling curves in 2D are illustrated in Fig. 3. As shown in the suppl. material, a combination of Sweep and Gilbert curves performs best. To this end, each hindcast block has its own curve. As shown in Fig. 3, we sweep in the first block over the horizontal direction. The spatial curves are connected over time by continuing the last point of the curve at t with the first point of the curve at $t+1$. The second block then sweeps over the vertical direction and we continue with the Gilbert curve and its transposed. These four space-filling curves are iterated. By altering the curves sequentially through the hindcast blocks, the sampled points will be connected and scanned from diverse spatial perspectives, enabling RiverMamba to capture different contextual features.

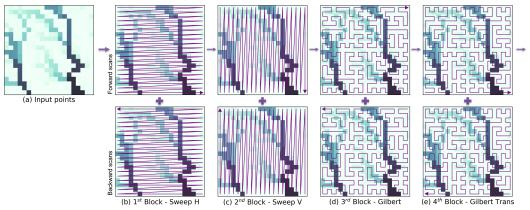

Figure 3: Illustration of the spatial scans in RiverMamba. Larger images are in the supp. material.

Location-aware adaptive normalization layer. In order to condition the model on static river attributes $\mathbf{X}_{static}$, the location-aware adaptive normalization layer (LOAN) [91] modulates the features $\mathbf{X}$ within the hindcast block:

$$\text{LOAN}(\mathbf{X}) = \left(\frac{\mathbf{X} - \mu}{\sigma}\right) + \text{GELU}(\text{Linear}(\mathbf{X}_{static}))\,, \tag{2}$$

where a linear layer projects $\mathbf{X}_{static} \in \mathbb{R}^{B,1,P,V_s}$, with V_s being the number of static variables, to $\mathbb{R}^{B,1,P,K}$. The output is then duplicated along the temporal dimension so that the output has the dimension $\mathbb{R}^{B,T,P,K}$. $\mu \in \mathbb{R}^{B,T,P,1}$ and $\sigma \in \mathbb{R}^{B,T,P,1}$ are the mean and standard deviation of $\mathbf{X}$ along the channel dimension, respectively, and $\mathbf{X} \in \mathbb{R}^{B,T,P,K}$ is the input to the LOAN layer. Both μ and σ are duplicated K times along the last dimension to match $\mathbf{X}$. The layer normalizes the features

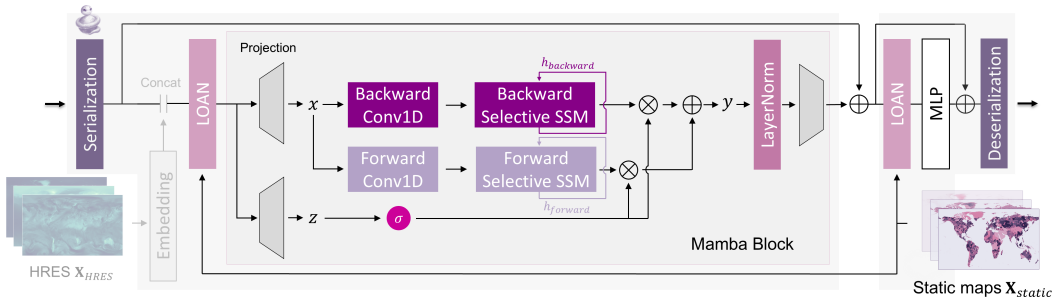

Figure 4: The structure of the hindcast block and forecast block. Both use a bidirectional Mamba block and the forecast block has the same structure as the hindcast block, but it additionally incorporates meteorological forecasts (HRES) by concatenation. The forecast block also includes LOAN layers although it is not shown in Fig. 2

and adds a systematic bias based on the attributes. For instance, the features are normalized and biased based on location attributes that have an impact on drainage and floods.

Mamba block. Fig. 4 shows a more detailed structure of the hindcast block with the elements of the Mamba block. After the input is serialized into a 1D sequence based on the block-specific space-filling curve and the features are normalized by the LOAN layer, the Mamba block processes the features of the sampled points along the sequence.

The Mamba block is based on a state-space model that transforms a 1D sequence of states $x(t)$ into another representation $y(t)$ through an implicit hidden latent state $h(t)$ and a first-order ordinary differential equation:

$$h'(t) = \mathbf{A}h(t-1) + \mathbf{B}x(t)\,, \quad y(t) = \mathbf{C}h(t) + \mathbf{D}x(t)\,. \tag{3}$$

To integrate Eq. (3) into a deep learning framework, S4 [26] parametrized the system with the matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ and discretized it with a timescale parameter Δ:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t\,, \quad y_t = \mathbf{C}h_t + \mathbf{D}x_t\,, \tag{4}$$

$$\bar{\mathbf{A}} = e^{(\Delta\mathbf{A})}\,, \quad \bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(e^{(\Delta\mathbf{A})} - \mathbf{I})\Delta\mathbf{B}\,, \tag{5}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are the discretized versions of the system. Recently, S6 [28] proposed to make Eqs. (4) and (5) time-variant. To this end, the parameters $\mathbf{B}(x)$, $\mathbf{C}(x)$, and $\Delta(x)$ become dependent on the input state x. This representation of a state-space model is called Mamba, which is an efficient alternative to transformers [92], particularly when processing many points as in our case.

Fig. 4 illustrates the steps of the Mamba block. The normalized sequence $\mathbf{X} \in \mathbb{R}^{B\times(T\times P)\times K}$ is projected into $\mathbf{x} \in \mathbb{R}^{B\times(T\times P)\times K}$ and $\mathbf{z} \in \mathbb{R}^{B\times(T\times P)\times K}$, where $T\times P$ is the length of the sequence. Note that the order of the elements in the sequence depends on the serialization, which differs between the hindcast blocks. We use a bi-directional approach that converts $\mathbf{x}$ into $\mathbf{x}'_o$ using a forward and a backward 1-D causal convolution, where $o \in \{f, b\}$ denotes the forward or backward pass. For each direction, $\mathbf{B}_o$, $\mathbf{C}_o$, and Δ_o are obtained by projection layers from $\mathbf{x}'_o$, and $\bar{\mathbf{A}}_o$ and $\bar{\mathbf{B}}_o$ are computed using Eq. (5). The selective SSM then uses Eq. (4) to obtain $\mathbf{y}_{forward}$ and $\mathbf{y}_{backward}$ for the forward and backward pass, respectively. The final output $\mathbf{y}$ is obtained by gating $\mathbf{y}_{forward}$ and $\mathbf{y}_{backward}$ via SiLU($\mathbf{z}$) and adding them up. Finally, $\mathbf{y}$ is normalized and projected back linearly to $\mathbb{R}^{B\times(T\times P)\times K}$. The complete algorithm for the Mamba block is described in the suppl. material. After the Mamba block, the hindcast block includes another LOAN layer followed by an MLP. The final output $\mathbf{X}$ is then deserialized at the end since the next hindcast block uses a different serialization.

Forecasting layer. While the hindcast layers encode the sequence of past input variables into a K-dimensional vector per sampled point, i.e., $\mathbf{X}_{hindcast} \in \mathbb{R}^{B\times 1\times P\times K}$, the forecasting layers forecast the difference of daily mean river discharge $\Delta\mathbf{X}^{t+l}_{dis24}$ for each lead time l, using $\mathbf{X}_{hindcast}$ and meteorological forecasts $\mathbf{X}^{t+1:t+L}_{HRES}$ as input, as shown in Fig. 2. The forecast blocks have the same structure as the hindcast blocks except that the forecast block incorporates the meteorological forcing (HRES). This is done by projecting $\mathbf{X}^{t+l}_{HRES}$ with a linear layer to 64 dimensions, serializing it, and concatenating it with the input $\mathbf{X}$ as illustrated in Fig. 4. The processing of HRES is done sequentially, i.e., we have L forecast blocks and the l-th forecast block processes $\mathbf{X}^{t+l}_{HRES}$. We argue

that this design is crucial to ensure that the temporal relationships between the meteorological forcing and the initial conditions are maintained.

The output of all forecast blocks is processed by L regression heads implemented as multi-layer perceptrons (MLP) where the output for the lead time $t+l$ is obtained as:

$$\Delta\mathbf{X}_{dis24}^{t+l} = \text{Linear}\Big(\text{ReLU}\Big(\text{Concat}\big(\text{Linear}(\mathbf{X}_{forecast}^{t+l}), \text{Linear}(\mathbf{X}_{forecast}^{t+1:t+L\setminus t+l})\big)\Big)\Big)\,, \tag{6}$$

where $\mathbf{X}_{forecast}^{t+l}$ are the features from the l-th forecast block and $\mathbf{X}_{forecast}^{t+1:t+L\setminus t+l}$ are the concatenated features from all forecast blocks except of the l-th block. The linear layers project the input $\mathbf{X}_{forecast}^{t+l}$ or $\mathbf{X}_{forecast}^{t+1:t+L\setminus t+l}$ to 32 dimensions and the last linear projection estimates finally $\Delta\mathbf{X}_{dis24}^{t+l} \in \mathbb{R}^{B\times 1\times P\times 1}$.

Training. As already mentioned, we sample P points around the globe for training. As a target value for training, we first use the river discharge data from the GloFAS reanalysis as ground truth and then fine-tune on sparse observations using data from the Global Runoff Data Centre (GRDC). For GRDC fine-tuning, we take P as the number of input points per sample and compute the loss only on points where GRDC observations are available without considering reanalysis data from GloFAS. We obtain the target values by $\Delta\hat{\mathbf{X}}_{dis24}^{t+l}(p) = \hat{\mathbf{X}}_{dis24}^{t+l}(p) - \hat{\mathbf{X}}_{dis24}^{t-1}(p)$, where $\hat{\mathbf{X}}$ are the values from GloFAS or GRDC. For the training loss, we propose a weighted version of the mean-squared error (MSE) loss:

$$\mathcal{L} = \frac{1}{B\times P\times L}\sum_{b=1}^{B}\sum_{p=1}^{P}\sum_{l=1}^{L} w^{b,t+l}(p)\|\Delta\hat{\mathbf{X}}_{dis24}^{b,t+l}(p) - \Delta\mathbf{X}_{dis24}^{b,t+l}(p)\|_2^2\,, \tag{7}$$

where B is the batch size. Since the severity of a flood is highly important for flood forecasting and severe floods occur rarely, the weighting factor $w^{b,t+l}(p)$ takes this into account. The severity of a flood is ranked by the statistical flood return period in years, which we denote by r and ranges from 1.5 to 500. These ranges are also used in GloFAS. We note that a high return period event simply reflects statistical rarity in streamflow magnitude, and should not be equated with a flood event without additional context, e.g., thresholds or inundation. The return period is used here as a proxy indicator of hydrological extremity, which we call flood. The severity of a flood is thus given by $\hat{r}^{t+l}(p) = \max_r \left\{r : \hat{\mathbf{X}}_{dis24}^{t+l}(p) \geq \theta_r(p)\right\}$, where θ_r is the statistical threshold for a given flood return period r. We also include the case $r{=}0$ with $\theta_r{=}0$ for defining events that are not floods. Using this notation, the weighting is thus given by

$$\hat{w}^{b,t+l}(p) = \begin{cases} \hat{r}^{b,t+l}(p) & \text{if} \quad \hat{r}^{b,t+l}(p) > 1 \\ 1 & \text{otherwise.} \end{cases} \tag{8}$$

We thus weight the loss based on the flood return period if a flood occurred at location p and time $t+l$, and we use 1 if there has been no flood. We further weight the loss with $\hat{u}^{b,t+l} = e^{\alpha(L-l+1)}$, where we give a higher weight to a shorter lead time l and use $\alpha{=}0.25$. This compensates for the sequential structure of the forecast blocks where each forecast block takes the features of the previous block as input. The final weight is thus given by $w^{b,t+l}(p) = \hat{u}^{b,t+l}\hat{w}^{b,t+l}(p)$. Since river discharge exhibits a very large dynamic with varying orders of magnitude, we transform the discharge values by $\text{sign}(\Delta\hat{x})\log(1+|\Delta\hat{x}|)$. We evaluate the impact of the weighting in Table 2 (a) and provide more details in the suppl. material. For inference, we can forecast floods for any set of points or densely as in Fig. 1.

4 Experimental results

Dataset. We obtain data for river discharge from the ECMWF GloFAS reanalysis [82]. It is generated by forcing the LISFLOOD hydrological model [93] using meteorological data from ERA5 [94]. GloFAS reanalysis combines physics-based simulation with observations to generate a consistent reconstruction of the past. The dataset is provided as a daily averaged discharge on a global coverage at 3 arcmin grid ($0.05°$). We use the GloFAS reanalysis as a target discharge for training and testing the model in Sec. 4.1. The ablation studies are done using GloFAS reanalysis over Europe. In addition, we fine-tune and test the model on observational GRDC river discharge data in Sec. 4.2. Flood thresholds are determined using return periods for individual points and are calculated from the long-term data. The thresholds allow for the identification of a flood when the threshold is surpassed.

Evaluation metrics. We evaluate the performance of RiverMamba on both GloFAS reanalysis and GRDC, where diagnostic GRDC stations are available (3366 stations). For evaluation, we use common metrics like the coefficient of determination (R2), Kling–Gupta efficiency (KGE), and the averaged F1-score for floods with return periods of 1.5 to 20 years. Details about these metrics can be found in the suppl. material. We train on the years 1979-2018, validate on 2019-2020, and test on 2021-2024. All evaluation points are gauged stations and temporally out-of-sample. Results on ungauged stations are also available in the suppl. material. The metrics are calculated on the time series at single grid points and then averaged over all points.

Baselines. We compare RiverMamba to persistence, climatology, and the state-of-the-art deep learning Encoder-Decoder LSTM of Google's operational flood forecasting system [24]. For the LSTM model, we followed the same protocol as originally proposed in [24], which considers only temporal context but does not include any spatial connections. The space filling curves are thus not used in combination with the LSTM baseline. To ensure a consistent evaluation, we train LSTM on the same input data as RiverMamba. All results in the paper are obtained with our trained LSTM. A comparison with the published reforecasts by Google's LSTM [24] is also available in the suppl. material. For evaluation on GRDC observations, we additionally compare our approach to the reforecast version of the state-of-the-art operational GloFAS forecasting system operated by ECMWF [15, 16]. More details about dataset, evaluation metrics and baselines are provided in the suppl. material.

4.1 Experiments on GloFAS river discharge reanalysis

The quantitative results are shown in Table 1. As can be seen, the climatology baseline performs poorly, as the dynamic in local river discharge varies a lot over time, highlighting the difficulty in predicting flows. We therefore exclude it in Fig. 5 (a) that shows F1-score for floods with a 1.5-year return period and (b) KGE for river discharge for different lead times from 24 to 168 hours. The boxes show distribution quartiles and the evaluation points are represented as points along the y-axis. Fig. 5 (d) shows the F1-score averaged over return periods of 1.5 to 20 years and (e) shows the median R2 for river discharge. The persistence baseline predicts the future discharge as the same value of the discharge at time t. This achieves good prediction for the short-term forecast, however, the prediction skill drops with lead time. While LSTM outperforms the persistence baseline, RiverMamba outperforms all baselines and methods on all metrics as shown in Table 1. In particular for lead times above 48 hours, the performance gap between RiverMamba and LSTM is large. We attribute this to the receptive field and the spatio-temporal modeling of RiverMamba. Fig. 5 (c) plots the F1-score averaged over 24 to 168 hours lead time for different flood return periods. The results show that RiverMamba outperforms the other approaches both for more frequent floods and rare severe floods that occur statistically only every 500 years. More results are in the suppl. material. In the following, we discuss a set of ablation studies that are not performed globally but over Europe.

Objective functions. In Table 2 (a), we evaluate the impact of the weighting factor in the loss (7), which is based on $\hat{w}$ Eq. (8) and $\hat{u}$. The results show that both terms improve the results. $\hat{w}$ is

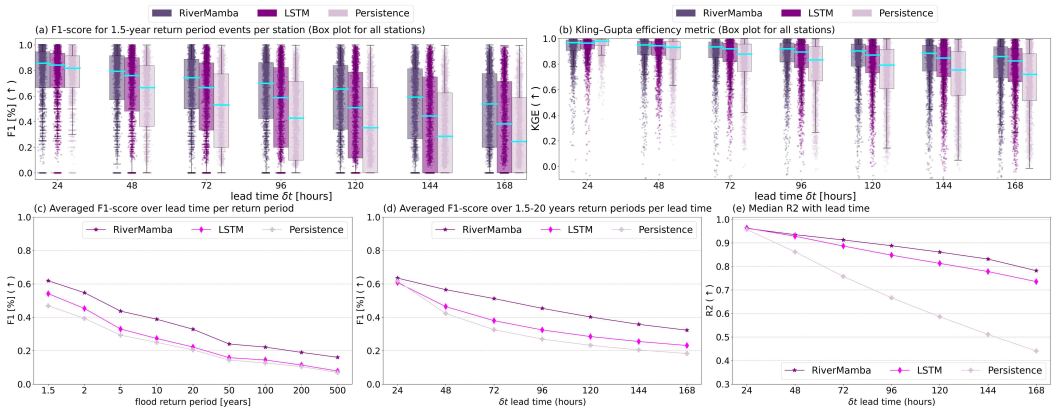

Figure 5: Results on GloFAS reanalysis across lead times and flood return periods.

Table 1: Results on GloFAS-Reanalysis. (±) denotes the standard deviation for 3 runs.

	Validation (2019-2020)			Test (2021-2024)		
Model	R2 (↑)	KGE (↑)	F1 (↑)	R2 (↑)	KGE (↑)	F1 (↑)
Climatology	0.1175	0.2618	–	0.1352	0.2449	–
Persistence	0.6778	0.8380	0.3138	0.6833	0.8412	0.3223
LSTM	0.8539±0.0031	0.8931±0.0034	0.3511±0.0068	0.8485±0.0021	0.8924±0.0029	0.3582±0.0058
RiverMamba	**0.8803**±0.0043	**0.9137**±0.0026	**0.4540**±0.0056	**0.8728**±0.0013	**0.9125**±0.0008	**0.4589**±0.0080

Table 2: Ablation studies on the validation set over Europe.

(a) Objective function

$\hat{w}$	$\hat{u}$	KGE | F1 (↑)
✗	✗	0.9086 | 0.2236
✓	✗	0.9127 | 0.2859
✗	✓	0.9136 | 0.2593
✓	✓	**0.9205** | **0.2875**

(b) Location Embedding

LOAN$_{(hind)}$	LOAN$_{(forc)}$	KGE | F1 (↑)
✗	✗	0.9183 | 0.2790
✓	✗	0.9160 | 0.2827
✗	✓	0.9166 | **0.2931**
✓	✓	**0.9205** | 0.2875

(c) Forecasting strategy

S-HRES	T-HRES	KGE | F1 (↑)
✗	✓	0.8862 | 0.2030
✓	✗	0.8869 | 0.2268
✓	✓	**0.9205** | **0.2875**

important to focus on rare and more severe floods, increasing the F1 metric substantially (second row). $\hat{u}$ gives more weight to the forecast in the near future where $\mathbf{X}_{HRES}$ is more reliable, which is important due to the sequential structure of the forecast module. Using only $\hat{u}$ (third row) improves the results on both KGE and F1 metrics. Using both $\hat{w}$ and $\hat{u}$ (fourth row) gives the best results.

Location embedding. In Table 2 (b), we show the benefit of using LOAN. In the first row, we duplicate the static features along the T dimension and concatenate them with the dynamic input. Using the LOAN layer in the hindcast (second row) or forecast blocks (third row) increases the F1 score but decreases KGE. Using LOAN in both hindcast and forecast blocks balances the metrics (fourth row).

Forecasting strategy. Table 2 (c) evaluates the impact of spatio-temporal modeling in the forecast module. In the first row, we remove the spatial relations in the forecast module by replacing the forecast blocks by point-wise MLPs. In this way, the data is processed after the last hindcast layer temporally but not spatially. This makes the model unaware of the spatial biases in the meteorological forcing $\mathbf{X}_{HRES}$. The second row denotes a setup where the forecast blocks do not get the features from the previous forecast block (Fig. 2) but directly from the last hindcast layer. In this case, we forecast river discharge for each lead time independently. The results show that in both cases the performance drops compared to our approach (third row), demonstrating the importance of spatio-temporal modeling. More ablation studies can be found in the suppl. material.

4.2 Experiments on GRDC observational river discharge

Table 3 reports the performance on GRDC river discharge observations at gauged stations, which also includes the physics-based GloFAS reforecast model. As previously, Fig. 6 compares the forecast performance across multiple lead times and flood return periods. Compared to the results on GloFAS reanalysis (Table 1), all models show a noticeable drop in performance when evaluated on GRDC observations (Table 3). This decline likely stems from the fact that GloFAS simulates primarily naturalized discharge, with simplified representations of major reservoirs [82, 95], whereas GRDC reflects fully regulated flow, influenced by complex and unobserved human activities, such as dam operations and irrigation. This introduces biases that models cannot learn, especially in the absence of globally available data representing human water management, highlighting the challenge of predicting discharge under human-modified conditions. The results show that traditional baselines such as Climatology and Persistence perform poorly. GloFAS performs much better than the baselines, but the R2 and KGE values are rather low due to the mentioned differences of physics-based models and observations. RiverMamba consistently outperforms the other methods for all metrics. Notably, RiverMamba shows less degradation in F1-score with increasing lead time, highlighting its strength in medium-range flood forecasting. More results are in the suppl. material.

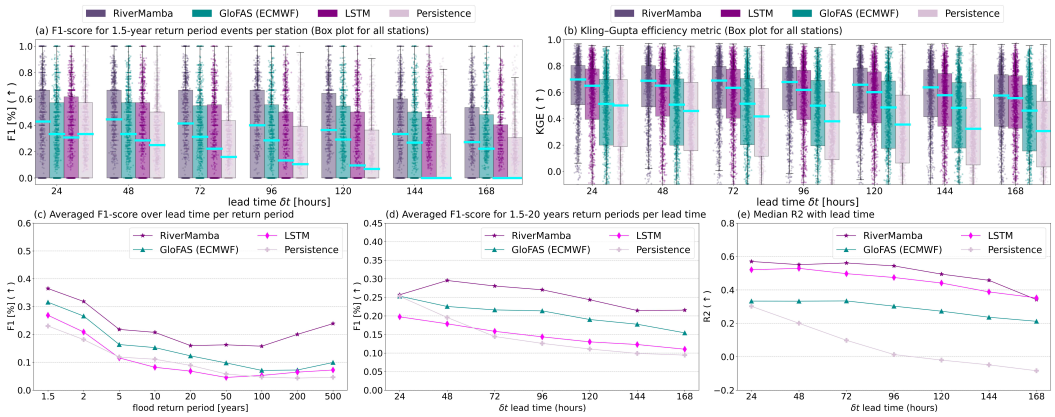

Figure 6: Results on gauged GRDC observations across lead times and flood return periods.

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

	Validation (2019-2020)			Test (2021-2023)		
Model	R2 (↑)	KGE (↑)	F1 (↑)	R2 (↑)	KGE (↑)	F1 (↑)
Climatology	-0.0002	0.1342	–	-0.0013	0.0870	–
Persistence	0.1682	0.4569	0.1626	0.0660	0.3918	0.1462
GloFAS	0.3713	0.5412	0.2135	0.2892	0.4944	0.2044
LSTM	0.5437±0.0025	0.6572±0.0010	0.1724±0.0017	0.4615±0.0039	0.6141±0.0018	0.1475±0.0014
RiverMamba	**0.5943**±0.0016	**0.7015**±0.0007	**0.2577**±0.0046	**0.5057**±0.0028	**0.6612**±0.0010	**0.2427**±0.0111

5 Conclusions and limitations

We introduced RiverMamba, a novel deep learning approach for global, medium-range river discharge and flood forecasting. Due to its efficient structure and specialized scanning paths, RiverMamba maintains a very large receptive field, while scaling linearly with respect to the number of sampled points. As a result, RiverMamba is capable of forecasting high-resolution ($0.05°$) global river discharge maps. Further, the spatio-temporal modeling of the forecast blocks incorporates meteorological forcing and ensures a consistent forecast through space and time. Our analysis reveals that RiverMamba outperforms operational state-of-the-art deep learning and physics-based models on both reanalysis and observational data. While the results show major advancements in river discharge and flood forecasting, the approach has some limitations. For a real operational setting, only data can be used that is available until the current day t. For instance, ERA5-Land is publicly available after 5 days whereas we assumed that ERA5-Land is already available after 1 day, i.e., $t - 1$. ERA5-Land, however, could be substituted by other near real-time reanalysis data that is earlier available or analysis data until day t. It also needs to be mentioned that observational data are affected by human interventions like dams and there is a need to integrate such interventions in the model. As it is the case for operational systems, floods are not always correctly forecast. The causes of the errors need to be analyzed more in detail. The forecast errors can be caused by human interventions, errors in the weather forecast for meteorological forcing or river attributes, the rarity of floods, or bias in the data and re-analysis. Given such errors, it is desirable to extend the model such that it estimates its uncertainty for the forecast as well.

Besides these limitations, RiverMamba has the potential for an operational medium-range river discharge and flood forecasting system that predicts flood risks, in particular extreme floods, more accurately and at higher resolution than existing systems. This is essential for stakeholders to make decisions for an effective flood risk mitigation strategy and an early warning system to protect citizens.

6 Acknowledgments and Disclosure of Funding

This work was supported by the Federal Ministry of Research, Technology, and Space under grant no. 16IS24075C RAINA and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1502/1–2022 – project no. 450058266 within the Collaborative Research Center (CRC) for the project Regional Climate Change: Disentangling the Role of Land Use and Water Management (DETECT). We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, through EuroHPC Regular Access Call - proposal No. EHPC-REG-2025R01-218. The authors also gratefully acknowledge the granted access to the Marvin cluster hosted by the University of Bonn. Furthermore, we acknowledge the ESM Testprojekt project for supporting this study by providing computing time on the esmtst partition of the supercomputer JUWELS-BOOSTER at Jülich Supercomputing Centre (JSC). Finally, we thank Lars Doorenbos for proofreading the manuscript. The Mamba logos were generated with Microsoft Designer.

References

[1] Francesco Dottori, Wojciech Szewczyk, Juan-Carlos Ciscar, Fang Zhao, Lorenzo Alfieri, Yukiko Hirabayashi, Alessandra Bianchi, Ignazio Mongelli, Katja Frieler, Richard A Betts, et al. Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change*, 8(9):781–786, 2018.

[2] Bruno Merz, Günter Blöschl, Sergiy Vorogushyn, Francesco Dottori, Jeroen C. J. H. Aerts, Paul Bates, Miriam Bertola, Matthias Kemter, Heidi Kreibich, Upmanu Lall, et al. Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment*, 2(9):592–609, 2021.

[3] Heidi Kreibich, Anne F Van Loon, Kai Schröter, Philip J Ward, Maurizio Mazzoleni, Nivedita Sairam, Guta Wakbulcho Abeshu, Svetlana Agafonova, Amir AghaKouchak, Hafzullah Aksoy, et al. The challenge of unprecedented floods and droughts in risk management. *Nature*, 608(7921):80–86, 2022.

[4] Jun Rentschler, Melda Salhab, and Bramka Arga Jafino. Flood exposure and poverty in 188 countries. *Nature communications*, 13(1):3527, 2022.

[5] Myanna Lahsen and Jesse Ribot. Politics of attributing extreme events and disasters to climate change. *WIREs Climate Change*, 13(1):e750, 2022.

[6] S. Jiang, E. Bevacqua, and J. Zscheischler. River flooding mechanisms and their changes in europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, 26(24):6339–6359, 2022.

[7] Shijie Jiang, Larisa Tarasova, Guo Yu, and Jakob Zscheischler. Compounding effects in flood drivers challenge estimates of extreme river floods. *Science Advances*, 10(13):eadl4005, 2024.

[8] Florian Pappenberger, Hannah L. Cloke, Dennis J. Parker, Fredrik Wetterhall, David S. Richardson, and Jutta Thielen. The monetary benefit of early flood warnings in europe. *Environmental Science & Policy*, 51:278–291, 2015.

[9] Gustau Camps-Valls, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, Adrian Höhl, Andrea Castelletti, Aytac Pacal, Claire Robin, Francesco Martinuzzi, Ioannis Papoutsis, Ioannis Prapas, et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16(1):1919, 2025.

[10] Rebecca E. Emerton, Elisabeth M. Stephens, Florian Pappenberger, Thomas C. Pagano, Albrecht H. Weerts, Andy W. Wood, Peter Salamon, James D. Brown, Niclas Hjerdt, Chantal Donnelly, Calum A. Baugh, and Hannah L. Cloke. Continental and global scale flood forecasting systems. *WIREs Water*, 3(3):391–418, 2016.

[11] H. A. P. Hapuarachchi, M. A. Bari, A. Kabir, M. M. Hasan, F. M. Woldemeskel, N. Gamage, P. D. Sunter, X. S. Zhang, D. E. Robertson, J. C. Bennett, and P. M. Feikema. Development of a national 7-day ensemble streamflow forecasting service for australia. *Hydrology and Earth System Sciences*, 26(18):4801–4821, 2022.

[12] 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.

[13] S. Nevo, E. Morin, A. Gerzi Rosenthal, A. Metzger, C. Barshai, D. Weitzner, D. Voloshin, F. Kratzert, G. Elidan, G. Dror, G. Begelman, G. Nearing, G. Shalev, H. Noga, I. Shavitt, L. Yuklea, M. Royz, N. Giladi, N. Peled Levi, O. Reich, O. Gilon, R. Maor, S. Timnat, T. Shechter, V. Anisimov, Y. Gigi, Y. Levin, Z. Moshe, Z. Ben-Haim, A. Hassidim, and Y. Matias. Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, 26(15):4013–4032, 2022.

[14] Husain Najafi, Pallav Kumar Shrestha, Oldrich Rakovec, Heiko Apel, Sergiy Vorogushyn, Rohini Kumar, Stephan Thober, Bruno Merz, and Luis Samaniego. High-resolution impact-based early warning system for riverine flooding. *Nature communications*, 15(1):3726, 2024.

[15] L. Alfieri, P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger. Glofas - global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3):1161–1175, 2013.

[16] 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.

[17] Anne Jones, Julian Kuehnert, Paolo Fraccaro, Ophélie Meuriot, Tatsuya Ishikawa, Blair Edwards, Nikola Stoyanov, Sekou L Remy, Kommy Weldemariam, and Solomon Assefa. Ai for climate impacts: applications in flood risk. *npj Climate and Atmospheric Science*, 6(1):63, 2023.

[18] Markus Reichstein, Vitus Benson, Jan Blunk, Gustau Camps-Valls, Felix Creutzig, Carina J Fearnley, Boran Han, Kai Kornhuber, Nasim Rahaman, Bernhard Schölkopf, et al. Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications*, 16(1):2564, 2025.

[19] 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. e2020WR028091 10.1029/2020WR028091.

[20] J. M. Frame, F. Kratzert, D. Klotz, M. Gauch, G. Shalev, O. Gilon, L. M. Qualls, H. V. Gupta, and G. S. Nearing. Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13):3377–3392, 2022.

[21] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[22] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

[23] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[24] 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.

[25] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *Advances in Neural Information Processing Systems*, volume 34, pages 572–585. Curran Associates, Inc., 2021.

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

[27] Jimmy T.H. Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[29] 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.

[30] J. Green, I. D. Haigh, N. Quinn, J. Neal, T. Wahl, M. Wood, D. Eilander, M. de Ruiter, P. Ward, and P. Camus. Review article: A comprehensive review of compound flooding literature with a focus on coastal and estuarine regions. *Natural Hazards and Earth System Sciences*, 25(2):747–816, 2025.

[31] T. Cache, M. S. Gomez, T. Beucler, J. Blagojevic, J. P. Leitao, and N. Peleg. Enhancing generalizability of data-driven urban flood models by incorporating contextual information. *Hydrology and Earth System Sciences*, 28(24):5443–5458, 2024.

[32] Julian Hofmann and Holger Schüttrumpf. floodgan: Using deep adversarial learning to predict pluvial flooding in real time. *Water*, 13(16), 2021.

[33] Tim Busker, Bart van den Hurk, Hans de Moel, and Jeroen CJH Aerts. The value of precipitation forecasts to anticipate floods. *Bulletin of the American Meteorological Society*, 2025.

[34] Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

[35] R. Bentivoglio, E. Isufi, S. N. Jonkman, and R. Taormina. Deep learning methods for flood mapping: a review of existing applications and future research directions. *Hydrology and Earth System Sciences*, 26(16):4345–4378, 2022.

[36] Zifeng Guo, João P. Leitão, Nuno E. Simões, and Vahid Moosavi. Data-driven flood emulation: Speeding up urban flood predictions by deep convolutional neural networks. *Journal of Flood Risk Management*, 14(1):e12684, 2021.

[37] Alexander Y. Sun, Zhi Li, Wonhyun Lee, Qixing Huang, Bridget R. Scanlon, and Clint Dawson. Rapid flood inundation forecast using fourier neural operator. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3735–3741, 2023.

[38] Omar Seleem, Georgy Ayzel, Arthur Costa Tomaz de Souza, Axel Bronstert, and Maik Heistermann and. Towards urban flood susceptibility mapping using data-driven models in berlin, germany. *Geomatics, Natural Hazards and Risk*, 13(1):1640–1662, 2022.

[39] Benjamin Burrichter, Julian Hofmann, Juliana Koltermann da Silva, Andre Niemann, and Markus Quirmbach. A spatiotemporal deep learning approach for urban pluvial flood forecasting with multi-source data. *Water*, 15(9), 2023.

[40] Haibo Chu, Wenyan Wu, Q.J. Wang, Rory Nathan, and Jiahua Wei. An ann-based emulation modelling framework for flood inundation modelling: Application, challenges and future directions. *Environmental Modelling & Software*, 124:104587, 2020.

[41] Fazlul Karim, Mohammed Ali Armin, David Ahmedt-Aristizabal, Lachlan Tychsen-Smith, and Lars Petersson. A review of hydrodynamic and machine learning approaches for flood inundation modeling. *Water*, 15(3), 2023.

[42] Matteo Pianforini, Susanna Dazzi, Andrea Pilzer, and Renato Vacondio. Floodsformer: A transformer-based data-driven model for predicting the 2-d dynamics of fluvial floods. *Environmental Modelling Software*, 193:106599, 2025.

[43] Björn Lütjens, Brandon Leshchinskiy, Océane Boulais, Farrukh Chishtie, Natalia Díaz-Rodríguez, Margaux Masson-Forsythe, Ana Mata-Payerro, Christian Requena-Mesa, Aruna Sankaranarayanan, Aaron Piña, Yarin Gal, Chedy Raïssi, Alexander Lavin, and Dava Newman. Generating physically-consistent satellite imagery for climate visualizations. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024.

[44] Saeid Janizadeh, Subodh Chandra Pal, Asish Saha, Indrajit Chowdhuri, Kourosh Ahmadi, Sajjad Mirzaei, Amir Hossein Mosavi, and John P. Tiefenbacher. Mapping the spatial and temporal variability of flood hazard affected by climate and land-use changes in the future. *Journal of Environmental Management*, 298:113551, 2021.

[45] Antara Dasgupta, Renaud Hostache, RAAJ Ramsankaran, Guy J.-P. Schumann, Stefania Grimaldi, Valentijn R. N. Pauwels, and Jeffrey P. Walker. A mutual information-based likelihood function for particle filter flood extent assimilation. *Water Resources Research*, 57(2):e2020WR027859, 2021. e2020WR027859 2020WR027859.

[46] Nikolaos Ioannis Bountos, Maria Sdraka, Angelos Zavras, Ilektra Karasante, Andreas Karavias, Themistocles Herekakis, Angeliki Thanasou, Dimitrios Michail, and Ioannis Papoutsis. Kuro siwo: 33 billion m^2 under the water. a global multi-temporal satellite dataset for rapid flood mapping. In *Advances in Neural Information Processing Systems*, volume 37, pages 38105–38121. Curran Associates, Inc., 2024.

[47] Wahid Palash, Ali S Akanda, and Shafiqul Islam. A data-driven global flood forecasting system for medium to large rivers. *Scientific Reports*, 14(1):8979, 2024.

[48] Liangjin Zhong, Huimin Lei, Zhiyuan Li, and Shijie Jiang. Advancing streamflow prediction in data-scarce regions through vegetation-constrained distributed hybrid ecohydrological models. *Journal of Hydrology*, 645:132165, 2024.

[49] 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.

[50] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[51] 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.

[52] Benedikt Heudorfer, Hoshin V. Gupta, and Ralf Loritz. Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, 52(6):e2024GL113036, 2025. e2024GL113036 2024GL113036.

[53] Yikui Zhang, Silvan Ragettli, Peter Molnar, Olga Fink, and Nadav Peleg. Generalization of an encoder-decoder lstm model for flood prediction in ungauged catchments. *Journal of Hydrology*, 614:128577, 2022.

[54] Binlan Zhang, Chaojun Ouyang, Peng Cui, Qingsong Xu, Dongpo Wang, Fei Zhang, Zhong Li, Linfeng Fan, Marco Lovati, Yanling Liu, and Qianqian Zhang. Deep learning for cross-region streamflow and flood forecasting at a global scale. *The Innovation*, 5(3):100617, 2024.

[55] 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.

[56] Yihan Wang, Lujun Zhang, N. Benjamin Erichson, and Tiantian Yang. Investigating the streamflow simulation capability of a new mass-conserving long short-term memory (mc-lstm) model across the contiguous united states. *Journal of Hydrology*, 658:133161, 2025.

[57] E. Acuña Espinoza, F. Kratzert, D. Klotz, M. Gauch, M. Álvarez Chaves, R. Loritz, and U. Ehret. Technical note: An approach for handling multiple temporal frequencies with different input dimensions using a single lstm cell. *Hydrology and Earth System Sciences*, 29(6):1749–1758, 2025.

[58] 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. e2025GL115705 2025GL115705.

[59] 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.

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

[61] Naghmeh Shafiee Roudbari, Shubham Rajeev Punekar, Zachary Patterson, Ursula Eicker, and Charalambos Poullis. From data to action in flood forecasting leveraging graph neural networks and digital twin visualization. *Scientific reports*, 14(1):18571, 2024.

[62] Arnold Kazadi, James Doss-Gollin, Antonia Sebastian, and Arlei Silva. Floodgnn-gru: a spatio-temporal graph neural network for flood prediction. *Environmental Data Science*, 3:e21, 2024.

[63] 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. e2024WR039764 2024WR039764.

[64] M. A. Vischer, N. Otero, and J. Ma. Spatially resolved rainfall streamflow modeling in central europe. *EGUsphere*, 2025:1–26, 2025.

[65] 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. e2023WR035337 2023WR035337.

[66] 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. e2024WR038928 2024WR038928.

[67] 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, 2024. e2023WR036170 2023WR036170.

[68] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pages 103031–103063. Curran Associates, Inc., 2024.

[69] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024.

[70] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[71] Hao Phung, Quan Dao, Trung Dao, Hoang Phan, Dimitris N. Metaxas, and Anh Tran. Dimsum: Diffusion mamba - a scalable and unified spatial-frequency method for image generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 32947–32979. Curran Associates, Inc., 2024.

[72] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model. In *European Conference on Computer Vision (ECCV)*, pages 148–166, Cham, 2025. Springer Nature Switzerland.

[73] Yicheng Xiao, Lin Song, Shaoli Huang, Jiangshan Wang, Siyu Song, Yixiao Ge, Xiu Li, and Ying Shan. Mambatree: Tree topology is all you need in state space model. In *Advances in Neural Information Processing Systems*, volume 37, pages 75329–75354. Curran Associates, Inc., 2024.

[74] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. Group-mamba: Efficient group-based visual state space model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14912–14922, June 2025.

[75] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision (ECCV)*, pages 237–255, Cham, 2025. Springer Nature Switzerland.

[76] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024.

[77] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision (ECCV)*, pages 265–282, Cham, 2025. Springer Nature Switzerland.

[78] Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, and Juergen Gall. Manta: Diffusion mamba for efficient and effective stochastic long-term dense action anticipation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[79] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. In *Advances in Neural Information Processing Systems*, volume 37, pages 32653–32677. Curran Associates, Inc., 2024.

[80] Guowen Zhang, Lue Fan, Chenhang He, Zhen Lei, Zhaoxiang Zhang, and Lei Zhang. Voxel mamba: Group-free state space models for point cloud based 3d object detection. In *Advances in Neural Information Processing Systems*, volume 37, pages 81489–81509. Curran Associates, Inc., 2024.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] Mingyue Chen, Wei Shi, Pingping Xie, Viviane BS 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), 2008.

[86] 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.

[87] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[88] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[89] Giuseppe Peano and G Peano. *Sur une courbe, qui remplit toute une aire plane*. Springer, 1990.

[90] 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*, pages 1–2, 1935.

[91] 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.

[92] 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.

[93] 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.

[94] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[95] Zuzanna Zajac, Beatriz Revilla-Romero, Peter Salamon, Peter Burek, Feyera A. Hirpa, and Hylke Beck. The impact of lake and reservoir parameterization on global streamflow simulation. *Journal of Hydrology*, 548:552–568, 2017.

[96] Mohamad Hakam Shams Eddin, Yikui Zhang, Stefan Kollet, and Juergen Gall. RiverMamba: A State Space Model for Global River Discharge and Flood Forecasting [data set], 2025.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: see Sec. 4.1 and 4.2.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: see Sec. 5 and supplementary material.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

Justification: the paper dose not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: see Sec. 3 and supplementary material.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: the code for the approach, processing the data, and evaluation are made publicly available. The code of RiverMamba, processing scripts, and pretrained models are available on GitHub at https://github.com/HakamShams/RiverMamba_code. The pre-processed data used in the study and RiverMamba reforecasts are available at https://doi.org/10.60507/FK2/T8QYWE [96]. GRDC data that has been used in this study is available for researchers after signing a license agreement with the owner of the data. Instructions on how the data can be obtained and used are provided in the source code.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: see supplementary material.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: see Tables 1 and 3 and Figs. 5 and 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: see supplementary material.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: the research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: see Introduction, Sec. 5, and supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: we think the paper does not pose such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: we cite the original publications for the raw data and refer to the URLs in the supplementary material when applicable.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: see supplementary material.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: we used LLMs for very limited editing (e.g., grammar, spelling, word choice). In addition, we used an LLM to generate the RiverMamba logo.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.