



Supplementary Material

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

Table of Contents

A	Dataset	3
A.1	GloFAS reanalysis data	3
A.2	GRDC observational river discharge data	3
A.3	ERA5-Land data	5
A.4	HRES	5
A.5	CPC data	6
A.6	LISFLOOD static features	7
A.7	Diagnostic river points	7
B	Return periods and flood definition	9
C	Implementation and training details	9
D	Mamba Block	11
E	Evaluation metrics	12
F	Ablation studies	14

F.1	Mamba vs. Transformer	14
F.2	Feature importance	14
F.3	Pretraining on reanalysis	15
F.4	Space-filling curves	15
F.5	Weighting in the objective function	16
F.6	Activation function in LOAN	16
G	Computational time	17
H	Baselines	18
H.1	Climatology	18
H.2	Persistence	18
H.3	LSTM	18
H.4	GloFAS Forecast	19
H.5	GloFAS Reforecast	19
I	Space-filling curves	20
J	Experiments on HydroRIVERS	23
K	Additional results	24
K.1	Comparison with Google reforecast on ungauged GRDC	24
K.2	Comparison with Google reforecast on gauged GRDC	27
K.3	Additional results on gauged GloFAS reanalysis	30
K.4	Additional results on gauged GRDC	39
K.5	Comparison to operational GloFAS forecasts on gauged GRDC	48
L	Case studies of extreme flood events	49
L.1	2021 Western Europe flood	49
L.2	2024 Southeast Europe floods	50
L.3	2024 Central European floods	51
L.4	2024 Spanish floods	51
L.5	2024 Saarland Germany flood	52
L.6	2024 Kenya-Tanzania flood	52
L.7	2024 California flood	53
L.8	2024 Central-South China floods	53
M	Code and data availability	54
N	Broader impacts	54

A Dataset

A.1 GloFAS reanalysis data

The Global Flood Awareness System (GloFAS) is an operational system developed by the European Commission’s Joint Research Centre (JRC) and operated by ECMWF under the Copernicus Emergency Management Service (CEMS) [1]. It provides real-time global-scale flood forecasts and a long-term hydrological reanalysis dataset, a key resource for flood risk assessment, climate impact studies, and machine learning applications. Fig. 1 shows the workflow of GloFAS to forecast river discharge and flood events. The GloFAS-ERA5 reanalysis is the long-term retrospective component of GloFAS [2]. It delivers daily river discharge estimates from 1979 to present at a spatial resolution of 0.05° (~ 5 km) and a global coverage (90°N - 60°S , 180°W - 180°E). The reanalysis is generated by coupling surface and subsurface runoff from the ERA5 reanalysis, produced by the H-TESSEL and surface model [3] with the LISFLOOD hydrological and river routing model [4]. While ERA5 runoff is computed at ~ 31 km resolution and lacks spatial connectivity, it is downscaled to 0.05° using a nearest-neighbour approach and routed through LISFLOOD to simulate realistic river discharge (dis_{24} , in m^3s^{-1}) across the global river network. The daily GloFAS reanalysis discharge data represents the mean value between 00:00 UTC previous day and 00:00 UTC current day. Similarly to GloFAS, there exists an early warning system for Europe (EFAS) with higher resolution [5]. In our work, GloFAS v4.0 is used for a global application. The dataset is publicly available on Climate Data Store and Early Warning Data Store (EWDs) <https://doi.org/10.24381/cds.a4fdd6b9>. Table 1 explains the details of four variables we took from the GloFAS reanalysis dataset as the model inputs. The GloFAS-ERA5 reanalysis supports the derivation of flood thresholds (i.e., 2-, 5-, and 20-year return periods) and serves as the initial condition for real-time forecasts such as GloFAS-30d and GloFAS-Seasonal. More details about GloFAS can be found in [2].

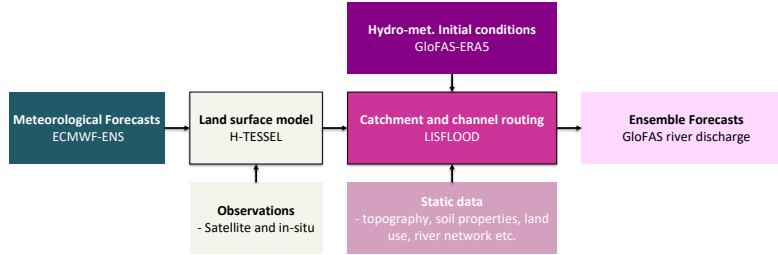


Figure 1: An overview of the key modules in the GloFAS forecasting system. GloFAS-ERA5 reanalysis uses ERA5 meteorological reanalysis data instead of ECMWF ensemble forecasts (ENS). Figure outline from [2].

Despite its broad applicability, the GloFAS-ERA5 reanalysis is subject to several limitations that researchers should be aware of. Regional biases have been identified, which may stem from uncertainties in the meteorological forcing provided by ERA5, the representation of runoff generation processes within the H-TESSEL land surface model, and limitations in the calibration of the LISFLOOD routing model. When sufficient observational discharge data are available, the LISFLOOD model is calibrated locally for each river catchment larger than 500 km^2 , and each calibrated catchment has its own optimized parameter set. This could give the model better performance at the local scale but reduce its generalization ability. Additionally, anthropogenic influences such as dams and reservoirs are incorporated using simplified operational rules, largely due to the lack of globally available real-time release data. Finally, as the dataset is entirely driven by ERA5, it inherits known deficiencies of the reanalysis, including biases in precipitation and the absence of river discharge data assimilation, which may affect the realism of simulated hydrological conditions in some regions.

A.2 GRDC observational river discharge data

We obtain observational river discharge from the Global Runoff Data Centre (GRDC) which is an international data repository that provides access to quality-controlled river discharge observation

Table 1: Details about the processed variables from GloFAS reanalysis [2].

Variable	Long name	Unit	Height	Surface parameters
acc_rod24	runoff water equivalent	kg/m ²	surface and subsurface	accumulated
dis24	river discharge in the last 24 hours	m ³ /s	surface	averaged over 24 hours
sd	snow depth water equivalent	kg/m ²	surface	instantaneous
swi	soil wetness index	-	root zone	instantaneous

data from around the world. The GRDC dataset contains time series of daily and monthly river discharge data from over 10000 hydrological gauging stations across more than 160 countries from small headwater catchments ($\sim 10 \text{ km}^2$ drainage area) to very large river catchment like the Amazon river (5 million km^2 drainage area). GRDC data can be obtained from <https://grdc.bafg.de/>. All GRDC daily time series measured the value set at 00:00 of the beginning of the day (left-labeled). To keep our evaluation consistent with [6], we used the GRDC dataset as the benchmark to evaluate the model performance and followed a similar data processing workflow as in [7, 6]. We first removed the catchments with a drainage area smaller than 500 km^2 and obtained 5524 GRDC stations to avoid very big discrepancies between the drainage area defined in GRDC and in the GloFAS dataset (A.1). Next, we geo-located the GRDC stations to compare them with the GloFAS drainage network and removed the GRDC stations with more than 10% of drainage area differences. For geo-location, we projected the points on the GloFAS grid, compared each point with its 9 nearest points, and took the location with the highest KGE value. Finally, the GRDC stations with no ERA5-Land reanalysis data were discarded. This resulted in 3366 stations for the global evaluation. This narrowed down the global median drainage area difference to 2.21% with an interquartile range of 0.86% to 4.73%. The discharge observations are recorded at a daily time scale, with the unit $\text{m}^3 \text{s}^{-1}$ and converted from local time zone to 00:00 UTC via linear interpolation. For evaluation, the GRDC observational time series, which are originally left-labeled, were explicitly converted to right-labeled time series to ensure a temporal consistency with the right-labeled predictions from the GloFAS simulations, RiverMamba, LSTM baseline, and Google reforecast. The F1 scores reported for the GRDC evaluation are based on synchronized detection windows between model predictions and observations.

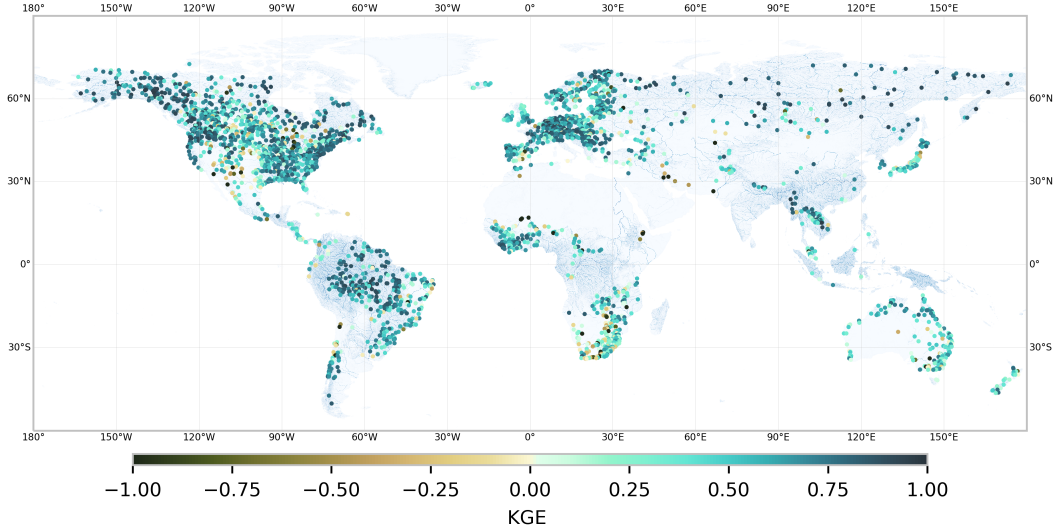


Figure 2: Locations of selected 3366 GRDC stations used for training and evaluating the RiverMamba model. The colorbar shows the KGE value of GloFAS reanalysis discharge data against the GRDC discharge observations.

Fig. 2 shows the KGE values of GloFAS reanalysis data against the GRDC observation at the locations of all 3366 selected stations. In general, there is a good agreement between GloFAS and GRDC data globally, with a median KGE at 0.61 and an interquartile range of 0.36 to 0.77. In regions like south

America, south Africa and Australia, the GloFAS reanalysis data have more inconsistency compared to the observations. Fig. 3 shows an exemplar hydrograph for a gauged station.

It is important to note that, compared to the GRDC observations, the GloFAS reanalysis dataset only simulates the naturalized flow without considering realistic human interventions such as dams, reservoirs, diversions, irrigation withdrawals, and other water management practices, and this can be a major source of bias in GloFAS compared to the GRDC data.

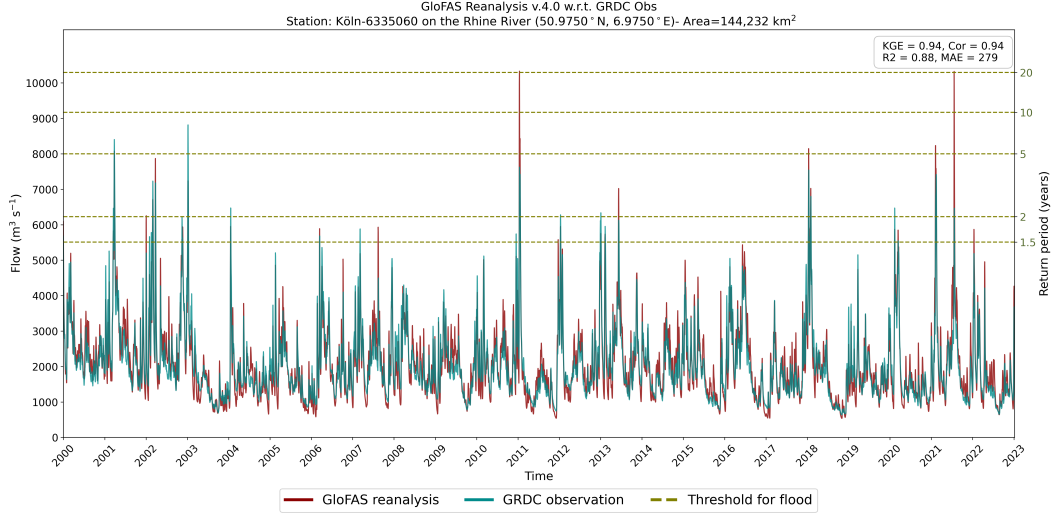


Figure 3: Hydrograph for GloFAS-reanalysis (red line) from 1 January 2000 to 31 December 2023 and observations (dark cyan line), for a gauging station on the Rhine River. The top-right box displays summary statistics from the reanalysis’s evaluation against the observations.

A.3 ERA5-Land data

In contrast to operational GloFAS which uses ERA5 as forcing data, we use ERA5-Land as an initial land surface condition for the forecast. The ERA5-Land reanalysis data are described in [8] and retrieved from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS). <https://doi.org/10.24381/cds.e2161bac>. We processed 14 instantaneous state variables at 00:00 UTC and 18 daily accumulated state variables (00:00 UTC previous day to 00:00 UTC current day). More details are provided in Table 2. ERA5-Land is provided at $0.1^\circ \times 0.1^\circ$. We mapped the data onto the GloFAS regular latitude and longitude (Plate Carrée projection) using bilinear mapping and implemented by Zhuang et al. [9].

A.4 HRES

Meteorological conditions serve as the driving forces behind hydrological processes. These are necessary for forecasting the river discharge and potential floods. We use the deterministic forecast of the ECMWF Integrated Forecast System (IFS) High Resolution (HRES) atmospheric model. The HRES data were obtained from the ECMWF Archive Catalogue <https://www.ecmwf.int/en/forecasts/dataset/operational-archive>. We use HRES up to 7 days lead time and once per day at 00:00 UTC. The processed data are similar to [6] except that we used total evaporation as an additional forcing variable. In addition, we do not use any forecast for nowcasting at time step t . The processed data include 2 instantaneous and 5 daily accumulated variables at the surface level forecasts. Technical details regarding the meteorological forcing variables are provided in Tables 3.

While the operational archive provides data from 1985, the quality and the resolution of the forecasts in the earlier years are not sufficient for our application. Therefore, we only processed and used data from 2010 to 2024. Data before 2010 were replaced by ERA5-Land. To match the resolution of the target GloFAS grid, we regridded HRES to $0.05^\circ \times 0.05^\circ$ regular grid.

Table 2: Details about the processed variables from ERA5-Land reanalysis [8].

Variable	Long name	Unit	Height	Surface parameters
d2m	2m dewpoint temperature	K	2m	instantaneous
e	total evaporation	m of water equivalent	surface	accumulated
es	snow evaporation	m of water equivalent	surface	accumulated
evabs	evaporation from bare soil	m of water equivalent	surface	accumulated
evaow	evaporation from open water surfaces excluding oceans	m of water equivalent	surface	accumulated
evatc	evaporation from the top of canopy	m of water equivalent	surface	accumulated
evavt	evaporation from vegetation transpiration	m of water equivalent	surface	accumulated
lai_hv	leaf area index high vegetation	m ² /m ²	2m	instantaneous
lai_lv	leaf area index low vegetation	m ² /m ²	2m	instantaneous
pev	potential evaporation	m	2m	accumulated
sf	snowfall	m of water equivalent	surface	accumulated
skt	skin temperature	K	surface	instantaneous
slhf	surface latent heat flux	J/m ²	surface	accumulated
smlt	snowmelt	m of water equivalent	surface	accumulated
sp	surface pressure	Pa	surface	instantaneous
src	skin reservoir content	m of water equivalent	surface	instantaneous
sro	surface runoff	m	surface	accumulated
sshf	surface sensible heat flux	J/m ²	surface	accumulated
ssr	surface net solar radiation	J/m ²	surface	accumulated
ssrd	surface solar radiation downwards	J/m ²	surface	accumulated
ssro	subsurface runoff	m	subsurface	accumulated
stl1	soil temperature	K	soil layer (0 - 7 cm)	instantaneous
str	surface net thermal radiation	J/m ²	surface	accumulated
strd	surface thermal radiation downwards	J/m ²	surface	accumulated
swvl1	volumetric soil water	m ³ /m ³	soil layer (0 - 7 cm)	instantaneous
swvl2	volumetric soil water	m ³ /m ³	soil layer (7 - 28 cm)	instantaneous
swvl3	volumetric soil water	m ³ /m ³	soil layer (28 - 100 cm)	instantaneous
swvl4	volumetric soil water	m ³ /m ³	soil layer (100 - 289 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

Table 3: Details about the processed variables from the ECMWF Integrated Forecast System (IFS) High Resolution (HRES) atmospheric model.

Variable	Long name	Unit	Height	Surface parameters
e	total evaporation	m of water equivalent	surface	accumulated
sf	snowfall	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
t2m	2m temperature	K	2m	instantaneous
tp	total precipitation	m	surface	accumulated

A.5 CPC data

Relying solely on the precipitation products from ERA5-Land reanalysis makes the model prone to the biases of the data assimilation which was used to derive the reanalysis. Similar to [6], we use precipitation estimates as observational input from the National Oceanic and Atmospheric

Administration (NOAA), Climate Prediction Center (CPC). The product is called Global Unified Gauge-Based Analysis of Daily Precipitation. The CPC precipitation product is accumulated daily and provided globally at $0.5^\circ \times 0.5^\circ$. To match the resolution of the target river discharge, we mapped CPC data onto the GloFAS domain using nearest point algorithm which preserves the original coarse grid structure but refines the resolution. We did not do any modification for the CPC time zones since it will be considered starting at two days in the past ($t - 2$) (see Sec. C). More details regarding the construction of the daily gauge analysis, the interpolation algorithm, and the gauge algorithm evaluation can be found in [10–12]. Operational CPC data can be obtained from <https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html>.

A.6 LISFLOOD static features

River attributes and static maps are crucial to capture the sub-grid variability for the river discharge. For consistency and to make a fair comparison with GloFAS, we used LISFLOOD input static maps [13] similar to the operational GloFAS. This includes 96 time-invariant variables from 7 different categories (Table 4). The maps are provided at the same resolution as GloFAS at 3 arcmin and covering the globe (90°N-60°S, 180°W-180°E). We excluded the lakes, reservoirs and some static water demand maps.

In addition, we add the Cartesian coordinates for the points on the WGS-84 ellipsoid to enhance the positional encoding:

$$x = (N + H) \cos \phi \cos \lambda, \quad y = (N + H) \cos \phi \sin \lambda, \quad z = ((1 - e^2) + H) \sin \phi, \quad (1)$$

$$N = \frac{a}{\sqrt{(1 - e^2 \sin^2(\phi))}}, \quad e^2 = \frac{a^2 - b^2}{a^2}, \quad (2)$$

where N is the radius of curvature in the prime vertical, H is the height from the elevation model, ϕ and λ are the geographic latitude and longitude, respectively, a and b are the semi-major and semi-minor axes of the ellipse, and e is the eccentricity. We set $a = 6,378.137$ km and $b = 6,356.752$ km.

The LISFLOOD static maps can be obtained from the Joint Research Centre Data Catalogue <http://data.europa.eu/89h/68050d73-9c06-499c-a441-dc5053cb0c86>.

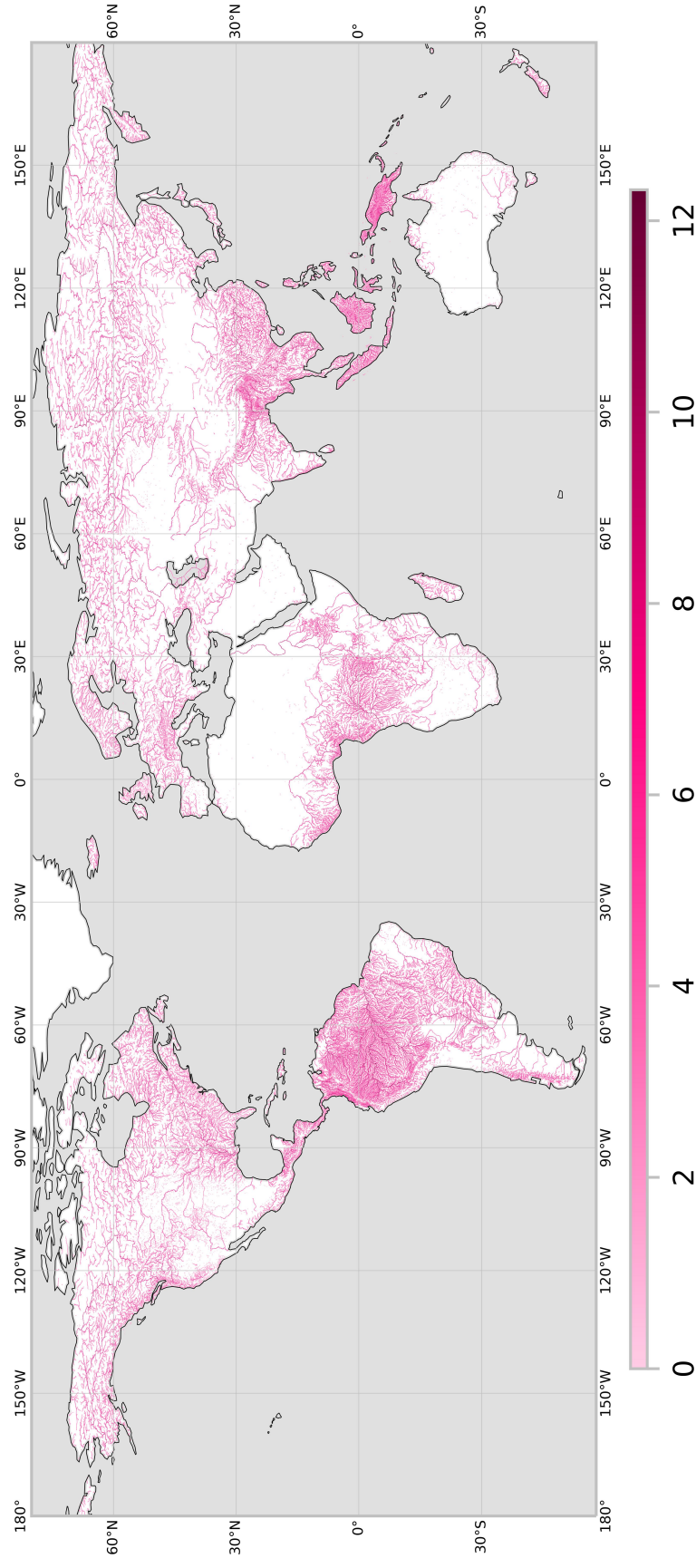
In Sec. J, we show experiments using the widely used HydroRIVERS river attributes data [14, 15].

Table 4: The processed LISFLOOD static and parameter maps [13].

Category	# Static features
catchment morphology and river network	12
grid	2
land use	6
vegetation properties	45
soil properties	14
water demand	3
GloFASv4.0 calibrated parameters	14

A.7 Diagnostic river points

The original resolution of GloFAS v.4.0 is 3 arcmin with an image resolution of 3000×7200 (21 million pix). In order to run experiments efficiently, we sampled points. For this, we remove all points that are not located on the land surface, i.e., points over ocean or sea. This reduced the points from 21,000,000 to 6,221,926 points. We excluded points with median river discharge less than $10 \text{ m}^3 \text{ s}^{-1}$ since river discharge is more relevant where there is a water flow, i.e., points that are located near to water bodies and not located over desert or glacier regions. Points which are close to rivers (distance 1 pix to points with discharge $> 10 \text{ m}^3 \text{ s}^{-1}$) were not excluded. We also do not exclude points defined as GRDC stations. This reduced the points further to 1,529,667 diagnostic river points on which we train and test. Figure 4 gives an overview of the filtered diagnostic river points used in this study. Note that the trained model can generate river discharge maps at full resolution as can be seen in Sec. L.



River discharge [$\text{m}^3 \text{s}^{-1}$] (log scale)

Figure 4: Overview of the selected river points.

B Return periods and flood definition

In hydrology, the return period (also known as the recurrence interval) is a statistical measure that estimates how often a given hydrological event such as a flood, drought, or heavy rainfall is expected to occur on average over a long period. In this study, the return periods refer to the flood frequency. For example, a 2-year flood has a 50% chance of being exceeded in any given year. The return period (RP) is defined as the inverse of the annual exceedance probability (AEP):

$$RP = \frac{1}{AEP}. \quad (3)$$

In practice, flood return periods are used to define flood thresholds, i.e., a flood warning is triggered when discharge exceeds the 2-year threshold. Note that a high return period i.e., 20 years does not necessarily imply actual flooding in all regions, particularly in highly regulated or flood-resilient areas. A high return period event simply reflects the statistical rarity in streamflow magnitude, and should not be equated with a flood event without additional context (e.g., thresholds, inundation). The return period is used here as a proxy indicator of hydrological extremity, which we call flood severity. In this study, we adapted the GloFAS approach to define flood thresholds corresponding to selected return periods (or recurrence intervals) of 1.5, 2, 5, 10, 20, 50, 100, 200, and 500 years. These thresholds are derived from the LISFLOOD reanalysis simulations, which are forced with ERA5 meteorological data. The return levels are estimated by fitting a Gumbel extreme value distribution to the annual maxima for the period 1979–2022, using the L-moments method. For the evaluation on GloFAS reanalysis data, we use the pre-defined return periods data from the Copernicus Emergency Management Service (<https://confluence.ecmwf.int/display/CEMS/Auxiliary+Data>). Fig. 3 shows the flood thresholds defined by different return periods. For the GRDC observation dataset, we calculated the return periods at individual stations from the first available observation date to 2022. To allow a fair evaluation of GloFAS reanalysis data on GRDC observation, the return period of GloFAS data is also calculated at GRDC stations but on the local available observation time period. Note that while we calculated return period thresholds separately from both the GRDC observations and the GloFAS reanalysis, we did not calculate return periods from the trained ML models reforecast as this would require generating a long reforecast climatology to fit a statistical extreme value distribution.

C Implementation and training details

The training was done on clusters with NVIDIA A100 80GB and 48GB GPUs. In Table 5, we highlight the main hyperparameters used for training RiverMamba.

To mimic a real operational setting, the initial conditions from CPC data starts at day $t - 2$, GloFAS reanalysis at day $t - 1$, and ERA5-Land reanalysis at day $t - 1$ in the past. All input data are normalized based on the computed mean and standard deviation from the training set. To handle missing data in the reanalysis, we first use the pre-computed statistics to normalize the data. Then, we replace the invalid pixels with zero values. HRES data is always used for validating and testing. During training, we replace IFS meteorological forcing by ERA5 if they are unavailable, i.e., before 2010. The training, validation, and testing splits are shown in Fig. 5.

To accelerate training and to fit the data into the memory, we use bfloat16 floating point precision. During inference, we use float32 floating point precision. Pre-training RiverMamba took about 3 days on 16 GPUs. Finetuning on GRDC data took about 4 hours on 16 GPUs.

Table 5: Implementation details of RiverMamba

Configuration	Pre-training (GloFAS-Reanalysis)	Fine-tuning (GRDC)
Optimizer	AdamW	AdamW
Learning rate	0.0006	0.0001
Minimum learning rate	0.00009	0.00009
Batch size (B)	1	1
Learning rate scheduler	Cosine annealing	Cosine annealing
Weight decay	0.001	0.01
Training epochs	60	20
Warmup epochs	4	6
Gradient clip	10	10
Input hindcast length (T)	4	4
Lead time (L)	7	7
α for \hat{u}	0.25	0.25
Number of input points P	245,954	245,954*
Embedding dimension for GloFAS reanalysis	48	48
Embedding dimension for ERA5-Land reanalysis	128	128
Embedding dimension for CPC	16	16
Number of hindcast layers	3	3
Hidden dimension in hindcast block (K)	192	192
Depth of hindcast layers	[2, 2, 2]	[2, 2, 2]
Curves in hindcast layers	{Sweep_H, Sweep_V, Gilbert, Gilbert trans}	{Sweep_H, Sweep_V, Gilbert, Gilbert trans}
Grouping size in hindcast block	[(4, 254945), (2, 254945), (1, 254945)]	[(4, 254945), (2, 254945), (1, 254945)]
Dropout in hindcast block	0.2	0.4
D_state in hindcast block	16	16
D_conv in hindcast block	4	4
Hidden dimension in forecast block (K)	192+64	192+64
Embedding dimension for HRES (K_{HRES})	64	64
Number of forecast layers	7	7
Depth of forecast layers	[1, 1, 1, 1, 1, 1, 1]	[1, 1, 1, 1, 1, 1, 1]
Curves in forecast layers	{Sweep_H, Sweep_V, Gilbert, Gilbert trans}	{Sweep_H, Sweep_V, Gilbert, Gilbert trans}
Grouping size in forecast block	[(1, 254945)] * 7	[(1, 254945)] * 7
Dropout in forecast block	0.2	0.4
D_state in forecast block	16	16
D_conv in forecast block	4	4
Hidden dimension in forecasting head (K_{head})	64	64
Dropout in head	0.1	0.3

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

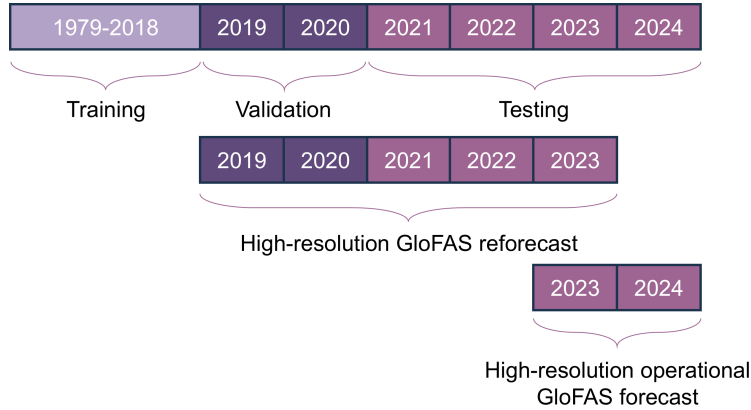


Figure 5: Details about the data splits.

D Mamba Block

Algorithm 1 Mamba block

Require:

- 1: token sequence $\mathbf{X}^l : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K})$
- 2: token sequence $\mathbf{X}_{static} : (\mathbf{B}, \mathbf{P}, \mathbf{K})$
- 3: curve ID (S_{block}) specific to the block l

Result:

- 4: transformed token sequence $\mathbf{X}^{l+1} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K})$
-
- 5: # serialize the input sequence along the P dimension
 - 6: $\mathbf{X}^l : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K}), \mathbf{X}_{static} : (\mathbf{B}, \mathbf{P}, \mathbf{V}_s) \leftarrow \text{Serialization}(\mathbf{X}^l, S_{block}), \text{Serialization}(\mathbf{X}_{static}, S_{block})$
 - 7: # adaptively normalize the input sequence \mathbf{X}^l
 - 8: $\mathbf{X}^{l'} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K}) \leftarrow \text{LOAN}_1(\mathbf{X}^l, \mathbf{X}_{static})$
 - 9: # projection of $\mathbf{X}^{l'}$ into \mathbf{x} and \mathbf{z} , here E is equal to K in our work since we do not expand the dimension
 - 10: $\mathbf{x} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}), \mathbf{z} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}) \leftarrow \text{Linear}^{\mathbf{xz}}(\mathbf{X}^{l'})$
 - 11: # process with different direction
 - 12: **for** o in {forward, backward} **do**
 - 13: # flip the curve along the spatial dimension P
 - 14: **if** $d = \text{'backward'}$ **then**
 - 15: $\mathbf{x} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}) \leftarrow \text{Flip}(\mathbf{x})$
 - 16: **end if**
 - 17: # flatten the curve along the temporal dimension 'spatial-first'
 - 18: $\mathbf{x}' : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E}) \leftarrow \text{Flatten}(\mathbf{x})$
 - 19: # selective state space model, here N is the D_{state}
 - 20: $\mathbf{x}'_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E}) \leftarrow \text{SiLU}(\text{Conv1d}_o(\mathbf{x}'))$
 - 21: $\mathbf{B}_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{N}), \mathbf{C}_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{N}) \leftarrow \text{Linear}_o^{\mathbf{B}}(\mathbf{x}'_o), \text{Linear}_o^{\mathbf{C}}(\mathbf{x}'_o)$
 - 22: # initialize D_o with ones
 - 23: $\mathbf{D}_o : (\mathbf{E}) \leftarrow \text{Parameter Ones} : (\mathbf{E})$
 - 24: # softplus ensures positive Δ_o
 - 25: $\Delta_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E}) \leftarrow \log(1 + \exp(\text{Linear}_o^{\Delta}(\mathbf{x}'_o) + \text{Parameter}_o^{\Delta}))$
 - 26: # shape of $\text{Parameter}_o^{\Delta}$ is (\mathbf{E}, \mathbf{N})
 - 27: $\overline{\mathbf{A}}_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E}, \mathbf{N}) \leftarrow \Delta_o \otimes \text{Parameter}_o^{\Delta}$
 - 28: $\overline{\mathbf{B}}_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E}, \mathbf{N}) \leftarrow \Delta_o \otimes \mathbf{B}_o$
 - 29: # initialize h_o and y_o with zeros
 - 30: $h_o : (\mathbf{B}, \mathbf{E}, \mathbf{N}) \leftarrow \text{Zeros} : (\mathbf{B}, \mathbf{E}, \mathbf{N})$
 - 31: $y_o : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E}) \leftarrow \text{Zeros} : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{E})$
 - 32: # SSM recurrent
 - 33: **for** i in $\{0, \dots, L-1\}$ **do**
 - 34: $h_o = \overline{\mathbf{A}}_o[:, i, :, :] \odot h_o + \overline{\mathbf{B}}_o[:, i, :, :] \odot \mathbf{x}'_o[:, i, :, \text{None}]$
 - 35: $y_o[:, i, :] = h_o \otimes \mathbf{C}_o[:, i, :] + \mathbf{D}_o[\text{None}, :] \odot \mathbf{x}'_o[:, i, :]$
 - 36: **end for**
 - 37: # reshape $(\mathbf{T} \times \mathbf{P})$ to (\mathbf{T}, \mathbf{P})
 - 38: $y_o : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}) \leftarrow \text{Reshape}(y_o)$
 - 39: # flip the curve along the spatial dimension P
 - 40: **if** $o = \text{'backward'}$ **then**
 - 41: $y_o : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}) \leftarrow \text{Flip}(y_o)$
 - 42: **end if**
 - 43: **end for**
 - 44: # get gated y
 - 45: $y'_{forward} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}), y'_{backward} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{E}) \leftarrow y_{forward} \odot \text{SiLU}(\mathbf{z}), y_{backward} \odot \text{SiLU}(\mathbf{z})$
 - 46: # post normalization and residual connection
 - 47: $\mathbf{X}^{l+1'} : (\mathbf{B}, (\mathbf{T} \times \mathbf{P}), \mathbf{K}) \leftarrow \text{Linear}^{\mathbf{X}}(\text{LayerNorm}((y'_{forward} + y'_{backward})/2)) + \mathbf{X}^l$
 - 48: # adaptively normalize the output sequence $\mathbf{X}^{l+1'}$
 - 49: $\mathbf{X}^{l+1} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K}) \leftarrow \text{LOAN}_2(\mathbf{X}^{l+1'}, \mathbf{X}_{static})$
 - 50: # feed-forward layer and residual connection
 - 51: $\mathbf{X}^{l+1} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K}) \leftarrow \text{MLP}(\mathbf{X}^{l+1}) + \mathbf{X}^{l+1'}$
 - 52: # resort the input sequence along the P dimension
 - 53: $\mathbf{X}^{l+1} : (\mathbf{B}, \mathbf{T}, \mathbf{P}, \mathbf{K}) \leftarrow \text{Resort}(\mathbf{X}^{l+1}, S_{block})$
 - 54: **Return:** \mathbf{X}^{l+1}
-

E Evaluation metrics

To assess model performance, we used 8 metrics that are commonly used for hydrological modeling and flood forecasting evaluation [16]. This includes MAE (Mean Absolute Error), RMSE (Root Mean Square Error), R (Pearson Correlation Coefficient), R2 (Coefficient of Determination), KGE (Kling–Gupta Efficiency), Precision, Recall and F1 score. Below are the details about the individual metrics:

Mean Absolute Error (MAE) represents the average of the absolute differences between the predicted and observed values. It provides a straightforward measure of model accuracy. MAE is less sensitive to outliers than RMSE:

$$\text{MAE} = \frac{1}{P} \sum_{p=1}^P \left| \mathbf{X}_p^{\text{obs}} - \mathbf{X}_p^{\text{pred}} \right|, \quad (4)$$

where $\mathbf{X}_p^{\text{obs}}$ is the observed river discharge at point p , $\mathbf{X}_p^{\text{pred}}$ is the predicted river discharge, and P is the total number of points.

Root Mean Square Error (RMSE) measures the square root of the average squared differences between predicted and observed values. It penalizes large errors more heavily than MAE:

$$\text{RMSE} = \sqrt{\frac{1}{P} \sum_{p=1}^P (X_p^{\text{obs}} - X_p^{\text{pred}})^2}. \quad (5)$$

Pearson Correlation Coefficient (R) measures the linear relationship between observed and predicted values, ranging from -1 (perfect negative correlation) to $+1$ (perfect positive correlation):

$$R = \frac{\sum_{p=1}^P (X_p^{\text{obs}} - \bar{X}^{\text{obs}})(X_p^{\text{pred}} - \bar{X}^{\text{pred}})}{\sqrt{\sum_{p=1}^P (X_p^{\text{obs}} - \bar{X}^{\text{obs}})^2} \sqrt{\sum_{p=1}^P (X_p^{\text{pred}} - \bar{X}^{\text{pred}})^2}}, \quad (6)$$

where \bar{X}^{pred} is the mean of predicted river discharge, and \bar{X}^{obs} is the mean of observed river discharge.

Coefficient of Determination (R2) evaluates the predictive power of a model relative to the observed mean. It has the same meaning as Nash–Sutcliffe Efficiency (NSE) which is commonly used in hydrology. Values closer to 1 indicate better performance, while values below 0 suggest that the model performs worse than using the observed mean:

$$\text{R2} = 1 - \frac{\sum_{p=1}^P (X_p^{\text{obs}} - X_p^{\text{pred}})^2}{\sum_{p=1}^P (X_p^{\text{obs}} - \bar{X}^{\text{obs}})^2}. \quad (7)$$

Kling–Gupta Efficiency (KGE) is a composite metric that combines correlation, bias, and variability. It addresses some weaknesses of NSE by ensuring balance across multiple aspects of model performance. Like NSE, values near 1 indicate good performance, while values below 0 indicate performance worse than the observed mean:

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}, \quad (8)$$

$$\beta = \frac{\bar{X}^{\text{pred}}}{\bar{X}^{\text{obs}}}, \quad \gamma = \frac{\text{CV}^{\text{pred}}}{\text{CV}^{\text{obs}}}, \quad r = \text{Pearson correlation coefficient}. \quad (9)$$

where r is the Pearson correlation between observed and simulated, β is the bias ratio, γ is the variability ratio, and $\text{CV}^{\text{obs}} = \sigma^{\text{obs}} / \bar{X}^{\text{obs}}$ and $\text{CV}^{\text{pred}} = \sigma^{\text{pred}} / \bar{X}^{\text{pred}}$ are the coefficients of variation, where σ^{obs} and σ^{pred} are the standard deviations of the observed and predicted river discharge, respectively.

Precision is the proportion of correctly identified positive cases (i.e., flood events) among all predicted positives. High precision indicates a low false-positive rate:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (10)$$

where TP is the number of true positives and FP is the number of false positives.

Recall is the proportion of correctly identified positives among all actual positives. High recall indicates a low false-negative rate:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (11)$$

where FN is the number of false negatives.

F1-score is the harmonic mean of precision and recall, particularly useful in imbalanced classification tasks (i.e., flood detection where flood events are rare):

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (12)$$

In this study, the metrics (4)-(8) are used to evaluate the agreement between observed and forecast discharge time series. During the evaluation, these metrics are calculated on the time series at single grid points and then averaged over all the grid points. The metrics (10)-(12) are applied to assess the model's ability to detect flood events at different return periods. For example, on a day classified as exceeding the 2-year return period threshold, a correct prediction of discharge above this threshold is considered a true positive. If not otherwise specified, we report F1-score averaged over 1.5-20 year return periods and all 3366 points.

F Ablation studies

For the ablation studies, we conducted experiments over the European domain (60°N 30°S, −10°W 40°E) which has 82,804 points from the filtered diagnostic river points defined in Sec. A.7 and includes 675 GRDC stations for evaluation.

F.1 Mamba vs. Transformer

In Fig. 6, we compare the model using Mamba blocks to a variant using Transformer blocks with Flash-Attention [17, 18]. Both Mamba and Flash-Attention are efficient compared to a typical self-attention. However, Mamba scales better with the number of input tokens (Fig. 6 (left)), important for global modeling. The Transformer-based approach becomes computationally infeasible regarding the runtime for a larger number of input points. Both approaches have similar memory consumption which scales linearly with the sequence length. Using the bidirectional Mamba block increases the memory consumption slightly (Fig. 6 (right)).

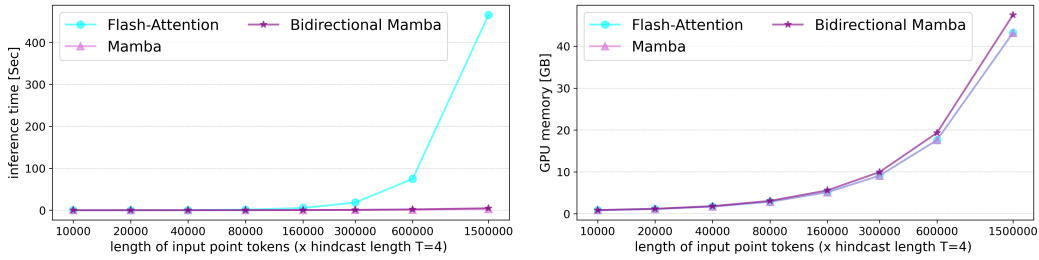


Figure 6: Comparison between Mamba and Transformer-based backbones.

Table 6 shows a comparison between Mamba and Flash-Attention regarding training time with different configurations. Since Flash-Attention does not scale with global data, we split the space-filling curve and do a local modeling (see Fig.8 (b)). This requires rearranging the curves at each block, which becomes the main bottleneck for Flash-Attention compared to Mamba.

Table 6: Training time on the reanalysis data for 1 epoch (1,529,667 points).

Model	Time (min)	GPUs	CPUs
Flash-Attention	~ 145	16 × A100	64 × 4
Mamba	~ 82	16 × A100	64 × 4

Table 7 shows the results for RiverMamba with different backbones. Flash-Attention and Mamba2 [19] achieve slightly lower performance compared to Mamba.

Table 7: Ablation studies for the RiverMamba backbone on the validation set over Europe.

Backbone	# Params	KGE F1 (↑)
Flash-Attention	5.03 M	0.9161 0.2804
Mamba	4.38 M	0.9205 0.2875
Mamba2	4.02 M	0.9169 0.2793

F.2 Feature importance

In Table 8, we study the performance of RiverMamba with different input features. All input features have an impact on the performance. Removing the observational CPC data only slightly reduces the results for GloFAS reanalysis, but the decrease is larger for observational GRDC. Removing GloFAS reanalysis and ERA5-Land initial conditions reduces the performance as well, but GloFAS reanalysis is more important. For GloFAS reanalysis (fourth row), we only drop GloFAS from the input and keep the data as a target to train the model. During inference, we still use the last step of GloFAS

reanalysis and add it to $\Delta \mathbf{X}_{dis24}$ to generate the discharge. This ensures consistency within the table and isolates the impact of input GloFAS reanalysis on the model. RiverMamba still works without taking GloFAS as input. This highlights that the model is more than a post-processor of discharge data and can in fact be used as a backbone for hydrological modeling. Note that if we want to drop GloFAS completely, we need to change the objective function, i.e., by predicting the absolute value of river discharge or the change of discharge w.r.t. climatology. Finally, removing the meteorological forcing forecasts HRES has the biggest impact since weather forecasting is an important source of information. The best performance is achieved when we use all input variables.

Table 8: Ablation studies for feature importance on the validation set over Europe.

Static (LISFLOOD)	CPC	ERA5	GloFAS-reanalysis	HRES	KGE F1 (↑) (Reanalysis)	KGE F1 (↑) (GRDC Obs)
✗	✓	✓	✓	✓	0.9091 0.2505	0.7633 0.1118
✓	✗	✓	✓	✓	0.9151 0.2842	0.7681 0.1102
✓	✓	✗	✓	✓	0.9077 0.2521	0.7731 0.1110
✓	✓	✓	✗	✓	0.9060 0.2450	0.7521 0.1226
✓	✓	✓	✓	✗	0.7972 0.1276	0.6757 0.0640
✓	✓	✓	✓	✓	0.9205 0.2875	0.7838 0.1335

F.3 Pretraining on reanalysis

In Table 9, we show the value of pretraining on GloFAS reanalysis data for the GRDC prediction. From our experiments, we can see a clear benefit of training on river discharge reanalysis before training the model on GRDC observations.

Table 9: Ablation studies for pretraining on GloFAS reanalysis on the validation set over Europe.

Pretrained on GloFAS reanalysis	KGE F1 (↑)
✗	0.7406 0.0882
✓	0.7838 0.1335

F.4 Space-filling curves

In Table 10, we investigate the impact of various serialization patterns for RiverMamba. Our experiments show that sweep curves perform better than the other curves. Iterating between sweep and Gilbert curves (fourth column) improves the F1-score further. Iterating over all curves improves F1-score, but decreases KGE. For simplicity, we use thus the combination of sweep and Gilbert curves for our experiments.

Table 10: Ablation studies for space-filling curves on the validation set over Europe. The columns indicate the serialization patterns: G for Gilbert, Z for Zigzag, S for Sweep, Shuffle represents shuffling the order inside the hindcast layers. S and Z curves use both direction H and V and G uses both regular and trans Gilbert versions.

Curve type	G	Z	S	S + G	S + G + Z
KGE F1 (↑)	0.9156 0.2733	0.9156 0.2719	0.9205 0.2826	0.9205 0.2875	0.9163 0.2962

In Table 11, we remove the spatial modeling in RiverMamba completely and do the scanning only along the temporal dimension (first row). The results show the importance of spatiotemporal modeling.

The design of the scanning also plays a role. From Table 12 (a), we found that sequential scanning along the spatial dimension (P) works better. This is represented as scanning from P to T ($P \rightarrow T$). In other words, the points are connected over time by scanning at time step t and continuing the scan at

Table 11: Ablation regarding the spatiotemporal modeling on the validation set over Europe.

Temporal modeling	Spatiotemporal modeling	KGE F1 (\uparrow)
✓	✗	0.8726 0.1952
✗	✓	0.9205 0.2875

Table 12: Ablation studies for scan patterns on the validation set over Europe.

(a) Curve order			(b) Curve type			(c) Bidirectional Curve		
T→P	P→T	KGE F1 (\uparrow)	Local	Global	KGE F1 (\uparrow)	SSM	Bi-SSM	KGE F1 (\uparrow)
✓	✗	0.9153 0.2807	✓	✗	0.9164 0.2893	✓	✗	0.9113 0.2482
✗	✓	0.9205 0.2875	✗	✓	0.9205 0.2875	✗	✓	0.9205 0.2875

$t + 1$. In the second case (T→P), each point will be scanned along the time dimension (T) and then connected to the next point along the spatial dimension P .

In Table 12 (b), we split the curve into local curves similar to PointTransformer [20] (Fig. 8 (b)). Using a larger receptive field gives the model more capability to extract up- and downstream features and to model adjacent catchments. In addition, local modeling needs more computations along the network i.e., sorting, resorting and padding. Finally, bidirectional Mamba (Table 12 (c)) collects information about the streamflow from both side of the curve thus covering the whole domain and achieving a better performance than unidirectional Mamba.

For training and inference on the global dataset, it is impractical to fit all the input points (~ 6 million points) into the memory. For this, we first define a Gilbert space-filling curve on the globe and then we split the curve into smaller curves along the space-filling curve, i.e., we split the curve into sequences with ~ 311 K points for each. A simplified version of the splits is shown in Fig. 7.



Figure 7: A simplified view of splitting along Gilbert space-filling curve.

F.5 Weighting in the objective function

In this experiment, we study the effect of weighting floods not just by their return period but also by augmenting it with an additive flood offset of 1. To this end, we trained a model with weighted flood events by their return periods + flood offset of 1. The F1 results are shown below in Table 13 for the reanalysis dataset and different return periods. Adding an offset of 1 does not improve the results.

F.6 Activation function in LOAN

We conducted an additional experiment where we replaced the activation function in the LOAN layer by ReLU activation. GELU [21] avoids the dying ReLU problem and improves optimization. As can be seen from Table 14, GELU performs slightly better.

Table 13: Ablation study on the validation set over Europe regarding the weighting in the objective function. Shown is F1-score for reanalysis data across different return periods.

Return period	1.5	2.0	5.0	10.0	20.0
Validation (2019-2020)					
W/ offset	0.4820	0.3760	0.2358	0.1790	0.1181
W/o offset	0.4870	0.3767	0.2516	0.2015	0.1208
Testing (2021-2024)					
W/ offset	0.6114	0.5080	0.3125	0.2486	0.1656
W/o offset	0.6122	0.5072	0.3031	0.2434	0.1669

Table 14: Ablation study on the validation set over Europe regarding the activation function in LOAN layers.

Activation function	ReLU	GELU
KGE F1 (\uparrow)	0.9143 0.2833	0.9205 0.2875

G Computational time

Neither Google [6] nor GloFAS [7] provided the compute time for the operational forecast. The inference time for RiverMamba is reported in supp. Fig. 6. In Table 15, we report the inference time (seconds) w.r.t. the number of input points for our model with 4 days as a hindcast (first row), a trained version of Google’s LSTM with 4 days as a hindcast (second row), and a trained Google’s LSTM version as in [6] with one year hincast (third row). We use one A100 GPU for all runs. All machine learning approaches are very fast. We expect that GloFAS is by several magnitudes slower, which is a practical advantage of machine learning approaches for this task.

Table 15: Inference time in seconds.

Model	10K	20K	40K	80K	160K	300K	600K	1500K
RiverMamba (4 days hindcast)	0.026	0.044	0.086	0.190	0.423	0.874	1.914	4.739
LSTM (4 days hindcast)	0.005	0.009	0.015	0.027	0.053	0.098	-	-
LSTM (one year hindcast)	0.069	-	-	-	-	-	-	-

H Baselines

H.1 Climatology

We followed [7] to define the climatology baseline. For this, we computed climatology for the long-term record of river discharge data (1979-2018) with a moving window of 31 days centered on the day-of-the-year. Then, we computed 11 fixed quantiles at 10% interval for each day-of-the-year. As a result, the climate distribution changes with lead time, reflecting the dynamic changes in local river discharge patterns over time. Climatology is commonly applied to medium- and extended range lead times, where seasonal patterns predominantly influence the river discharge forecast [7].

H.2 Persistence

We defined the persistence baseline as the daily river discharge of GloFAS reanalysis from the day preceding the day at which the forecast was issued, i.e., for a forecast starting at 00:00 UTC \mathbf{X}^t , the persistence is defined as the averaged river discharge between 00:00 UTC \mathbf{X}^{t-1} and 00:00 UTC \mathbf{X}^t . This value was used as a prediction for the entire lead time. Persistence is primarily applied to short lead times, where the correlation of sequential river discharge values predominantly influences the forecasts [7]. Note that this baseline is unrealistic since no reanalysis is available directly at time t .

H.3 LSTM

We adopted the same LSTM architecture as described in [6]. The model follows an encoder–decoder structure, where the encoder is a bi-directional “hindcast” LSTM that processes historical input data, and the decoder is a uni-directional “forecast” LSTM that generates predictions over a 7-day forecast horizon based on forecast inputs. To ensure fair comparison and benchmarking, we used the same input data (i.e., we include GloFAS reanalysis and exclude IMERG and nowcasting data for LSTM), train–test split, and normalization strategies as in the RiverMamba model. To remain consistent with [6], we trained the model only at locations with available gauge observations, specifically the 3366 GRDC stations (see Sec. A.2) rather than using a global training setup. Thus for LSTM, we do not include any spatial connections and space filling curves are not used in combination with the LSTM baseline.

The model leverages both dynamic and static inputs. For the hindcast LSTM, we used a 14-day sequence of dynamic inputs including CPC precipitation, GloFAS reanalysis, and ERA5-Land reanalysis data. At each time step, static attributes derived from the LISFLOOD model are embedded and concatenated with the dynamic inputs. For the forecast LSTM, we used ECMWF HRES forecasts as dynamic inputs over the 7-day horizon, with the static attributes concatenated in the same manner.

To connect the encoder and decoder, we employed a “state” layer consisting of two transfer networks (<https://neuralhydrology.github.io/>): a linear cell-state transfer network and a nonlinear hidden-state transfer network (a fully connected layer with hyperbolic tangent activation). A linear output head is applied at each forecast step to predict streamflow, and the model is trained using the mean squared error (MSE) loss. Unlike [6], we focus on deterministic prediction, so we do not implement a probabilistic output head or probabilistic loss function. In total, the model has 834,421 parameters.

In [6], an input sequence length of 365 days was used. This is because the model in [6] has to simulate the states (i.e., soil moisture) and current runoff from the meteorological forcing input. In our experiments, since the states and the streamflow already integrate the meteorological signal of the past, we trained the LSTM model using a range of input sequence lengths from 4 to 90 days. We observed only marginal performance gains beyond a certain point, and identified 14 days as an optimal input sequence length.

The reported LSTM results are averaged over an ensemble of three independently trained models, each initialized with a different random seed. Each training batch contains data from all 3366 GRDC stations at a given time step, with a batch size of 1—effectively training on 3366 samples per mini-batch. Training takes approximately 12 hours on four NVIDIA A100 GPUs for 35 epochs. More details about the model architecture can be found in [6], as well as in the NeuralHydrology GitHub repository (<https://neuralhydrology.github.io/>). Table 16 summarizes the key hyperparameters used in our implementation of the LSTM model.

Table 16: Implementation details of the LSTM model.

Configuration	Value
Hidden size in hindcast LSTM	256
Hidden size in forecast LSTM	128
Hidden size in static embedding layer	20
Hidden size in dynamic embedding layer	20
Hidden size in state layer	128
Number of layers	1
Dropout at output regression head	0.4
Dropout at state layer	0.1
Learning rate	0.0003
Learning rate scheduler	Cosine annealing
Batch size	1 with (3366 samples)
Optimizer	Adam
beta1 momentum term	0.9
beta2 momentum term	0.999
weight decay	0

It is important to note that [6] did not release the full code or the full hyperparameter configurations of their final model, but only the pretrained checkpoints were made available. Although the saved models can be loaded for inference using the original inputs, it is not possible to retrain or adapt these models to a different input setup, which was required for our experiments. We therefore used the published checkpoints and the NeuralHydrology GitHub repository as a reference to re-implement and train the LSTM.

All results shown in the paper for the LSTM baseline are obtained by our trained LSTM, except in sections K.1 and K.2, where we compare with the published reforecast of Google’s LSTM obtained from [6].

H.4 GloFAS Forecast

Operational forecast from GloFAS was obtained from the ECMWF Early Warning Data Store (EWDS) <https://doi.org/10.24381/cds.ff1aef77>. This represents real-time data from the official system for operational flood forecasting from the Copernicus Emergency Management Service (CEMS) and managed and developed by the European Commission’s Joint Research Centre. GloFAS forecast is produced by forcing the LISFLOOD model with the ECMWF ensemble forecast (ENS) up to 30 days. GloFAS forecast uses ENS meteorological forcing twice a day at 00:00 UTC. The high-resolution GloFAS v.4.0 forecast is available from 2023-07-26. We compare to this baseline in Sec. K.5.

H.5 GloFAS Reforecast

This baseline is similar to GloFAS forecast (Sec. H.4), however, GloFAS reforecast are forecasts run over the past with the new system version 4.0. The reforecast is available until 2023 and does not span the full testing split. We use this baseline for the main comparison with GloFAS in the main paper and in Sec. K.4.

I Space-filling curves

Serialized encoding maps a point’s position into an integer index representing its order within the given space-filling curve. Each point is stored as a 64-bit integer. For simplicity, we define the curves on the 2D PlateCarree projection of the Earth. As illustrated in Figs. 9 and 10, the serialization is done according to the sorted serialized encoding of all points with $\Phi : \mathbb{Z}^3 \rightarrow \mathbb{N}$. Due to the nature of the bijective transformation, there is an inverse mapping $\Phi^{-1} : \mathbb{N} \rightarrow \mathbb{Z}^3$ which allows for the mapping of the encoded index back into the point’s position $p_i \in \mathbb{Z}^3$ (or $p_i \in \mathbb{R}^3$ in case of a continuous space). This inverse mapping is called the serialized decoding or the deserialization. In the following, we describe the mapping for each curve:

Sweep. This curve fills in the space like a spherical helix or a Luxodrome around the sphere.

Zigzag. This curve is similar to the Sweep curve. The main difference is that the transformation ensures that every neighboring points on the curve are also neighboring in the physical space.

Generalized Hilbert. Generalized Hilbert (Gilbert) is a Hilbert space-filling curve [22] for rectangular domains of arbitrary non-power of two sizes [23]. We used the numpy implementation of (<https://github.com/jakubcerveny/gilbert>) to generate the curves. Transposed Gilbert is generated as $y_{(transpose)} = H - y$, where $y \in [1, H]$.

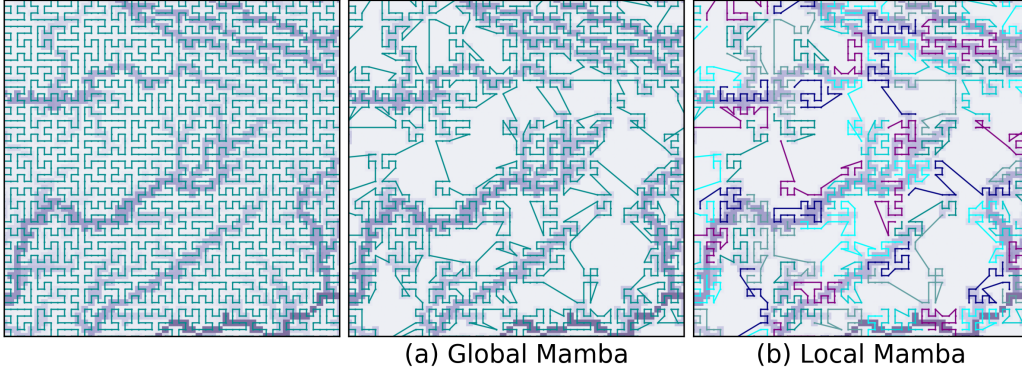


Figure 8: Illustration of the difference between global curve (middle) and grouped local curves (right). The left image shows a Gilbert space-filing curve for all points. In our experiments, the global curve is used (middle). For the experiments with Flash-Attention, we use the local curves (right).

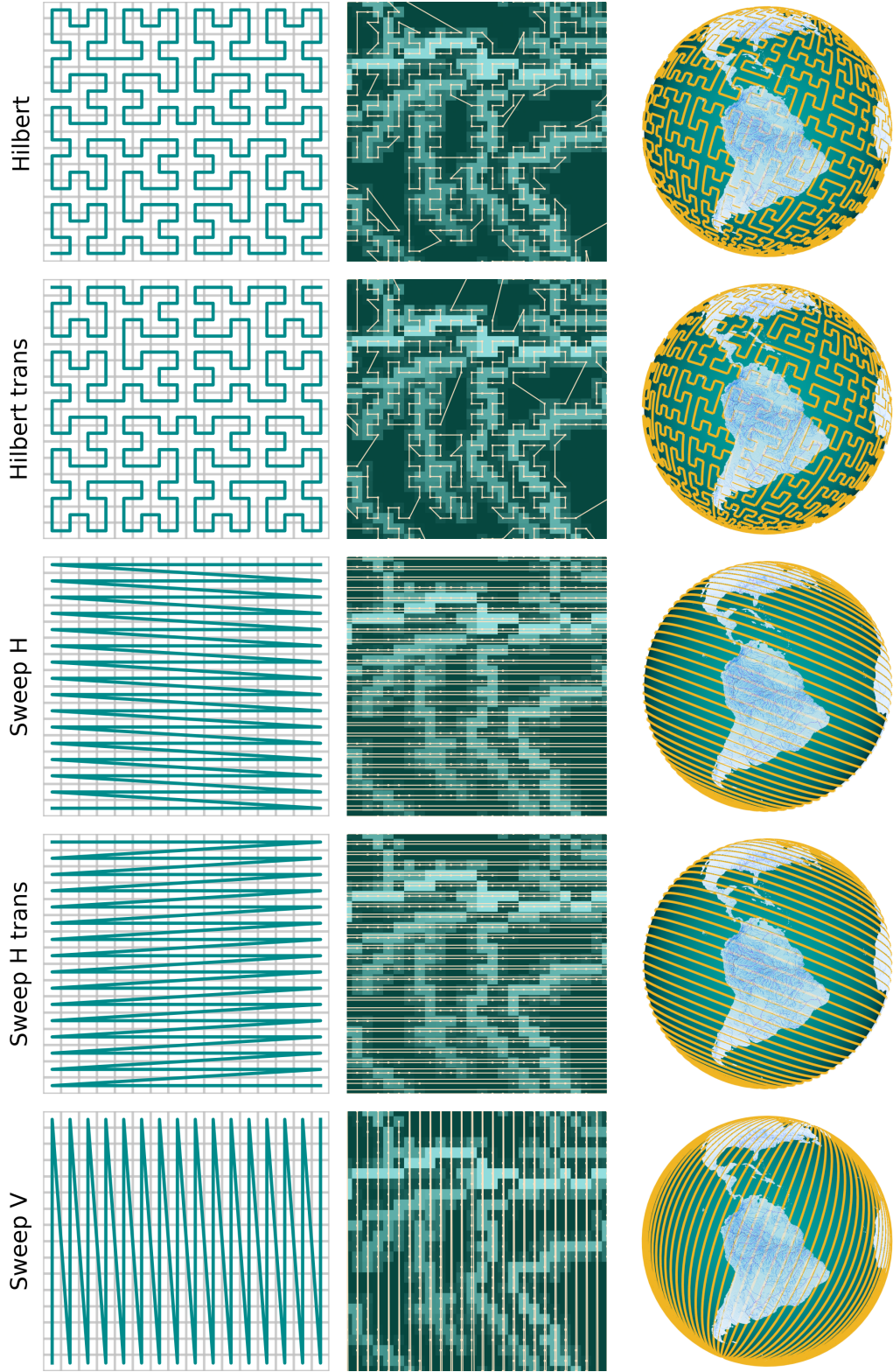


Figure 9: Visualization of different types of space-filling curves. For each type, we show the space-filling curve over a 2D discrete space (left), zoomed in version over the Earth where the points are sorted via a specific serialization order within the space-filling curve (middle), and simplified 3D visualization of the curve over the Earth (right).

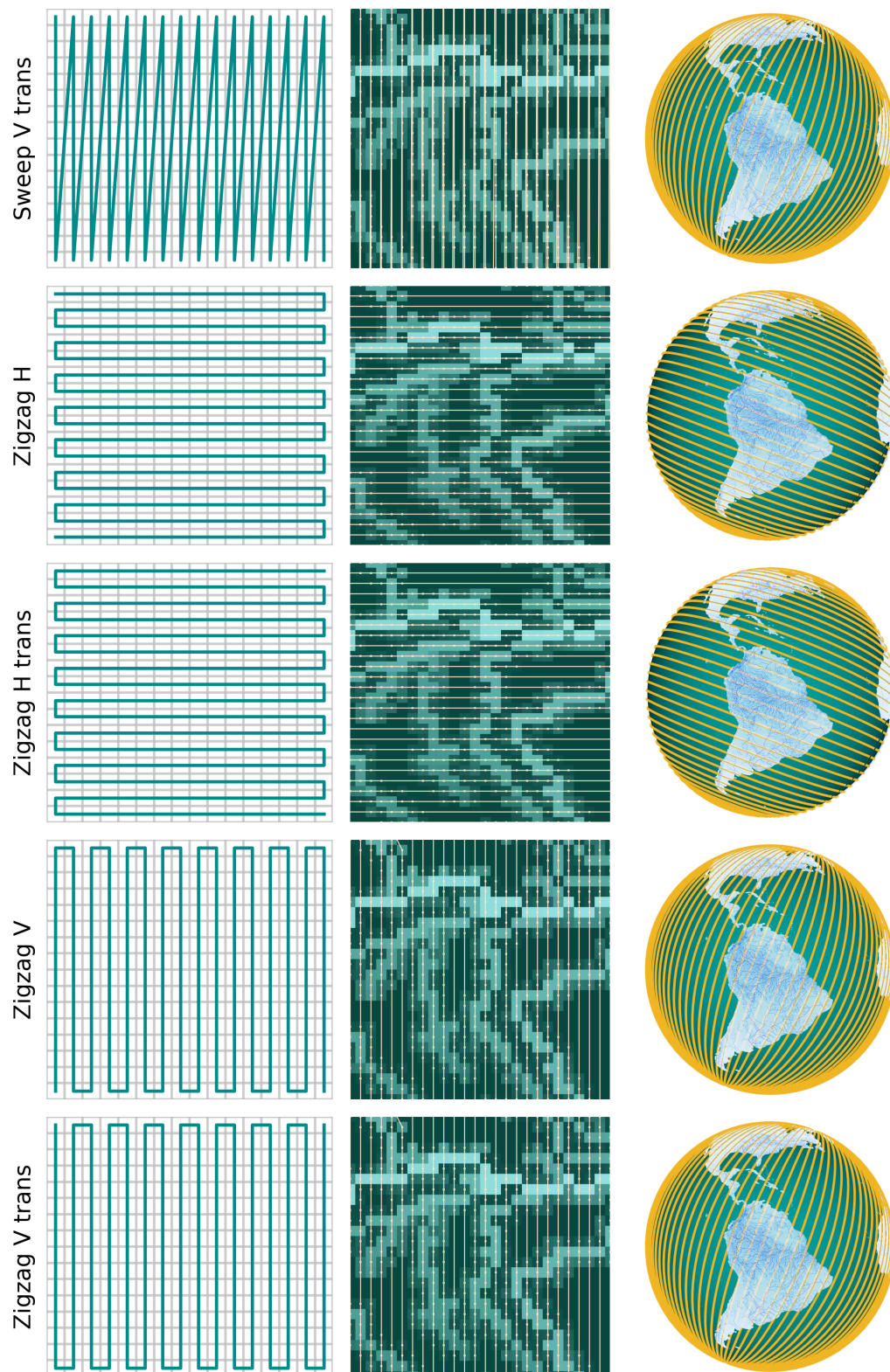


Figure 10: Visualization of different types of space-filling curves. For each type, we show the space-filling curve over a 2D discrete space (left), zoomed in version over the Earth where the points are sorted via a specific serialization order within the space-filling curve (middle), and simplified 3D visualization of the curve over the Earth (right).

J Experiments on HydroRIVERS

HydroRIVERS data are widely used to train deep learning models in hydrology. In this section, we explore the performance of RiverMamba with HydroRIVERS [14, 15]. For this, we obtained static river attributes from <https://www.hydrosheds.org/products/hydrorivers>. The data is stored as a shape file. To map them onto the GloFAS domain, we first extract the coordinates of the rivers and then project them with the grid points on the WGS-84 ellipsoid (Eq. 1 and 2). Then, for each GloFAS grid point, depending on the attribute type, we either average the attributes or take the most frequent attribute within a radius of 5 km. If no attributes were found, we increase the radius to 12 km, and 24 km, respectively. We processed 299 river feature attributes overall and experimented with 103 features, i.e., we removed the monthly attribute statistics from the static features. Fig. 11 gives an overview of the processed HydroRIVERS data.

In Table 17, we compare the LISFLOOD with the HydroRIVERS static maps for prediction on both GloFAS reanalysis and GRDC data. Using HydroRIVERS performs worse than using LISFLOOD static maps.

Table 17: Ablation studies on the validation set over Europe.

HydroRIVERS	LISFLOOD	KGE F1 (\uparrow) (Reanalysis)	KGE F1 (\uparrow) (GRDC Obs)
\times	\times	0.9091 0.2505	0.7633 0.1118
\checkmark	\times	0.9174 0.2622	0.7406 0.1227
\times	\checkmark	0.9205 0.2875	0.7838 0.1335

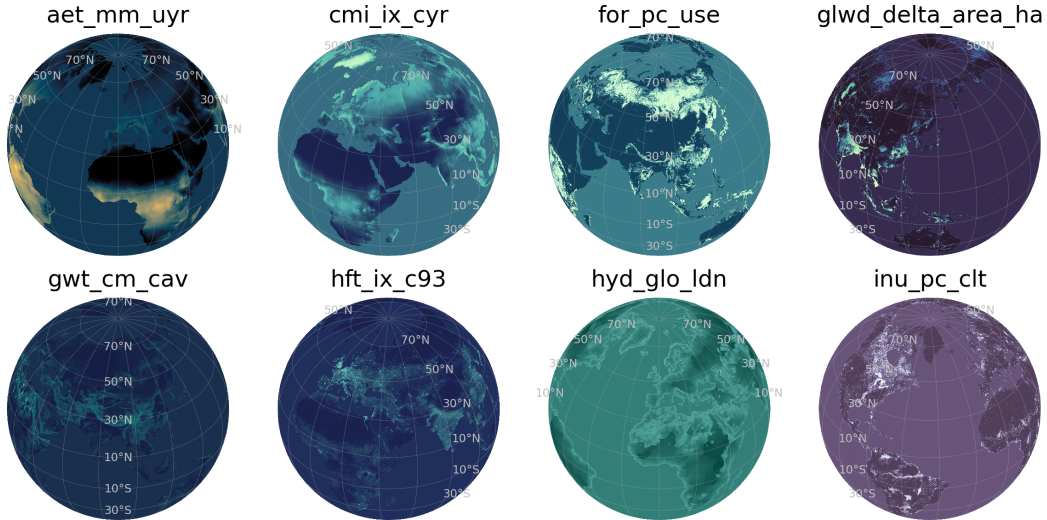


Figure 11: An overview of the processed HydroRIVERS static features. HydroRIVERS is mapped into the GloFAS domain.

K Additional results

K.1 Comparison with Google reforecast on ungauged GRDC

In this section, we evaluate RiverMamba against the published reforecast by [6] and available from [24]. For this, we split our data into 8 folds and evaluate on ungauged stations. All stations were predicted similar to [6] where each station was evaluated out-of-sample in both time and space. Note that the LSTM model for Google reforecast used more stations (~ 5680), while we used much less stations (3366). In addition, LSTM takes one year input as a hindcast, while RiverMamba takes only 4 days as input. Furthermore, RiverMamba does not use nowcasting data at time t and starts the input initial conditions at $t - 1$ to mimic an operational forecast. There are also differences in the input initial conditions, i.e., LSTM uses precipitation estimates from the NASA Integrated Multi-satellite Retrievals for GPM (IMERG) early run as input. In addition, it uses HydroATLAS [15] as geophysical and anthropogenic basin attributes. RiverMamba uses GloFAS reanalysis as an initial condition and LISFLOOD as static basin attributes.

Table 18 shows the overall performance for the years 2014-2021. The F1-score is averaged for all lead times and 1.5-20 year return periods. We expect that adding nowcasting (analysis data) and IMERG as input and an ensemble would improve the results of RiverMamba on ungauged basins further. The ungauged streamflow forecast becomes also better when the number of stations increases. More results are shown in Figs.12-14.

Table 18: Comparison to Google reforecast on ungauged GRDC stations for the years 2014-2021. Shown is the averaged F1-score (\uparrow) for all lead times and 1.5-20 year return periods.

LSTM (Google reforecast from [6])	RiverMamba
0.2164	0.2355

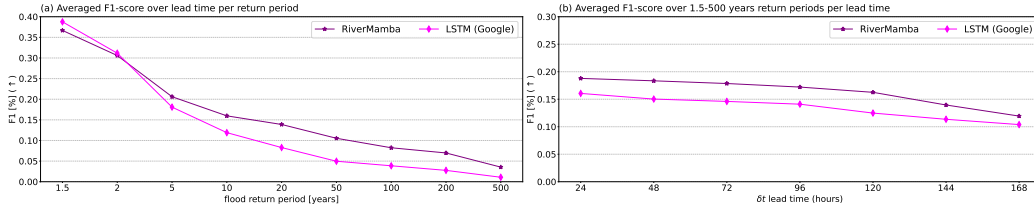


Figure 12: Comparison to Google reforecast on ungauged GRDC stations (test set 2014-2021 out-of-sample in space and time).

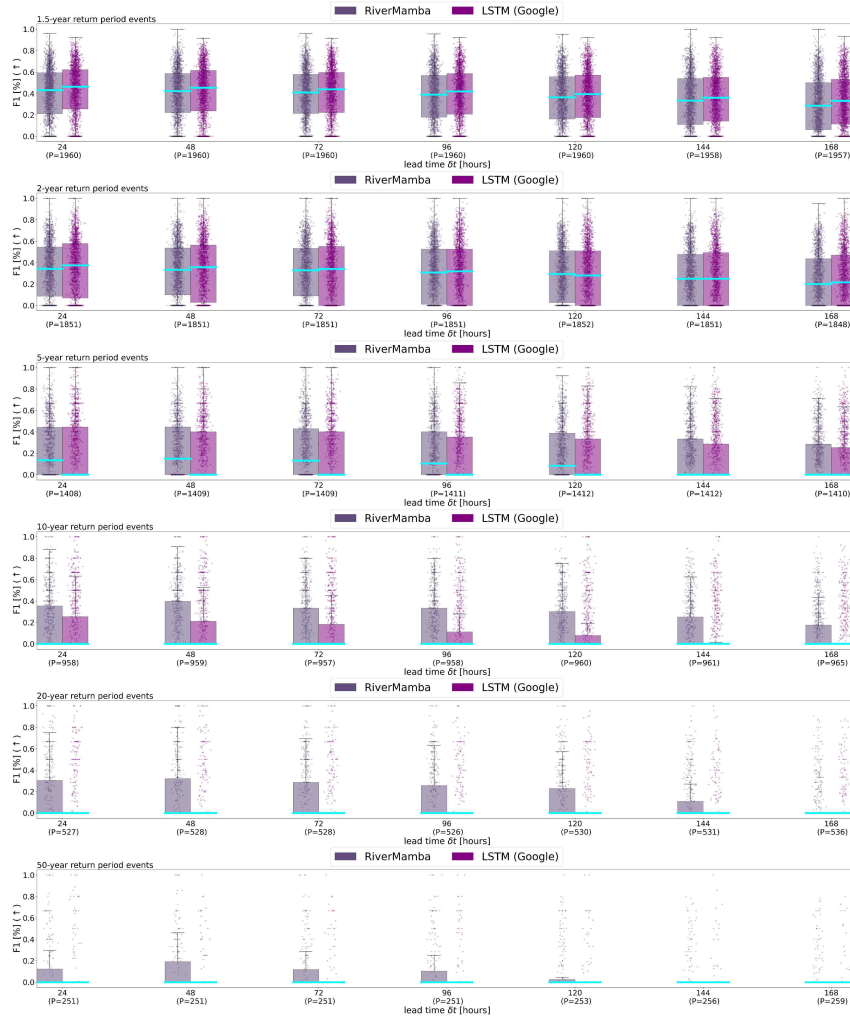


Figure 13: Comparison to Google reforecast. Shown is F1-score of flood forecasting for different return periods and lead time on ungauged GRDC stations (test set 2014-2021 out-of-sample in space and time). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.



Figure 14: Comparison to Google reforecast. Shown is F1-score of flood forecasting for different lead time and return periods (1.5 - 50 years) on ungauged GRDC stations (test set 2014-2021 out-of-sample in space and time). 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.

K.2 Comparison with Google reforecast on gauged GRDC

Similar to Section K.1, here we compare to the published reforecast by [6] and available from [24] but on gauged stations where all stations were evaluated out-of-sample in time for the years 2019-2021. For the differences between RiverMamba and the LSTM model by [6], see Sec. K.1.

Table 19 shows the overall performance for the years 2019-2021. The F1-score is averaged for all lead times and 1.5-20 year return periods. More results are shown in Figs. 15-17.

Table 19: Comparison to Google reforecast on gauged GRDC stations for the years 2019-2021. Shown is the averaged F1-score (\uparrow) for all lead times and 1.5-20 year return periods.

LSTM (Google reforecast from [6])	RiverMamba
0.2318	0.2587

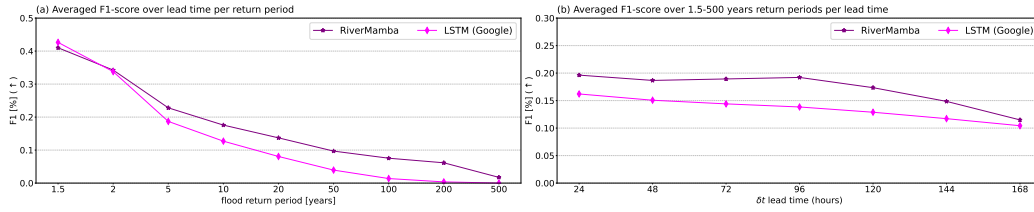


Figure 15: Comparison to Google reforecast on gauged GRDC stations (test set 2019-2021 out-of-sample in time).

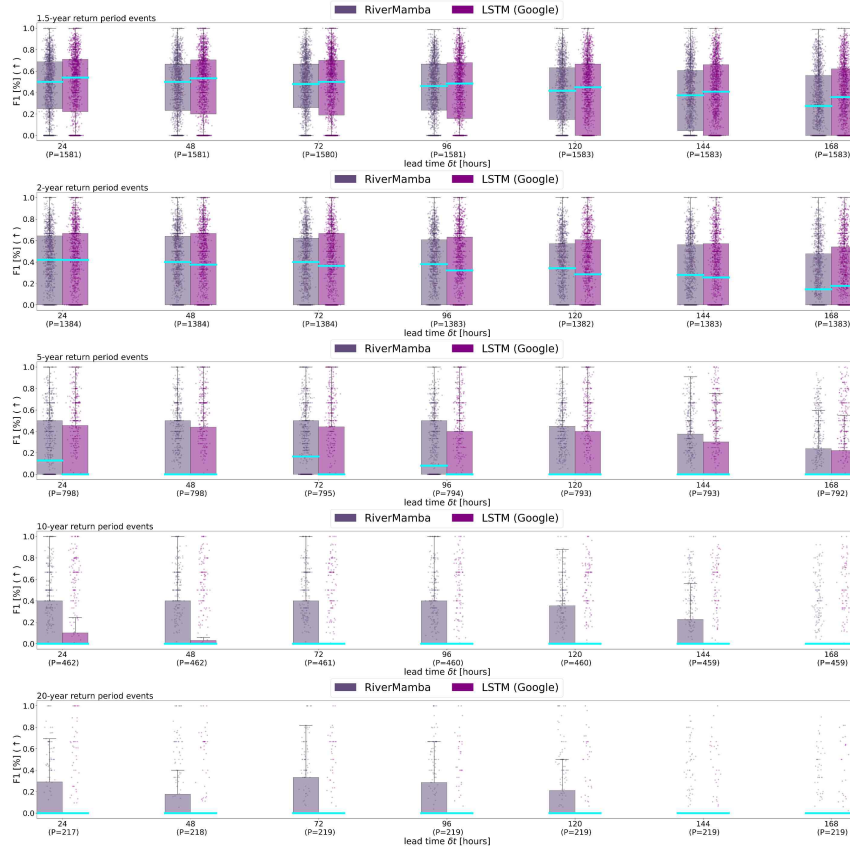


Figure 16: Comparison to Google reforecast. Shown is F1-score of flood forecasting for different return periods and lead time on gauged GRDC stations (test set 2019-2021 out-of-sample in time). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

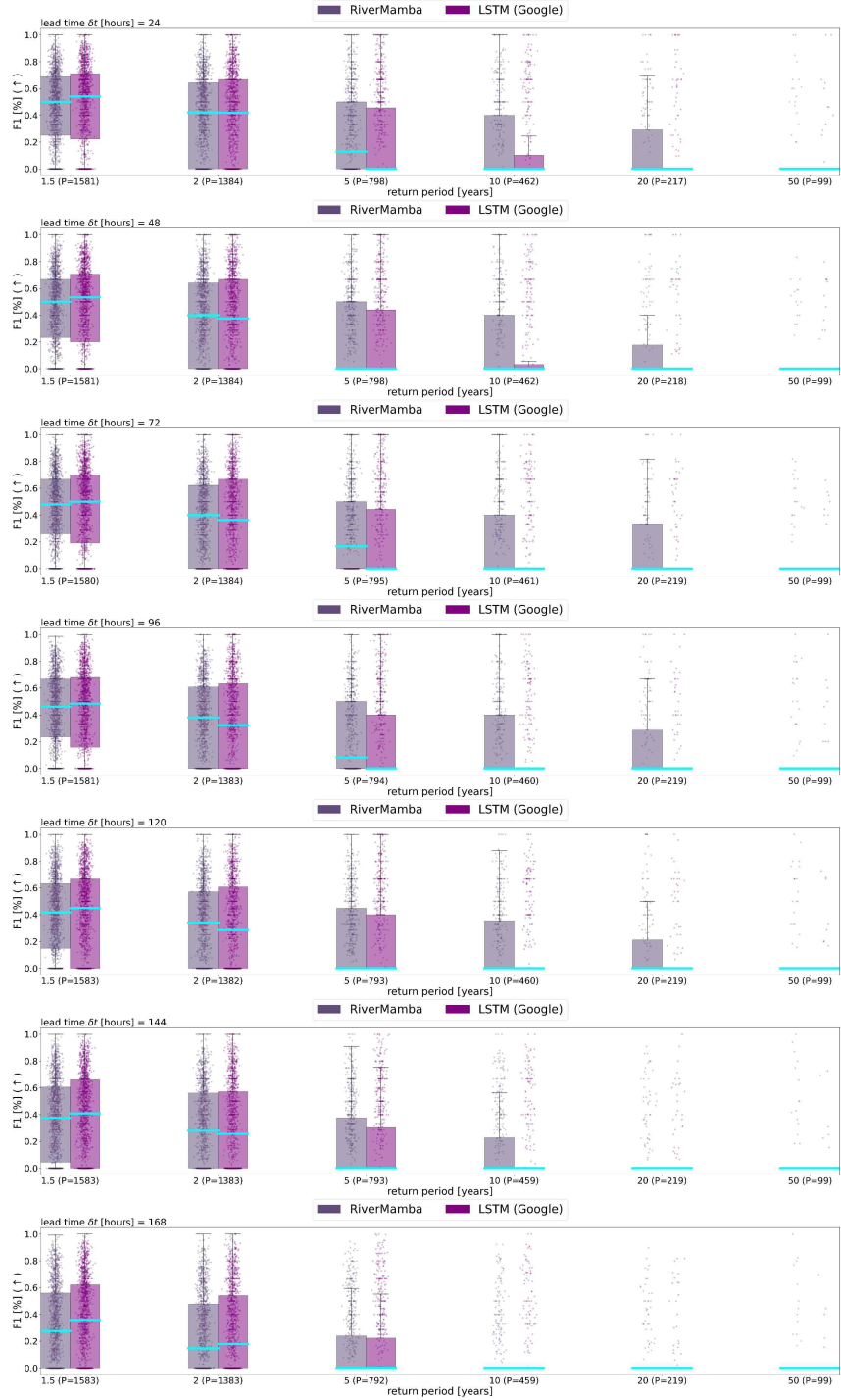


Figure 17: Comparison to Google reforecast. Shown is F1-score of flood forecasting for different lead time and return periods (1.5 - 50 years) on gauged GRDC stations (test set 2019-2021 out-of-sample in time). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

K.3 Additional results on gauged GloFAS reanalysis

In this section, we plot additional results for the experiments on GloFAS river discharge reanalysis. In Figs. 18-21, we report the results for MAE, RMSE, R2, and R metrics with lead time. In Figs. 22-27, we report the results of F1-score, Precision, and Recall metrics for different return periods and lead times. In Figs. 28 and 29, we compare the results between RiverMamba and LSTM for F1-score and KGE metrics spatially. Finally, Fig. 30 shows the confusion matrix for both RiverMamba and LSTM.

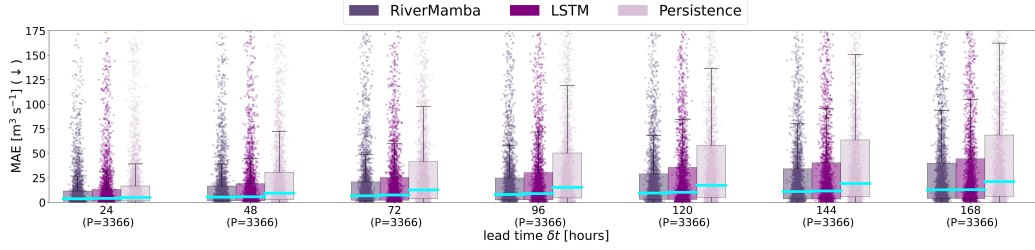


Figure 18: MAE of the river discharge forecasting with different lead time on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample).

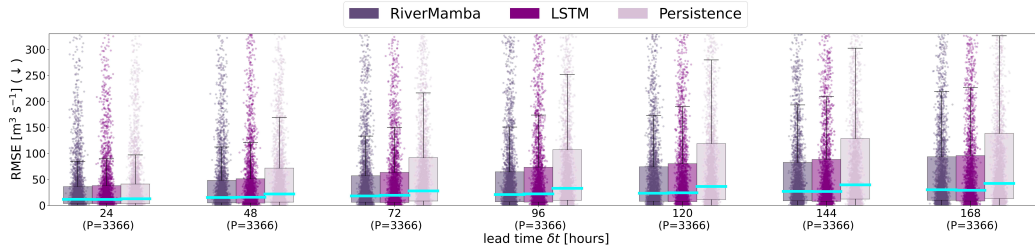


Figure 19: RMSE of the river discharge forecasting with different lead time on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample).

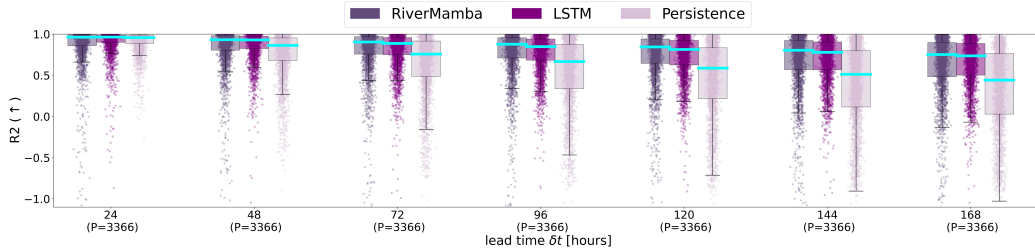


Figure 20: R^2 (NSE) of the river discharge forecasting with different lead time on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample).

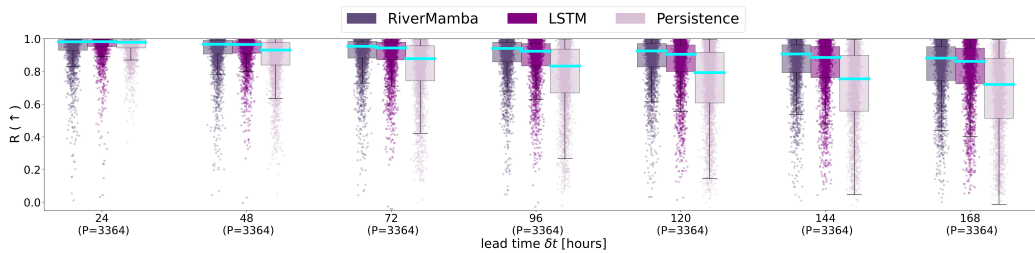


Figure 21: Pearson correlation (R) of the river discharge forecasting with different lead time on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample).



Figure 22: F1-score of flood forecasting for different return periods and lead time on GloFAS reanalysis (test set 2021-2024 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 23: Precision of flood forecasting for different return periods and lead time on GloFAS reanalysis (test set 2021-2024 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 24: Recall of flood forecasting for different return periods and lead time on GloFAS reanalysis (test set 2021-2024 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.



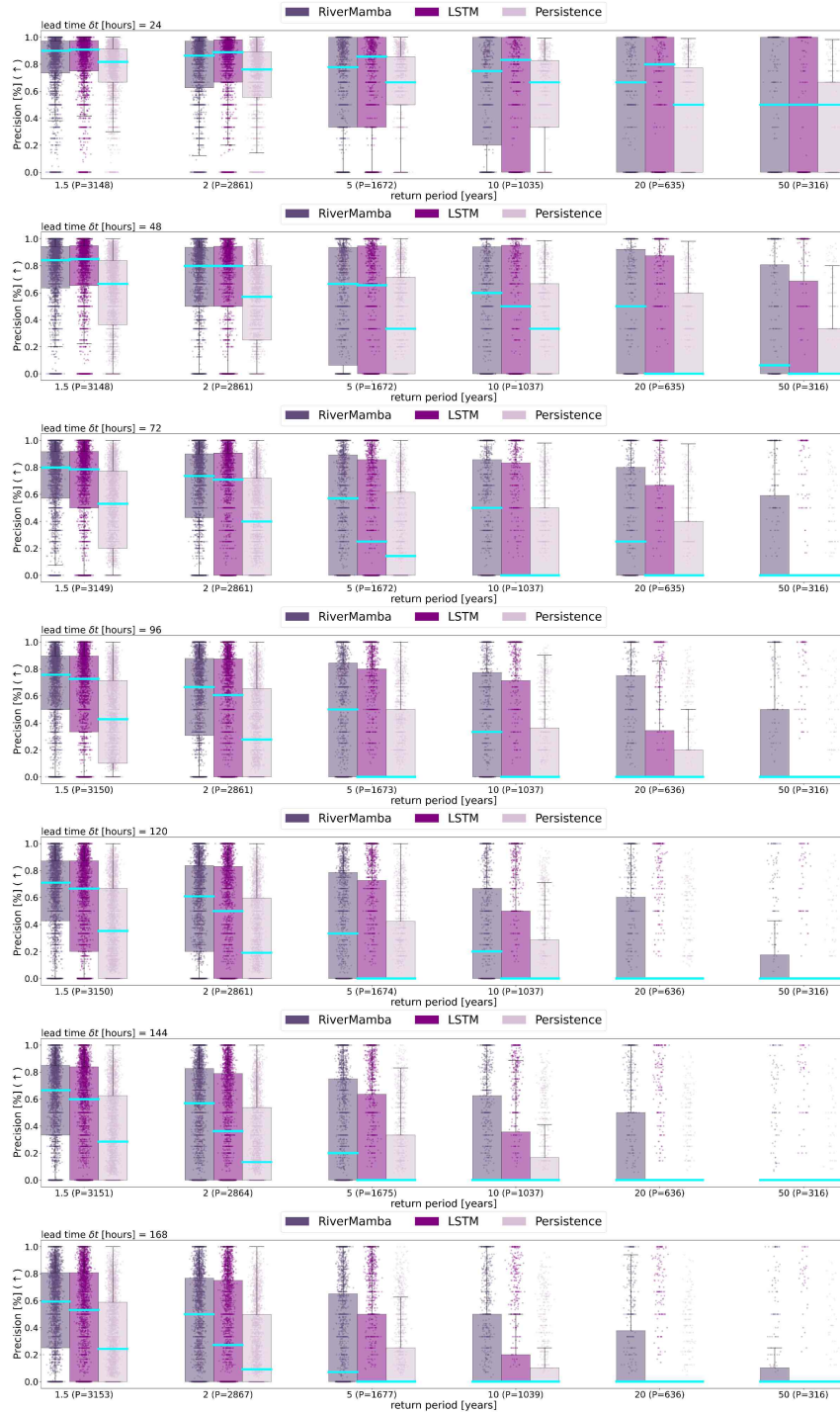


Figure 26: Precision of flood forecasting for different lead time and return periods (1.5 - 50 years) on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

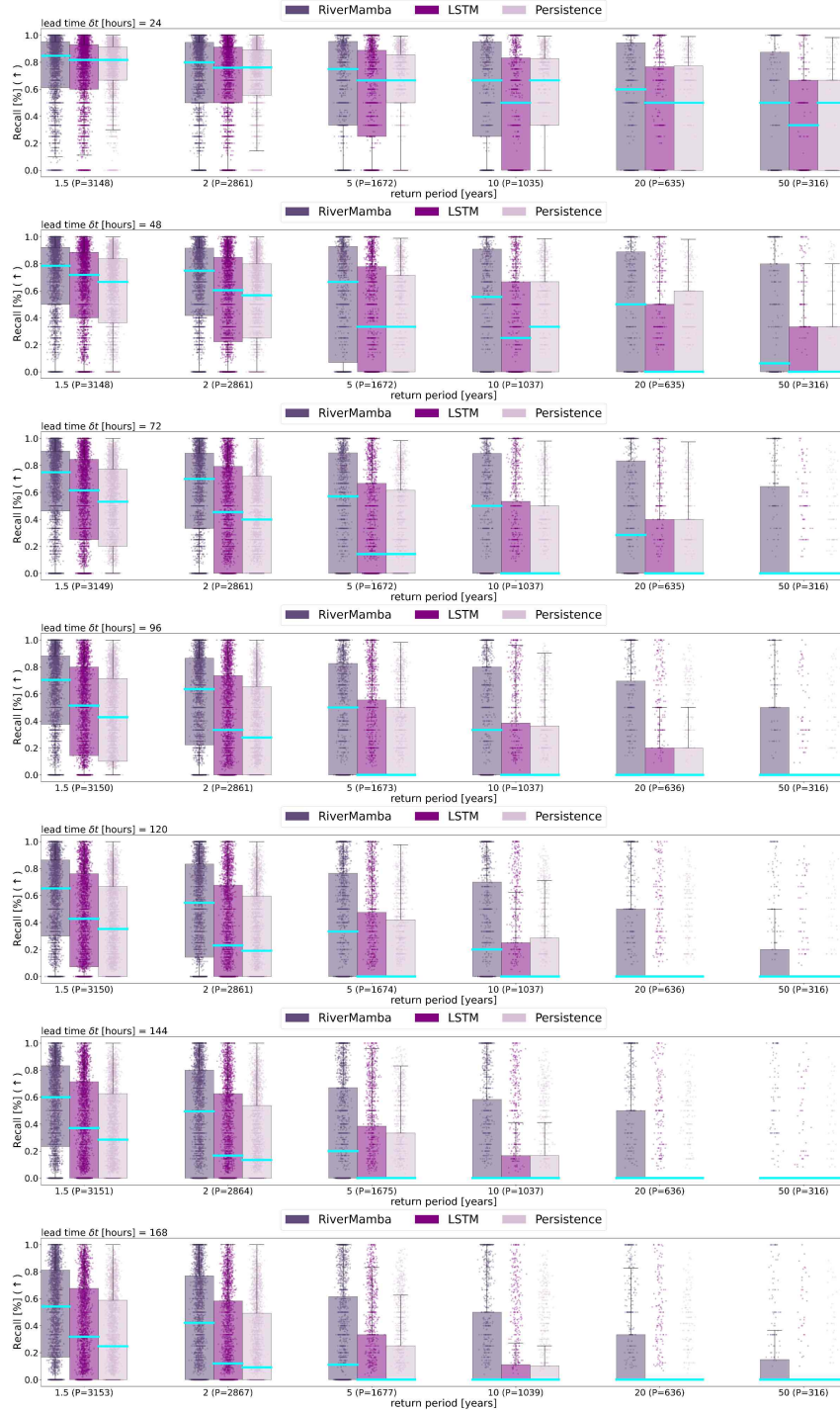


Figure 27: Recall of flood forecasting for different lead time and return periods (1.5 - 50 years) on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.

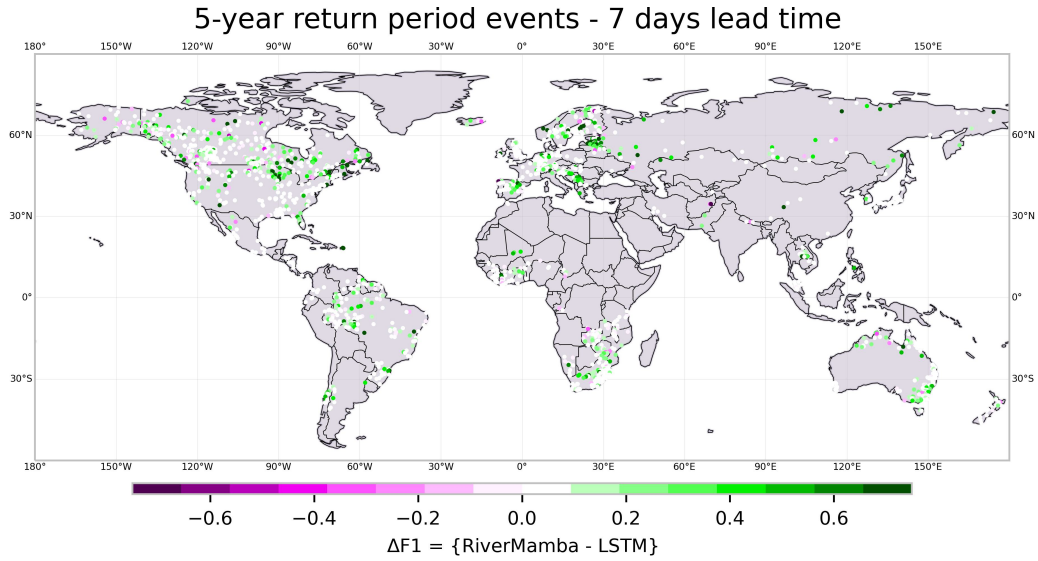


Figure 28: Comparison of F1-score between RiverMamba and LSTM on GloFAS reanalysis for the 5-year return period events (test set 2021-2024 temporally out-of-sample). RiverMamba improves over LSTM in 41% of the stations ($P=1677$) and being better or equally better in 89% of the stations.

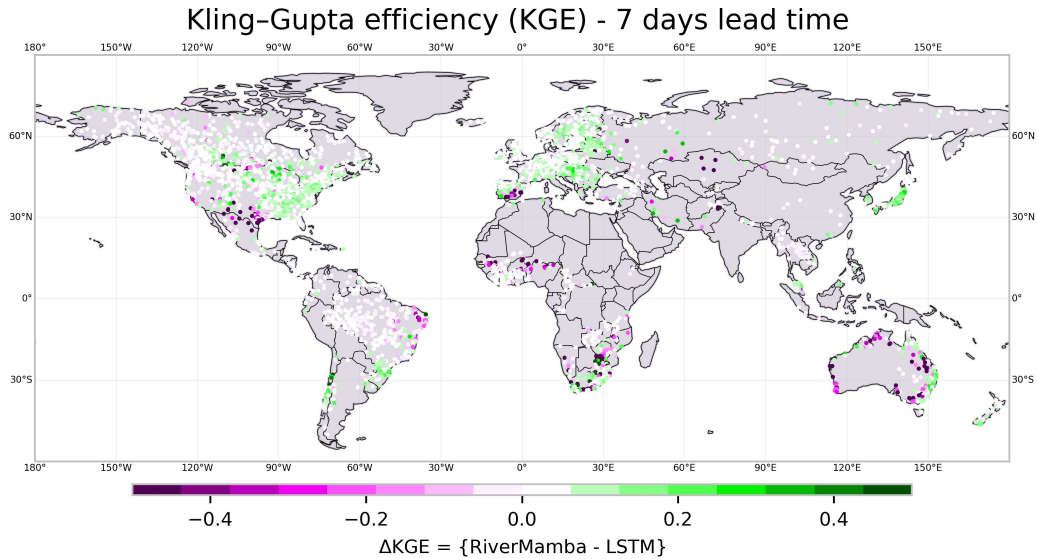


Figure 29: Comparison of KGE between RiverMamba and LSTM on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample). RiverMamba improves over LSTM in 73% of the stations ($P=3364$).

Confusion matrix - 4 days lead time (P=3366)

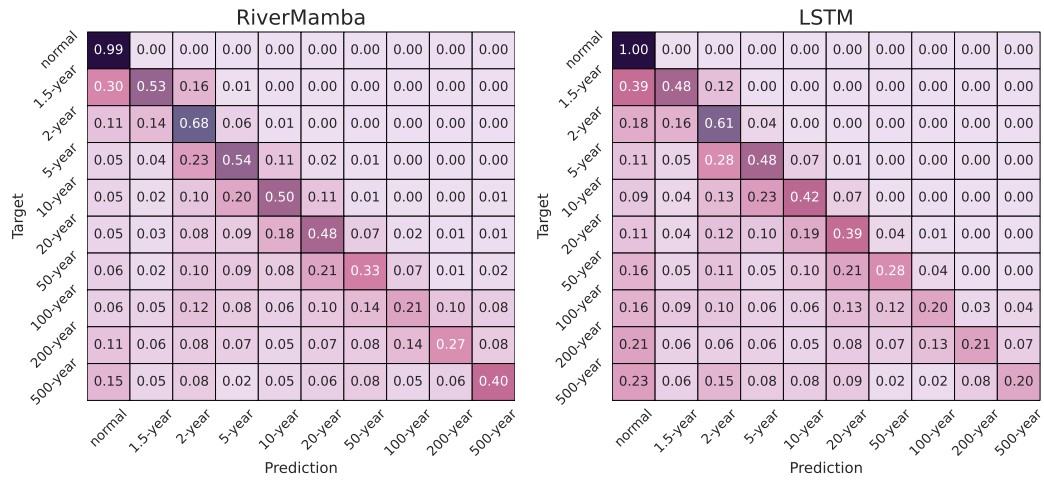


Figure 30: Comparison to LSTM on GloFAS reanalysis (test set 2021-2024 temporally out-of-sample).

K.4 Additional results on gauged GRDC

In this section, we plot additional results for the experiments on GRDC observational river discharge. In Figs. 31-34, we report the results for MAE, RMSE, R2, and R metrics with lead time. Figs. 35 and 36 show the confusion matrix for RiverMamba, LSTM, and GloFAS. In Figs. 37-42, we report the results of F1-score, Precision, and Recall metrics for different return periods and lead times. Finally, in Figs. 43-45, we compare the results between RiverMamba, LSTM, and GloFAS for F1-score and KGE metrics spatially.

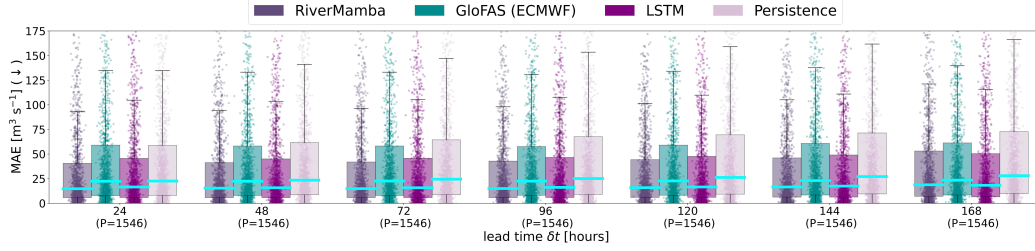


Figure 31: MAE of the river discharge forecasting with different lead time on GRDC observations (test set 2021-2023 temporally out-of-sample).

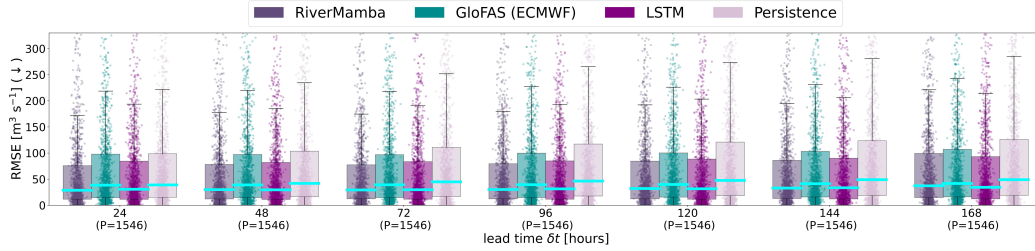


Figure 32: RMSE of the river discharge forecasting with different lead time on GRDC observations (test set 2021-2023 temporally out-of-sample).

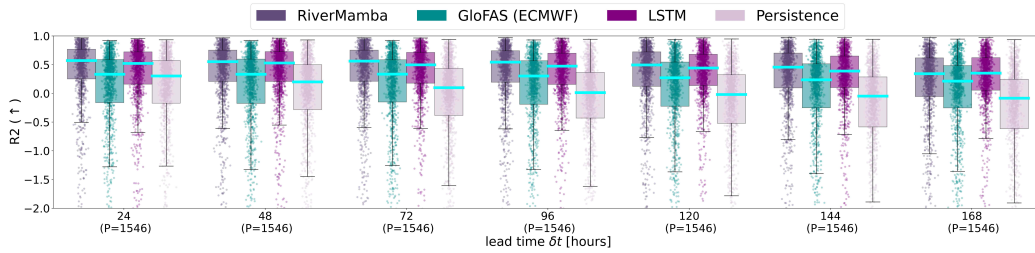


Figure 33: R^2 (NSE) of the river discharge forecasting with different lead time on GRDC observations (test set 2021-2023 temporally out-of-sample).

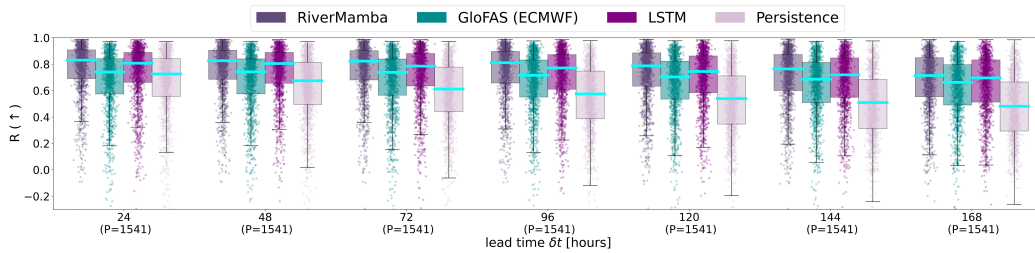


Figure 34: Pearson correlation (R) of the river discharge forecasting with different lead time on GRDC observations (test set 2021-2023 temporally out-of-sample).

Confusion matrix - 4 days lead time (P=1552)

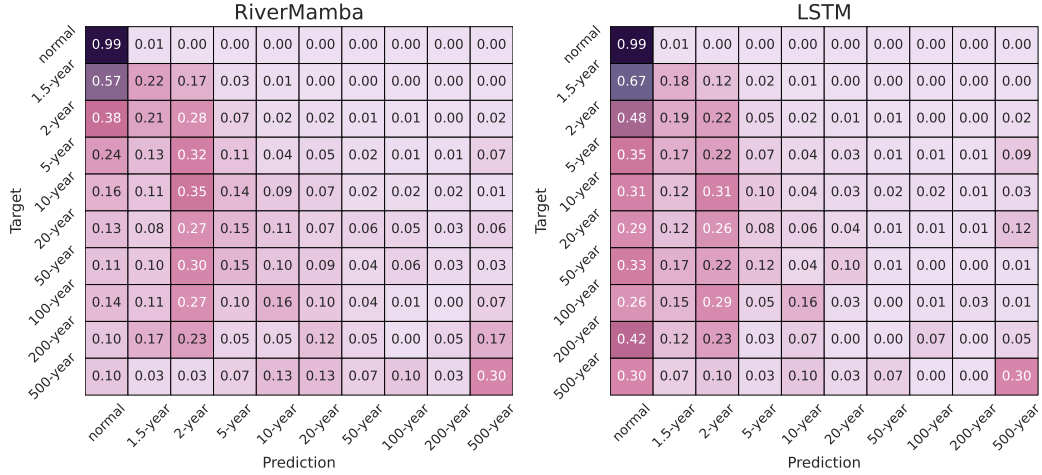


Figure 35: Comparison to LSTM on GRDC observations (test set 2021-2024 temporally out-of-sample).

Confusion matrix - 4 days lead time (P=1551)

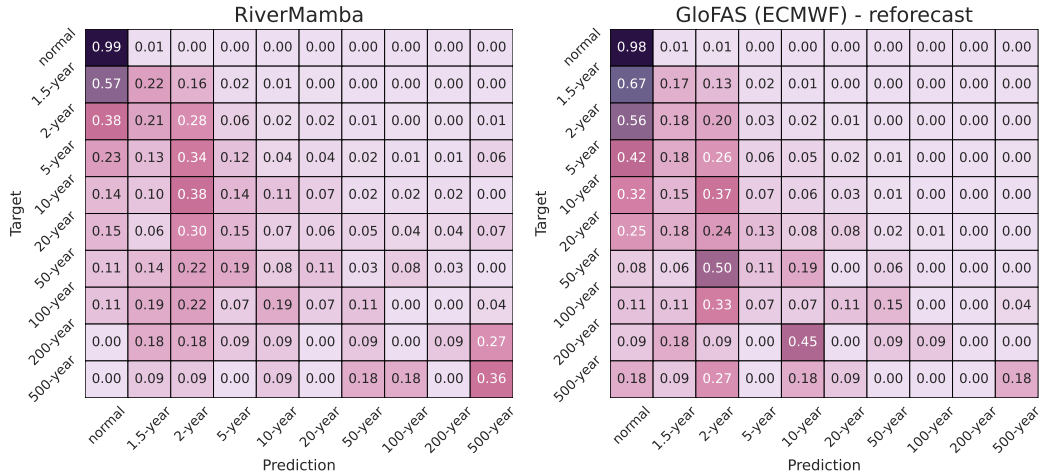


Figure 36: Comparison to GloFAS (ECMWF) - reforecast on GRDC observations (test set 2021-2023 temporally out-of-sample).



Figure 37: F1-score of flood forecasting for different return periods and lead time on GRDC observations (test set 2021-2023 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 38: Precision of flood forecasting for different return periods and lead time on GRDC observations (test set 2021-2023 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 39: Recall of flood forecasting for different return periods and lead time on GRDC observations (test set 2021-2023 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.



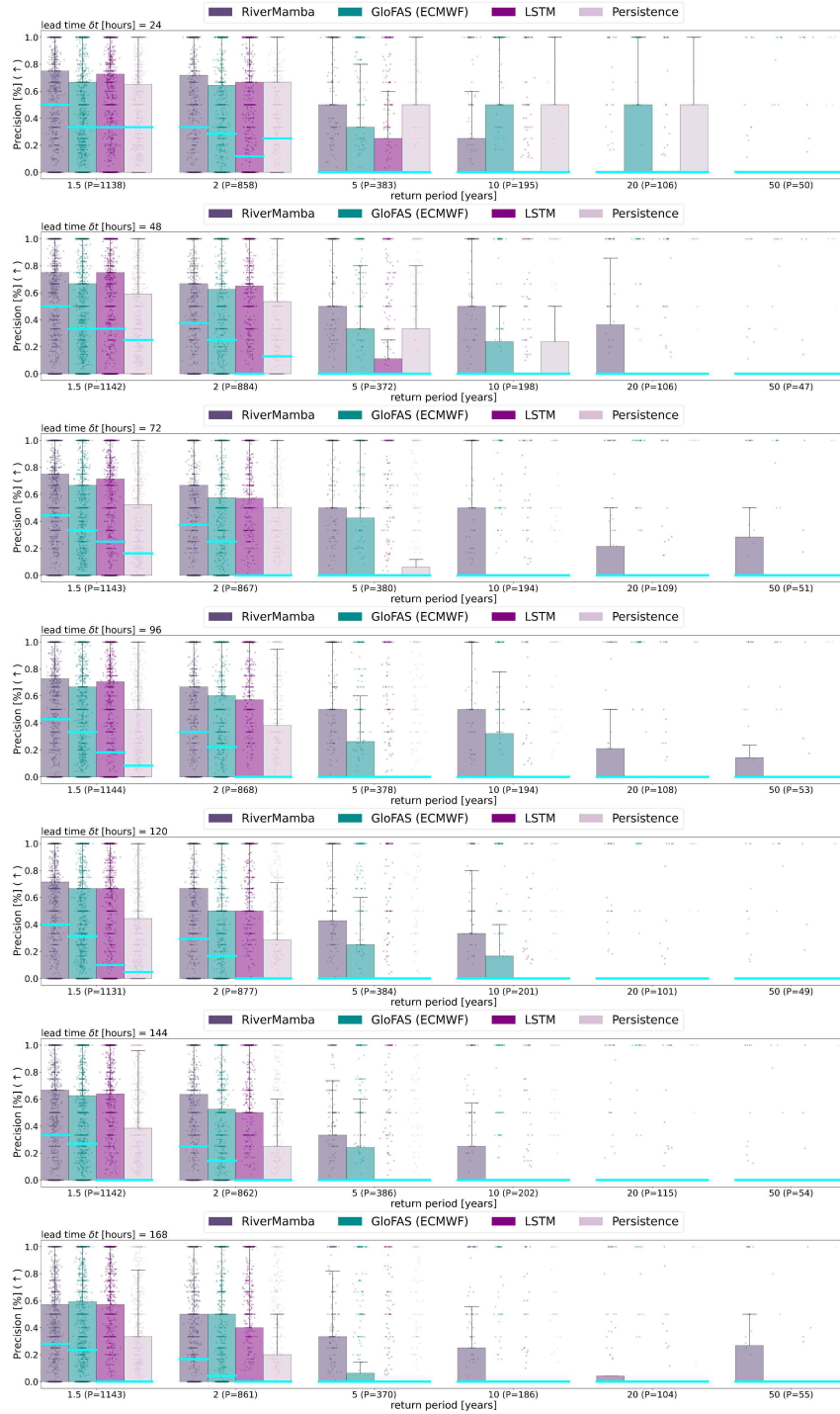


Figure 41: Precision of flood forecasting for different lead time and return periods (1.5 - 50 years) on GRDC observations (test set 2021-2023 temporally out-of-sample). Distribution quartiles are displayed in boxes, and the entire range excluding outliers is displayed in whiskers. The median score for the model is shown by the cyan line in the box.



Figure 42: Recall of flood forecasting for different lead time and return periods (1.5 - 50 years) on GRDC observations (test set 2021-2023 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.

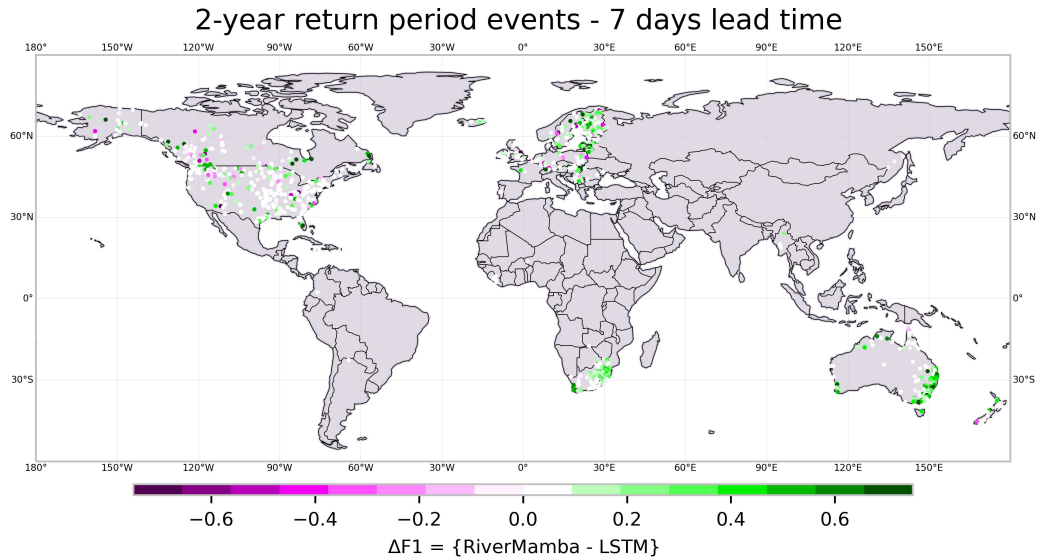


Figure 43: Comparison of F1-score between RiverMamba and LSTM on GRDC observations for the 2-year return period events (test set 2021-2023 temporally out-of-sample). RiverMamba improves over LSTM in 42% of the stations (P=861) and is better or equally better in 86% of the stations.

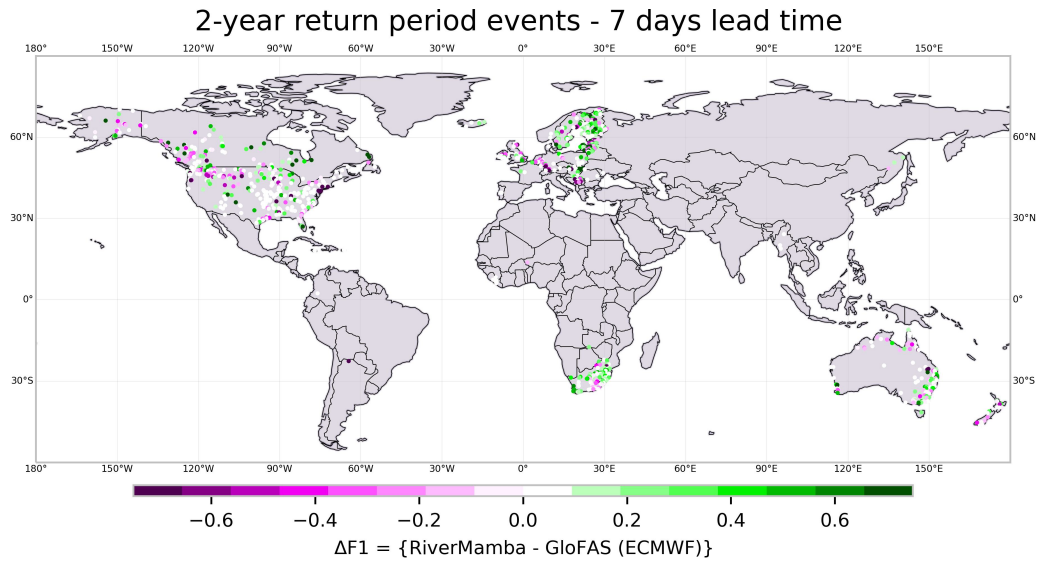


Figure 44: Comparison of F1-score between RiverMamba and GloFAS reforecast on GRDC observations for the 2-year return period events (test set 2021-2023 temporally out-of-sample). RiverMamba improves over GloFAS reforecast in 38% of the stations (P=861) and is better or equally better in 73% of the stations.

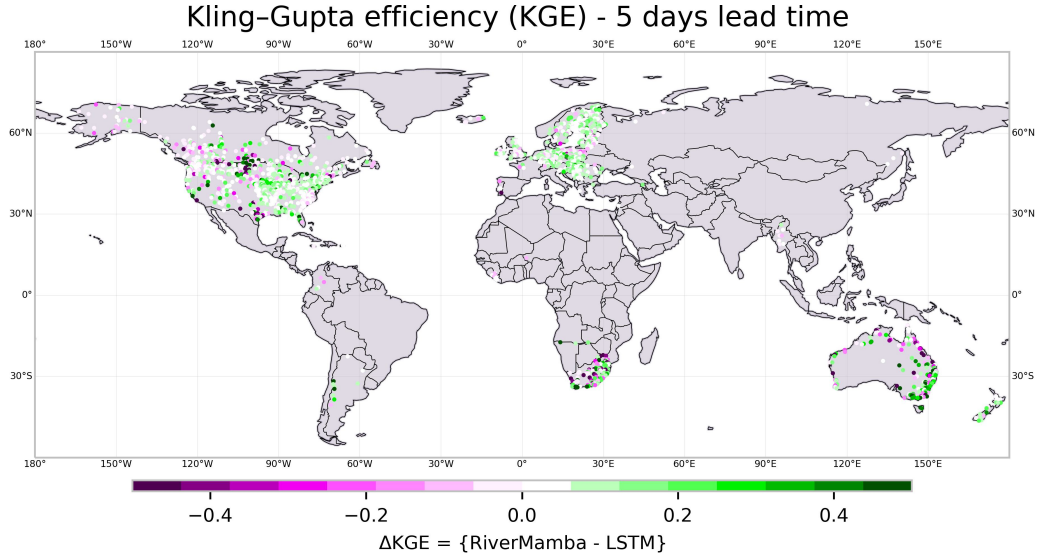


Figure 45: Comparison of KGE between RiverMamba and LSTM on GloFAS reanalysis (test set 2021-2023 temporally out-of-sample). RiverMamba improves over LSTM in 68% of the stations (P=1542).

K.5 Comparison to operational GloFAS forecasts on gauged GRDC

Here, we compare to the archival operational forecast from GloFAS ECMWF (Sec. H.4). This is different from the GloFAS reforecast and gives a more realistic assessment of the physics-based model.

Table 20: Results on GRDC gauged stations. (\pm) denotes the standard deviation for 3 runs.

Test (2023-2024)				
Model	MAE (\downarrow)	R2 (\uparrow)	KGE (\uparrow)	F1-score (\uparrow)
GloFAS*	65.35	-0.0063	0.0439	0.1795
LSTM	53.07 \pm 0.39	-0.0005 \pm 0.0001	0.3147 \pm 0.0026	0.1120 \pm 0.0065
RiverMamba	49.32\pm0.18	0.0001\pm0.0001	0.3821\pm0.0076	0.2358\pm0.0211

*GloFAS operational forecast [7]

L Case studies of extreme flood events

L.1 2021 Western Europe flood

In this section, we present the daily river discharge at a gauge station on the Sauer River—located at the border of Germany, France, and Luxembourg—for the year 2021. Shown is the river discharge signal as predicted by RiverMamba, LSTM, and GloFAS reanalysis, and compared against GRDC observations. Particular attention is given to the extreme flood event in July 2021 [25, 26], highlighted by the grey-shaded area in Fig. 46. To illustrate the meteorological drivers of this flood, we also show 7-day precipitation from ERA5 reanalysis and ECMWF HRES forecasts in Fig. 47.

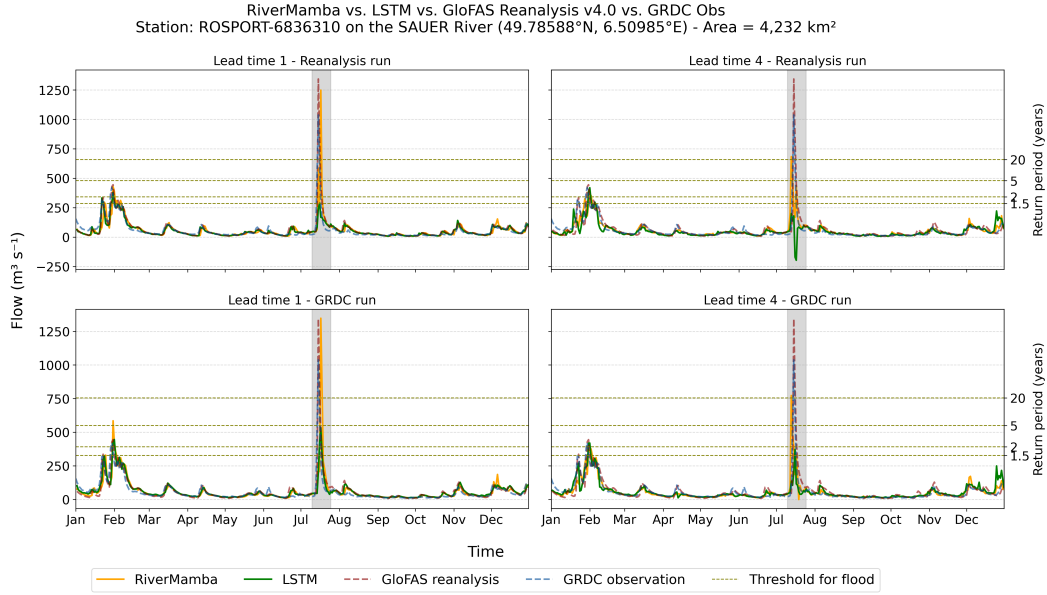


Figure 46: River discharge of the Sauer river in 2021 for RiverMamba (orange), LSTM (green), GloFAS reanalysis (dashed red), and GRDC observation (dashed green). The grey shaded area highlights the 2021 Germany flood between July 10 to July 20. The olive dashed lines represent the 1.5, 2, 5, and 20-year return periods calculated over reanalysis and GRDC observation data, respectively.

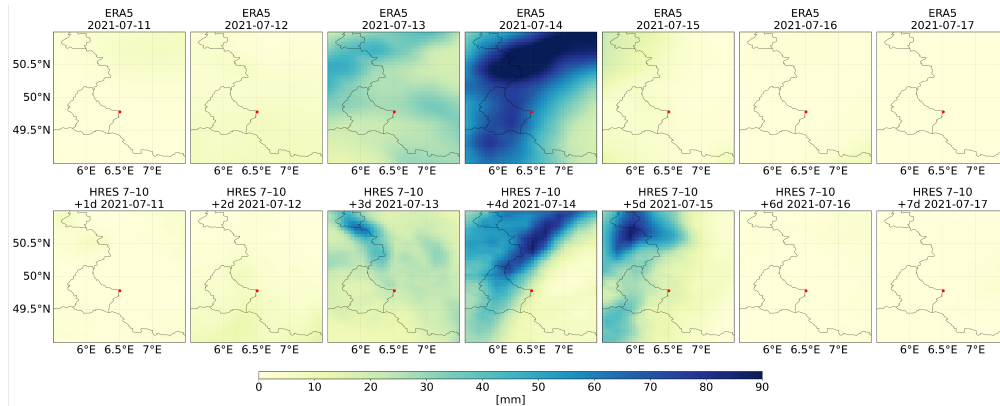


Figure 47: Daily total precipitation from 2021-07-11 to 2021-07-17 during the 2021 flood event in the target domain. First row shows precipitation as simulated by ERA5 Reanalysis and second row is the ECMWF HRES forecast issued at 2021-07-10 with 7-day lead time. The red dot is the location of the river discharge gauge station.

L.2 2024 Southeast Europe floods

In the following sections, we compare the daily flood severity map from GloFAS reanalysis as a ground truth and RiverMamba model at big flood events with different causes in 2024 and from different places around the Earth. These maps are usually used in the operational flood forecast service like GloFAS to provide a quick overview of the ongoing and upcoming flood events. As shown from these flood severity maps, RiverMamba can provide useful flood risk information at high spatial resolution to support decision-making. To the best of our knowledge, RiverMamba is the first AI model to demonstrate strong performance in predicting flood return periods globally under varying climate conditions at 5 km resolution across Europe, USA, Africa, and China. This further demonstrates its potential as a valuable component of operational flood early warning systems. It is important to note that the quality of the ECMWF HRES data driving the forecast is a key factor influencing RiverMamba performance.

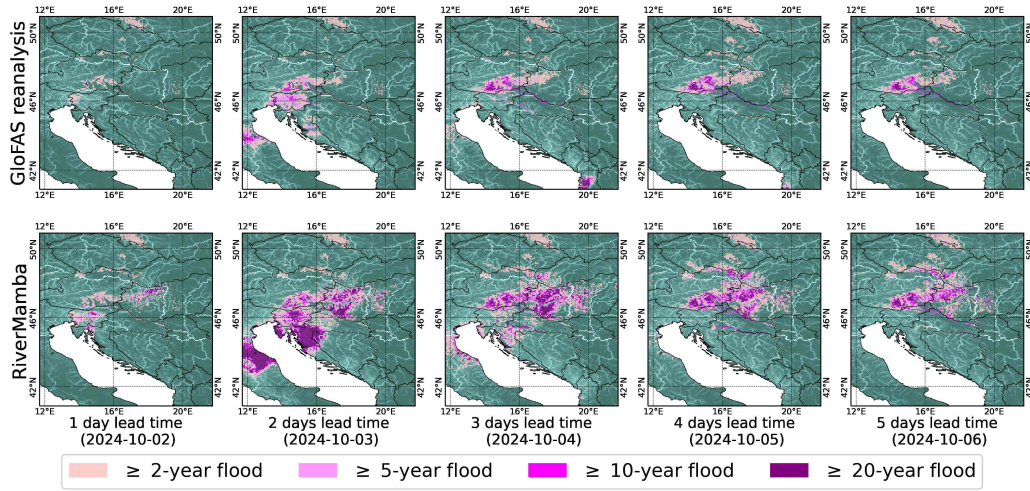


Figure 48: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the Southeast European flood in October 2024. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days.

L.3 2024 Central European floods

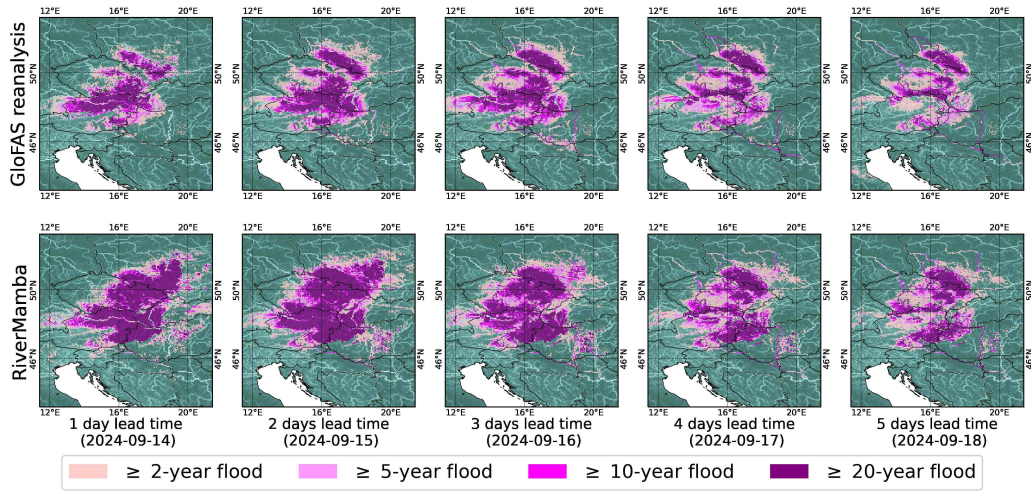


Figure 49: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the Central European floods event in 2024 [27]. In September 2024, Storm Boris brought record-breaking rainfall to Central Europe, causing devastating floods across Austria, the Czech Republic, Poland, Romania, Slovakia, Germany, and Hungary. Studies indicate that climate change doubled the likelihood and increased the intensity of such extreme rainfall events, highlighting the growing impact of global warming on severe weather patterns. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days.

L.4 2024 Spanish floods

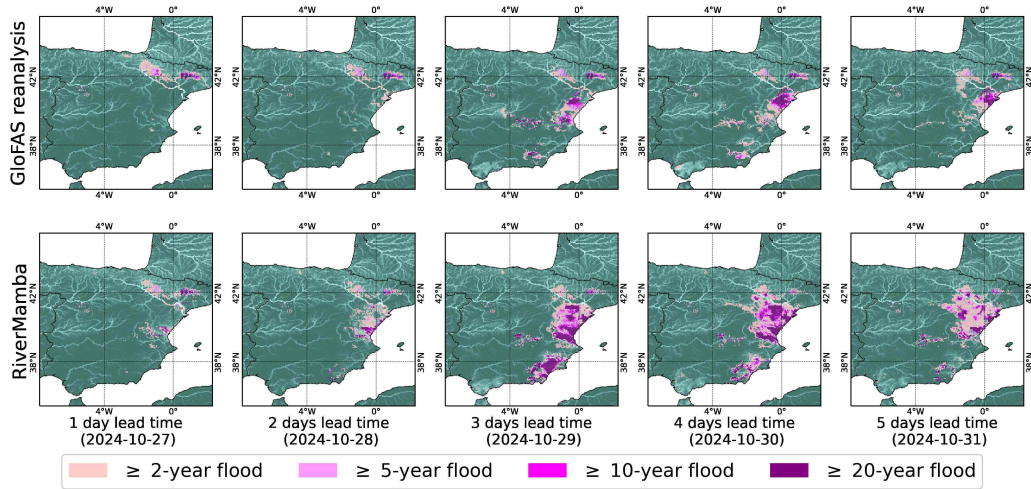


Figure 50: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the Spanish flood event in October 2024. The flood event primarily affecting the Valencia region, was caused by a cold drop (DANA) weather system intensified by climate change. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days.

L.5 2024 Saarland Germany flood

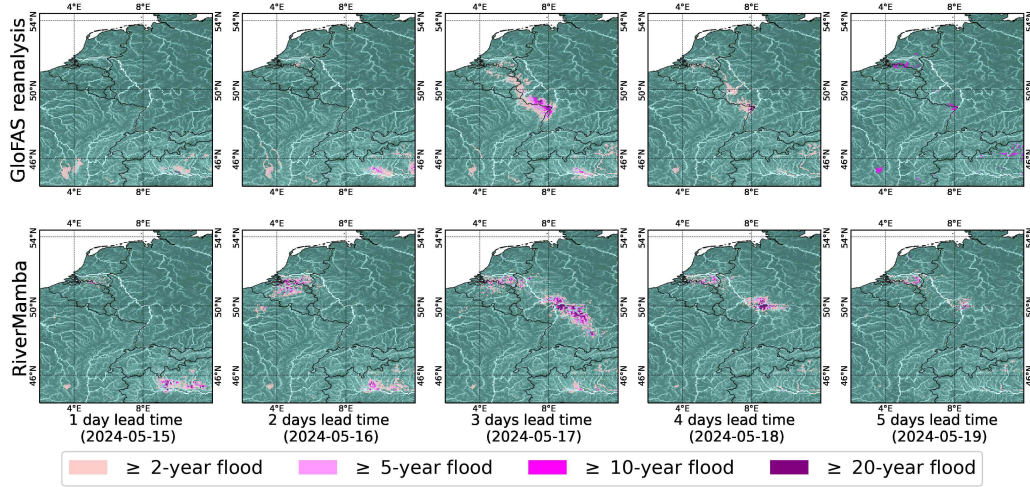


Figure 51: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the Saarland Germany flood event in May 2024. It was caused by thunderstorms and extreme rainfall, resulting in deadly floods and landslides across Saarland and Rheinland-Pfalz. However, RiverMamba predicted this flood event at a nearby location in Saarland, and the reason could be attributed to the inaccurate real-time ECMWF HRES forecast that drives the flood forecasting. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days.

L.6 2024 Kenya-Tanzania flood

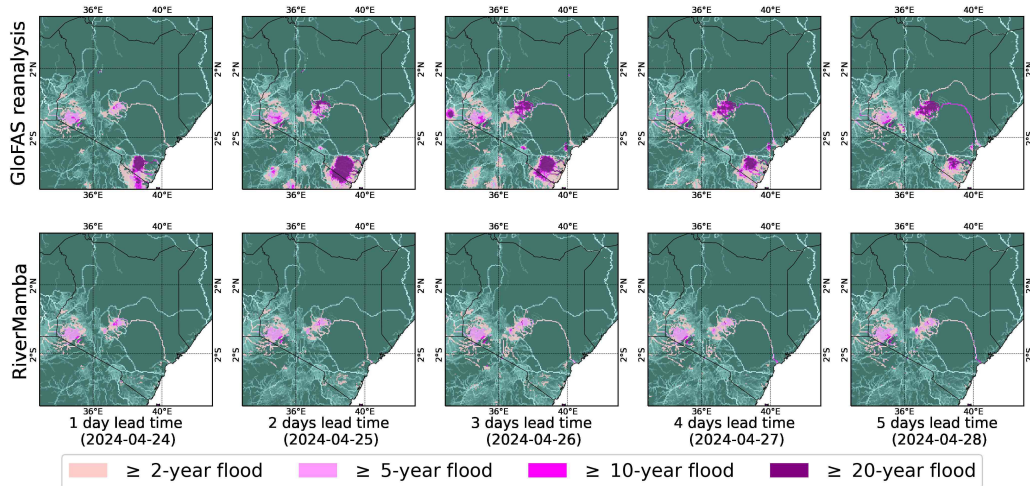


Figure 52: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the Kenya-Tanzania flood event in April 2024. This was the consequence of a combination of El Niño and a positive Indian Ocean Dipole, resulting in deaths, widespread displacement, and significant infrastructure damage. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days. Here points with less than $1 \text{ m}^3/\text{s}$ discharge have been removed to erase artifacts on desert grids.

L.7 2024 California flood

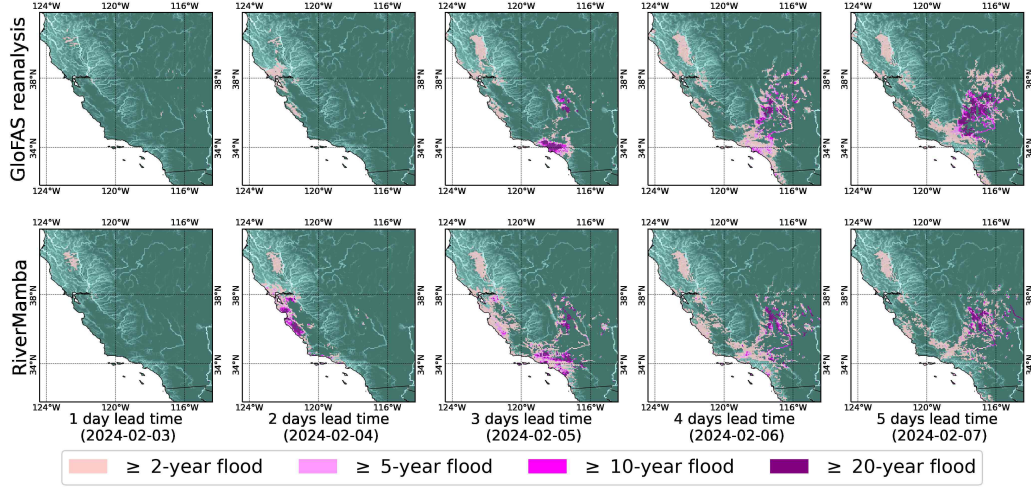


Figure 53: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the California, USA flood in February 2024. It was caused by two powerful atmospheric rivers. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days. Here points with less than $0.01 \text{ m}^3/\text{s}$ discharge have been removed to erase artifacts on desert grids.

L.8 2024 Central-South China floods

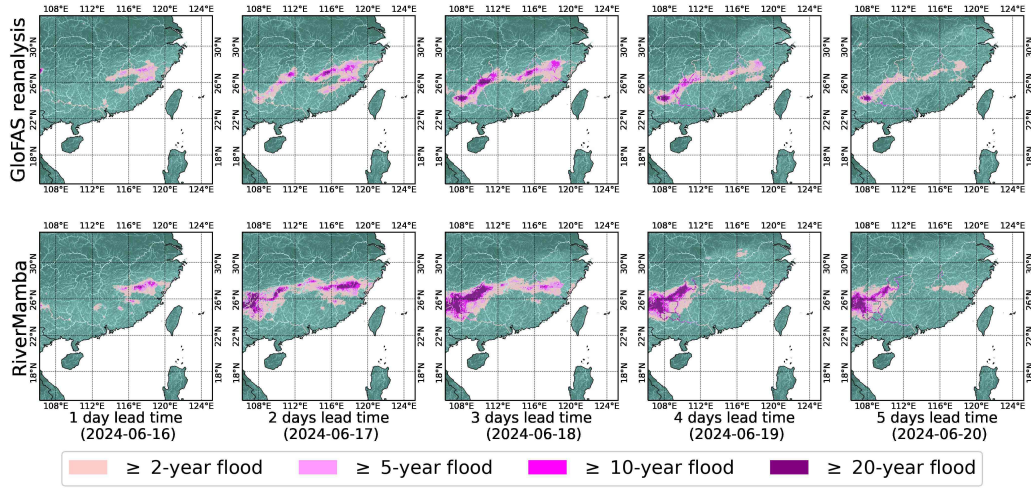


Figure 54: Comparison between GloFAS reanalysis (top, used as reference) and RiverMamba forecast (bottom) during the Central-South China flood event in June 2024 due to unprecedented rainfall. Shown are flood severity maps at 5-km resolution where each panel shows flood extent at different lead times from 1 to 5 days.

M Code and data availability

The code of RiverMamba and processing scripts are available on GitHub at https://github.com/HakamShams/RiverMamba_code. The pre-processed data used in this study are available at <https://doi.org/10.60507/FK2/T8QYWE> [28]. 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.

N Broader impacts

Extreme flood events, characterized by longer return periods, are expected to become more frequent and intense due to climate change. Traditional hydrology models often struggle to accurately predict such events. To improve flood detection, it is crucial to develop computationally efficient and precise deep learning models capable of forecasting key hydrological variables, such as river discharge. In this study, we demonstrated the potential of RiverMamba for predicting extreme riverine floods and in Appendix Sec. L, we presented case studies for extreme flood events. We see this as the primary motivation for developing RiverMamba. However, it is important to acknowledge that early warning systems may sometimes fail, leading to inaccurate forecasts. This limitation should be considered when deploying early warning systems.

References

- [1] 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.
- [2] 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.
- [3] Gianpaolo Balsamo, Anton Beljaars, Klaus Scipal, Pedro Viterbo, Bart van den Hurk, Martin Hirschi, and Alan K. Betts. A revised hydrology for the ecmwf model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, 10(3):623 – 643, 2009.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Jiawei Zhuang, raphael dussin, André Jüling, and Stephan Rasp. JiaweiZhuang/xESMF: v0.3.0 Adding ESMF.LocStream capabilities, March 2020.
- [10] 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.
- [11] 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.

- [12] 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.
- [13] 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.
- [14] Bernhard Lehner and Günther Grill. Global river hydrography and network routing: baseline data and new approaches to study the world’s large river systems. *Hydrological Processes*, 27(15):2171–2186, 2013.
- [15] Simon Linke, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, Vicente Burchard-Levine, Sally Maxwell, Hana Moidu, et al. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific data*, 6(1):283, 2019.
- [16] Martin Gauch, Frederik Kratzert, Oren Gilon, Hoshin Gupta, Juliane Mai, Grey Nearing, Bryan Tolson, Sepp Hochreiter, and Daniel Klotz. In defense of metrics: Metrics sufficiently encode typical human preferences regarding hydrological model performance. *Water Resources Research*, 59(6):e2022WR033918, 2023. e2022WR033918 2022WR033918.
- [17] 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.
- [18] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.
- [19] 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.
- [20] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, June 2024.
- [21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [22] 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.
- [23] Jian Zhang, Sei-ichiro Kamata, and Yoshifumi Ueshige. A pseudo-hilbert scan algorithm for arbitrarily-sized rectangle region. In *Advances in Machine Vision, Image Processing, and Pattern Analysis*, pages 290–299, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [24] Grey Nearing. Global prediction of extreme floods in ungauged watersheds, December 2023.
- [25] 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.
- [26] Magdalena Kracheletz, Ziyu Liu, Anne Springer, Jürgen Kusche, and Petra Friederichs. Would the 2021 western europe flood event be visible in satellite gravimetry? *Journal of Geophysical Research: Atmospheres*, 130(3):e2024JD042190, 2025. e2024JD042190 2024JD042190.
- [27] C Hauer, M Paster, U Pulg, T Ofenböck, and H Habersack. Critical flows at the wien river during the 1000-years event in september 2024—causes, consequences and possible management options for urban river flood management. *Natural Hazards*, pages 1–13, 2025.
- [28] 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.