Aligning and Assimilating Multi-source Data for Flood Forecasting

Kun Wang[†1,2], Gabriele Bertoli[†3,4], Sibo Cheng[5], Kai Schröter[6], Enrica Caporali[3], Matthew D. Piggott[2], Yanghua Wang[1,2], and Rossella Arcucci[*2,4]

[1] Resource Geophysics Academy, Imperial College London, South Kensington, London SW7 2AZ, UK
[2] Department of Earth Science and Engineering, Imperial College London, South Kensington, London SW7 2AZ, UK
[3] Department of Civil and Environmental Engineering, University of Florence, via di Santa Marta 3, Firenze, Italy
[4] Data Science Imperial, Imperial College London, London, UK
[5] CEREA, ENPC and EDF R&D, Institut Polytechnique de Paris, Île-de-France, France
[6] Leichtweiß-Institute for Hydraulic Engineering and Water Resources, Division Hydrology and River Basin Management, Technische Universität Braunschweig, Beethovenstr. 51a, 38106 Braunschweig, Germany

Abstract. Floods are among the most destructive natural hazards worldwide and frequently cause severe economic losses and significant loss of life. Therefore, reliable and timely forecasting plays a crucial role in disaster risk reduction. In recent years, approaches based on machine learning and data assimilation have attracted increasing attention for this purpose. However, a fundamental challenge remains. Predictive models generally require large volumes of high quality data, yet such data are often scarce, spatially and temporally limited, and heterogeneous in structure. This limitation substantially restricts the development of advanced flood forecasting methods. To address this problem, we elaborate three independent datasets derived from different sources, namely EFAS [27], EMO-1 [28], and LamaH-CE [15], which include meteorological, remote sensing, hydrological, and topographic information. Based on these datasets, we design two test cases consisting of a two-dimension forecasting experiment and a data assimilation experiment. Overall, the elaborated datasets and test cases provide a solid foundation for advancing flood forecasting research using machine learning and data assimilation techniques.

Keywords: Flood forecasting · Machine learning · Data assimilation.

1 Introduction

Floods, as a pervasive natural hazard, are characterized by their rapid onset and severe impacts, leading to substantial loss of life and significant economic

[†] Co-first authors.
[*] Corresponding author.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

damages worldwide each year [25, 2, 12]. Flood forecasting is one of the key approaches to mitigating flood impacts, placing critical emphasis on the timeliness and accuracy of predictive models [24, 7, 13]. In recent years, machine learning methods have been increasingly adopted for flood forecasting [24, 10, 4]. This growing adoption is largely due to the strong capability of machine learning models to capture complex nonlinear relationships, using a lower amount of variables (with respect to process based modelling), making them particularly suitable for flood forecasting as a nonlinear regression problem. Furthermore, with advances in computational hardware, machine learning models provide faster forecasting speeds compared to traditional hydrological models [36, 37]. Consequently, machine learning-based approaches are well-positioned to meet the two essential requirements of flood forecasting: accuracy and timeliness. In addition, there is a growing trend of integrating DA techniques with machine learning methods for flood forecasting [20, 14, 5, 19]. In this context, DA provides an effective means of combining observations from river gauging stations with predictive models, thereby enhancing forecasting accuracy. However, both machine learning and DA are inherently data-driven approaches, and their performance strongly depends on the quality and quantity of available data. In practice, flood-related datasets often face critical limitations, including restricted accessibility due to the lack of open data, inconsistencies in data formats and types, and pronounced spatio-temporal constraints. Consequently, the availability of high-quality datasets is essential for supporting the design and advancement of next-generation flood forecasting models.

1.1 Related Work and Contribution

Currently, widely used datasets for flood forecasting can generally be categorized into two types: simulated datasets and observational datasets. Simulated datasets are typically generated using hydrological models driven by meteorological, hydrological, and topographic inputs. A representative example is the Global Flood Awareness System (GloFAS) dataset [11], which provides gridded river discharge data generated by the LISFLOOD hydrological model. GloFAS offers a temporal resolution of 1 day and a spatial resolution of $0.05^\circ \times 0.05^\circ$, delivering acceptable spatio-temporal coverage and encompassing almost the entire globe except Antarctica, thereby serving as a valuable resource for flood forecasting studies. Nonetheless, several limitations persist. Owing to its relatively coarse spatial resolution, simulated discharges for smaller tributaries often diverge considerably from observed values, which is a common drawback of model-based datasets. Moreover, the daily temporal resolution is insufficient to capture the rapid dynamics of flood events, and the high computational cost of hydrological modeling further constrains its real-time applicability.

The second category is observational datasets derived directly from hydrological monitoring stations. For instance, the NRFA [9] provides discharge records from more than 1,000 gauging stations across the United Kingdom. While such datasets offer direct observations, they are also subject to limitations: station distribution is often sparse, and river discharge data cannot be obtained for

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

ungauged basins, restricting their spatial representativeness. In addition, some gauging stations suffer from missing data, which further reduces the reliability and completeness of such datasets.

Furthermore, several other datasets have been proposed for more specialized purposes. For instance, FloodCastBench [35] is specifically designed for machine learning applications to capture the dynamic processes of floods, while de Bruijn et al. [6] introduces a novel Twitter-based dataset for real-time global flood detection. Although these datasets are valuable for their intended tasks, their task-specific nature limits their generalization capability and restricts their applicability to broader machine learning applications in flood forecasting.

This study elaborates three different hydrological and meteorological datasets for machine learning and data assimilation applications in flood forecasting. These datasets provide a comprehensive representation of the complex processes involved in rainfall-runoff generation. Furthermore, two test cases for flood forecasting are defined using these datasets, providing reliable data resources and a solid research foundation for research on flood forecasting and related fields. The main contributions of this work are:

- A harmonized multi-source hydrological dataset across EFAS, EMO-1, and LamaH-CE.
- A reproducible two-case benchmarking framework with a systematic evaluation under operationally relevant constraints.

2 Data

2.1 Data sources

As introduced, flood forecasting tasks require multiple variables to describe the complex processes involved in rainfall-runoff generation. We selected three standalone datasets from different domains: one providing rich information on catchment characteristics through in-situ observations and geo-morphological indicators (LamaH-CE), another offering detailed meteorological variables essential for predicting rainfall-runoff patterns (EMO-1), and a third contributing historical simulations of spatially distributed runoff values to inform model calibrations and validations (EFAS). By carefully combining these datasets, we have created a rich and comprehensive resource that supports hydrological forecasting with machine learning and data assimilation techniques.

LamaH-CE LamaH-CE (LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe) is a large-sample catchment hydrology dataset designed for comparative hydrology and data-driven modeling in Central Europe. LamaH-CE compiles harmonized hydrometeorological time series and static descriptors for 859 gauged catchments spanning the upper Danube and all Austrian basins (including foreign upstream areas), covering 170,000 km^2 across nine countries. The domain ranges from continental lowlands to high-alpine,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

snow- and glacier-influenced headwaters, providing broad hydro-climatic variability that is crucial for testing the robustness of flood models under diverse forcing and response regimes. For each catchment, the dataset provides >60 attributes describing topography, climatology, hydrology, land cover/vegetation, soils and geology, alongside runoff series and meteorological forcings at both daily and hourly resolution; most series extend for >35 years, capturing multiple flood-rich periods and rare extremes. Runoff records are additionally annotated with >20 metadata attributes, including indicators of human impacts and data quality/completeness, enabling informed station selection and stratified validation. Unlike many CAMELS-style collections, LamaH-CE includes both independent basins and intermediate catchments and supplies river-network/topology information, which supports network-aware forecasting and upstream–downstream consistency checks. The long, high-frequency records, rich physiographic context, and benchmarking baseline model outputs make LamaH-CE particularly relevant for flood forecasting research, including event-based verification, regionalization/transfer learning, and comparisons between process-based and data-driven approaches.

EMO-1 We use EMO-1 (European Meteorological Observations, 1 arcmin) as the primary gridded meteorological forcing for flood forecasting experiments. EMO-1 is a pan-European, observation-based, multi-variable dataset produced within the Copernicus Emergency Management Service and distributed as 1 arcmin × 1 arcmin grids over Europe and surrounding areas, spanning 1 Jan 1990 to 31 Dec 2023 with annual updates (latest release v3.0.0). It provides daily fields of total precipitation, minimum and maximum air temperature, wind speed, solar radiation and water vapour pressure, and additionally offers 6-hourly precipitation and mean temperature, enabling both continuous simulation and event-focused analyses at sub-daily scales. The gridded fields are derived from quality-controlled station observations and generated using an Angular Distance Weighted interpolation approach, yielding spatially consistent forcing that better resolves sharp precipitation gradients in complex terrain than coarser products. Conceptually, EMO-1 is the higher-density successor to EMO-5 (5 km), which is no longer maintained, allowing direct comparison of the impact of forcing resolution on peak-flow timing and magnitude while retaining a common methodological lineage. For flood forecasting, this combination of long record length (supporting calibration on multiple extreme events), fine spatial detail (improving representation of convective and orographic rainfall hot spots), and sub-daily precipitation (supporting fast catchment response) is particularly valuable.

EFAS EFAS (European Flood Awareness System) is the Copernicus Emergency Management Service's pan-European flood forecasting and monitoring system. It provides complementary, added-value early warnings—especially for large, transboundary river basins—by running hydrological models over the "greater European domain" independent of administrative borders, and disseminating information to national/regional hydrological services and EU civil protection

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

actors to support preparedness before major floods. EFAS historical simulations provide a spatially continuous, model-consistent hydrological information across the EFAS domain, generated by forcing the open-source LISFLOOD hydrological model with gridded observational precipitation and temperature at 1 arcmin × 1 arcmin resolution (approximately 1.5 km at EFAS latitudes). The current dataset includes 6-hourly and daily gridded time series from 1 Jan 1992 to near real time (with a short latency for the most recent period), comprising core variables used in flood forecasting such as river discharge, surface and subsurface runoff, snow water equivalent, volumetric soil moisture (three soil layers) and a root-zone soil wetness index, complemented by static auxiliary layers (e.g., upstream area, elevation, soil depth and soil hydraulic capacities) that support catchment interpretation and feature engineering. This historical simulation archive is directly relevant because EFAS is an operational Copernicus flood forecasting and monitoring system that couples meteorological forecasts to a continental-scale hydrological model, producing probabilistic flood information and related products. EFAS historical simulations are a key reference for benchmarking forecast skill against a stable baseline, and training or calibrating data-driven models on spatially complete target fields where gauge coverage is sparse or heterogeneous.

2.2 Study region

The study region follows the LamaH-CE domain in Central Europe. LamaH-CE covers approximately 170,000 km^2 across nine countries (mainly Austria, Germany, the Czech Republic, Switzerland). Spatially, it includes the upper Danube up to the Austrian Slovak border, as well as all Austrian catchments and their adjacent upstream areas in neighboring countries. The domain spans pronounced hydro-climatic gradients, from lowland continental settings to high-alpine environments dominated by snow and ice, and includes a large interconnected river network partitioned into multiple river regions based on major tributary systems.

2.3 Data Preprocessing

Regarding the hydrological datasets, following the preprocessing in [30], we (i) compiled the LamaH-CE gauge discharge series into a single, consistent time-indexed dataset and (ii) extracted EFAS historical simulated discharge for the corresponding region. Each catchment was then paired with a representative EFAS river-grid cell using an automated outlet/basin matching procedure with basic quality checks (and a small number of manual fixes/removals where matches were implausible). Finally, both sources were harmonized to a common 6-hourly timeline by aggregating the higher-frequency observations to the EFAS time step and restricting the analysis to the period where both datasets overlap, yielding aligned observed–simulated discharge pairs for all selected sites.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

3 Methodology

3.1 CNNLSTM Forecasting Model

In flood forecasting, it is essential to predict the evolution of the river discharge field. Machine learning methods have gradually become an effective approach for modeling such spatiotemporal dynamics [24, 23]. In this study, we employ a CNN-based LSTM model to forecast the spatiotemporal variations of the river discharge, named CNNLSTM [26]. CNNLSTM is a neural network architecture that integrates convolutional neural networks (CNN) and long short term memory (LSTM), and is commonly used for two-dimension spatiotemporal prediction tasks. The CNN component is responsible for extracting spatial features, which are then provided as input to the LSTM to model temporal dependencies and memory effects, as shown in Eq. 1.

$$\mathbf{X}_{t:t+3} \xrightarrow{\text{CNN}} \mathbf{z}_{t:t+3} \xrightarrow{\text{LSTM}} \mathbf{X}^{pred}_{t+4:t+7}, \tag{1}$$

where $\mathbf{X}_{t:t+3}$ denotes the input sequence of fields from time step t to $t+3$, $\mathbf{z}_{t:t+3}$ represents the corresponding spatial feature representations extracted by the CNN, and $\mathbf{X}^{pred}_{t+4:t+7}$ denotes the predicted fields from time step $t+4$ to $t+7$.

During the training of the CNNLSTM model, the mean squared error (MSE) is adopted as the loss function. The Adam optimizer is employed with a learning rate of $1e^{-3}$. The model is trained for 200 epochs, and the model parameters corresponding to the minimum validation loss are selected as the final model.

3.2 Hybrid Data Assimilation Framework

In flood forecasting, both predictive model outputs and in situ station observations are typically available. To improve forecasting accuracy, data assimilation is employed to combine observational data with model predictions, and it is therefore widely used in flood forecasting applications [38, 33, 29, 32, 31]. Data assimilation methods can be classified into sequential data assimilation and variational data assimilation. Variational data assimilation can simultaneously utilize observations over a time window, resulting in better spatiotemporal consistency and satisfaction of dynamical equations. It provides improved balance in the analysis field, enables effective propagation of sparse observational information, does not rely on large ensemble sizes, and avoids sampling errors. In addition, the analysis results are smoother and more stable, and physical constraints can be incorporated. Therefore, in flood forecasting applications, variational data assimilation is commonly adopted to improve forecast accuracy and reliability [1, 18, 8]. Within variational data assimilation, the primary approaches are 3D-Var and 4D-Var. Compared with 4D-Var, 3D-Var requires less computational cost and provides faster execution, making it more suitable for the timeliness requirements of flood forecasting [34, 22, 21]. The 3D-Var method obtains the analysis field by minimizing a cost function, which effectively combines observational data with model forecasts to produce a state estimate that is closer to the true state. The cost function of 3D-Var is formulated as shown in Eq. 2,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

$$\mathcal{J}(\mathbf{X}) = \left\|\mathbf{X} - \mathbf{X}^b\right\|^2_{\mathbf{B}^{-1}} + \left\|\mathbf{y} - \mathcal{H}(\mathbf{X})\right\|^2_{\mathbf{R}^{-1}}, \tag{2}$$

where $\mathbf{X}$ denotes the state field, $\mathbf{X}^b$ represents the background field, and $\mathbf{B}$ is the background error covariance matrix. The vector $\mathbf{y}$ denotes the observations, $\mathcal{H}$ is the observation operator that maps the model state to the observation space, and $\mathbf{R}$ is the observation error covariance matrix.

However, in flood forecasting, the dimensionality of the state field is relatively high, which leads to substantial computational cost when applying 3D-Var. To address this issue, we employ a compression model that compresses the state field into a latent space. The 3D-Var procedure is then performed in this latent space, a strategy commonly referred to as latent data assimilation, thereby reducing computational cost and improving forecasting efficiency [3]. In this study, the selected compression method is truncated singular value decomposition (TSVD), as formulated in Eq. 3,

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathcal{E}_{TSVD}(\mathbf{X}) \\ \hat{\mathbf{X}} &= \mathcal{D}_{TSVD}(\tilde{\mathbf{x}}), \end{aligned} \tag{3}$$

where $\tilde{\mathbf{x}}$ represents the latent state vector, $\mathcal{E}_{\text{TSVD}}$ denotes the TSVD encoding operator, $\mathcal{D}_{\text{TSVD}}$ is its corresponding decoding (inverse) operator, and $\hat{\mathbf{X}}$ denotes the reconstructed state field.

Within this hybrid data assimilation framework, an LSTM model is employed as the forward model in the TSVD-compressed latent space. Compared with traditional hydrological models, the LSTM does not rely on explicitly inferred internal basin states; instead, it learns catchment memory directly from data. Consequently, it often outperforms physical-based hydrological models in basins with complex or poorly understood mechanisms [16, 17]. In addition, the computational speed of the LSTM is significantly higher than that of conventional hydrological models, thereby satisfying the accuracy and timeliness requirements of flood forecasting. The LSTM model used in this study is formulated as shown in Eq. 4.

$$\tilde{\mathbf{x}}^{pred}_{t+4:t+7} = \mathcal{M}_{LSTM}(\tilde{\mathbf{x}}_{t:t+3}), \tag{4}$$

where $\tilde{\mathbf{x}}_{t:t+3}$ denotes the latent state sequence from time step t to $t+3$ obtained after TSVD compression, $\mathcal{M}_{LSTM}(\cdot)$ represents the LSTM forward model, and $\tilde{\mathbf{x}}^{pred}_{t+4:t+7}$ denotes the predicted latent states from time step $t+4$ to $t+7$.

In summary, the hybrid data assimilation framework consists of TSVD as the compression model, an LSTM as the forward model, and the 3D-Var scheme. The cost function of the hybrid data assimilation framework can be written as shown in Eq. 5,

$$\mathcal{J}(\tilde{\mathbf{x}}_{t:t+3}) = \left\|\tilde{\mathbf{x}}_{t:t+3} - \tilde{\mathbf{x}}^b_{t:t+3}\right\|^2_{\tilde{\mathbf{B}}^{-1}} + \left\|\mathbf{y}_{t:t+3} - \mathcal{H}(\tilde{\mathbf{x}}_{t:t+3})\right\|^2_{\mathbf{R}^{-1}}, \tag{5}$$

where $\tilde{\mathbf{B}}$ denotes the background error covariance matrix in the latent space.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

4 Case Studies and Results

We designed two case studies: (i) a 2D forecasting scenario that leverages historical river-discharge simulations together with meteorological and morphological data for the study area, and (ii) a data-assimilation framework that leverages historical simulations and observed river-discharge measurements.

4.1 Experimental Setup

All experiments are conducted on a server equipped with a NVIDIA A100 GPU running Ubuntu 22.04.5. The models are implemented in Python 3.10 using PyTorch 2.1.0. During training, the dataset is split chronologically, with the first 70% used for training, the following 15% for validation, and the final 15% for testing. After training, the model is evaluated on the test dataset. The performance is assessed using MSE, coefficient of determination R^2, structural similarity index measure (SSIM) for two-dimension forecasting, and execution time. The MSE is adopted because the squaring operation emphasizes large deviations, making it particularly sensitive to peak discharge errors that are critical in flood detection, as defined in Eq. 6,

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i^{gt} - \mathbf{X}_i^{pred}\right)^2, \tag{6}$$

where N denotes the total number of grid points, $\mathbf{X}_i^{gt}$ represents the ground truth discharge value at the i_{th} grid point, and $\mathbf{X}_i^{pred}$ denotes the corresponding predicted discharge value. The R^2 is used to evaluate how well the predicted discharge field reproduces the overall variability and temporal dynamics of the ground truth, as defined in Eq. 7,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\mathbf{X}_i^{gt} - \mathbf{X}_i^{pred})^2}{\sum_{i=1}^{N}(\mathbf{X}_i^{gt} - \bar{\mathbf{X}^{gt}})^2}, \tag{7}$$

where $\bar{\mathbf{X}^{gt}}$ denotes the mean of the ground truth discharge values. Values closer to 1 indicate better agreement between prediction and reference. The SSIM is adopted to assess the spatial consistency between the predicted discharge fields and ground truth, as defined in Eq. 8,

$$\text{SSIM} = \frac{(2\mu_{gt}\mu_{pred} + C_1)(2\sigma_{gt,pred} + C_2)}{(\mu_{gt}^2 + \mu_{pred}^2 + C_1)(\sigma_{gt}^2 + \sigma_{pred}^2 + C_2)}, \tag{8}$$

where μ_{gt} and μ_{pred} denote the mean values of the ground truth and predicted fields, σ_{gt}^2 and σ_{pred}^2 denote the corresponding variances, and $\sigma_{gt,pred}$ denotes their covariance. C_1 and C_2 are small constants introduced to stabilize the division. Values closer to 1 indicate higher structural similarity.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

4.2 Case Study (i): Two-dimension Forecasting: EFAS and EMO-1

In the first case study, data from EFAS and EMO-1 are used to evaluate the performance of the proposed CNNLSTM model in two-dimension forecasting. The input variables consisted of normalized river discharge and soil wetness index from EFAS, as well as precipitation, solar radiation, temperature, water vapor pressure, wind speed, and digital elevation model (DEM) from EMO-1. The output variable is the predicted normalized river discharge field, and the ground truth is given by the river discharge from EFAS. The input sequence length is four time steps, corresponding to one day, and the model produced forecasts for the subsequent four time steps. The trained model is evaluated on the test dataset, and the performance metrics are averaged over the entire test set. The results yielded an MSE of 86.82, an R^2 of 0.99, an SSIM of 0.964, and an execution time of approximately 5 ms.

Based on the numerical metrics obtained from the test dataset, the CNNLSTM model demonstrates strong performance in the two-dimension discharge forecasting task. Specifically, the SSIM and R^2 values are both close to 1, indicating high spatial similarity and strong correlation between the predicted field and ground truth. The relatively low MSE further confirms the model's accuracy, particularly in capturing peak discharge variations. In addition, the execution time of approximately 5 ms highlights the computational efficiency of the model, meeting the timeliness requirements of flood forecasting.

To more clearly illustrate the results, we visualized the predicted fields at four consecutive time steps randomly selected from the test set, together with the corresponding ground truth fields and the absolute error between the predictions and ground truth, as shown in the Fig. 1.

As shown in Fig. 1, the absolute errors between the predicted and reference fields are generally small, indicating that the predicted discharge fields closely match the ground truth. Overall, the CNNLSTM model effectively leverages EFAS and EMO-1 data to achieve accurate and efficient two-dimension forecasting. Hydrologically, the high R^2 and SSIM suggest that the CNNLSTM reproduces the dominant spatiotemporal organization of the EFAS discharge field, consistent with coherent flood-wave propagation along the river network.

4.3 Case Study (ii): Data Assimilation: EFAS and LamaH-CE

In the second case study, data from EFAS and LamaH-CE are used to evaluate the performance of the proposed hybrid data assimilation framework. The input variable is the river discharge from EFAS, which is treated as the state field. The input sequence consisted of four consecutive time steps, and the model produced forecasts for the subsequent four time steps. After TSVD compression, the state field is transformed into a lower-dimensional state vector in the latent space and provided as input to the LSTM model. The LSTM output represented the background vector. Subsequently, river discharge observations from LamaH-CE are assimilated with the background vector to obtain the analysis vector through the data assimilation procedure. Finally, the analysis vector is reconstructed

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

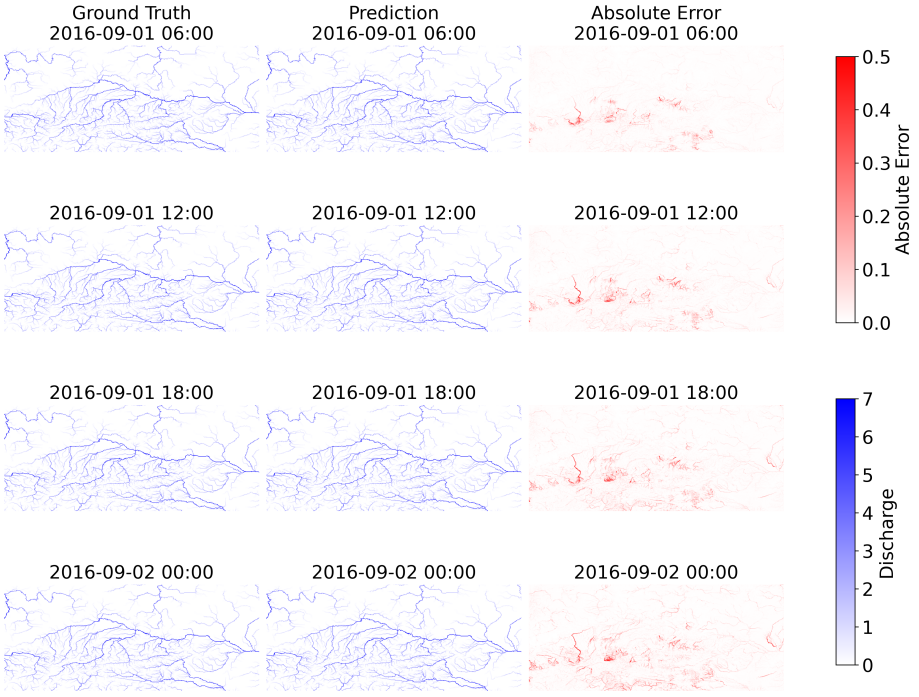

Fig. 1. Results of CNNLSTM at four consecutive time steps from 2016-09-01 06:00 to 2016-09-02 00:00. The first column shows the ground truth fields, the second column presents the corresponding predictions generated by the CNNLSTM model, and the third column illustrates the absolute error between the predictions and the ground truth.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

via the inverse TSVD to produce the analysis field, which constituted the final prediction, and the ground truth corresponded to the observations from the LamaH-CE dataset.

The proposed framework is evaluated on the test dataset, and the performance metrics are averaged over the entire test dataset. The numerical evaluation results are presented in Table 1.

Table 1. Averaged evaluation metrics of three methods for the data assimilation task on the test dataset. Values highlighted in red indicate the best-performing model for each metric.

Model	MSE ↓	R^2 ↑	Execution time (s) ↓
LSTM	1023.43	0.34	≈ 0.06
Hybrid data assimilation framework	356.73	0.83	≈ 20
EFAS	458.55	0.79	–

As shown in Table 1, the hybrid data assimilation framework results exhibits substantial improvements in both MSE and R^2 compared with the LSTM predictions and the original EFAS data. This indicates that hybrid data assimilation framework enhances forecasting accuracy and reliability in flood forecasting. In addition, hybrid data assimilation framework requires only approximate 20 s to complete a one-day forecast, demonstrating computational efficiency that satisfies the timeliness requirements of operational flood forecasting.

In addition, three months of results are randomly selected from the test dataset for visualization, as shown in Fig. 2.

As shown in the figure, the hybrid data assimilation framework results outperform both the LSTM and EFAS in terms of MSE and R^2, which is consistent with the averaged numerical evaluation metrics obtained on the test dataset. These results indicate that the proposed hybrid data assimilation framework achieves higher accuracy and reliability in flood forecasting by effectively integrating the river discharge field provided by EFAS with the observations from the LamaH-CE. Moreover, the behaviour of MSE and R^2 across the highlighted periods suggest that the assimilation step helps keep the forecast closer to the observed hydrograph evolution, even during high-flow dynamics. This points to a reduced sensitivity to mismatches in antecedent conditions and flow propagation when EFAS-based LSTM forecast are periodically constrained by discharge observations with data assimilation framework here provided.

5 Conclusion

This study addresses the data limitations that constrain the development of machine learning and data assimilation approaches for flood forecasting. We elaborate three independent datasets, namely EFAS, EMO-1, and LamaH-CE, which integrate meteorological, hydrological, remote sensing, and topographic data to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

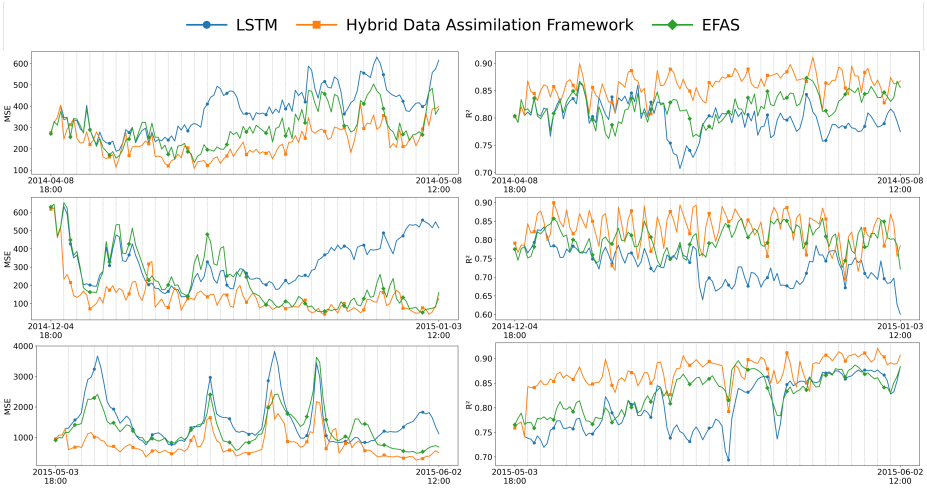

Fig. 2. Forecasting performance for three representative periods (April 2014, December 2014, and May 2015). Blue circles denote the LSTM results, orange squares denote the hybrid data assimilation framework results, and green diamonds denote the EFAS results. The left column shows MSE and the right column shows R^2. Vertical dashed lines indicate the data assimilation times.

provide long-term, multi-source data resources. Based on these datasets, two experimental test cases are designed: a two-dimension forecasting task and a data assimilation task. These experiments enable systematic evaluation of predictive performance and the benefit of integrating model forecasts with observations, while simultaneously assessing the capability of machine learning and data assimilation approaches in flood forecasting. Regarding the 2D case study, because the reference fields are taken from EFAS, the reported scores should be interpreted primarily as the model's ability to reproduce EFAS-consistent spatiotemporal discharge patterns from meteorological forcing and physiographic descriptors, rather than as a direct measure of agreement with independent observations. From a hydrological perspective, this is nevertheless a valuable result: it indicates that the CNNLSTM can act as an efficient surrogate of the EFAS discharge dynamics at the field level, preserving the large-scale organization of flow and its temporal evolution. This surrogate capability is promising for rapid scenario screening and real-time applications where computational cost is critical, and it motivates further work to assess generalization under out-of-sample events and to benchmark performance against gauge-based discharge observations (and/or alternative hydrological reanalyses) to quantify added value beyond reproducing the EFAS target. Regarding the data assimilation framework, a key limitation is that the current setup assimilates discharge only (with fixed TSVD/LSTM and error assumptions), so future work should examine sensitivity to the latent-space truncation and uncertainty specification and extend the framework to multi-source observations (e.g., soil moisture, water level, and potentially, weather

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

forecasts) and adaptive updating to further improve robustness across regimes and events. The datasets elaborated in this study, together with the proposed test cases, indeed establish a reproducible framework for benchmarking flood forecasting methods and support future research at the intersection of hydrology and data-driven modeling.

References

1. Alvarado-Montero, R., Schwanenberg, D., Krahe, P., Helmke, P., Klein, B.: Multi-parametric variational data assimilation for hydrological forecasting. Advances in Water Resources **110**, 182–192 (2017)
2. Apel, H., Aronica, G.T., Kreibich, H., Thieken, A.H.: Flood risk analyses—how detailed do we need to be? Natural hazards **49**(1), 79–98 (2009)
3. Arcucci, R., Mottet, L., Pain, C., Guo, Y.K.: Optimal reduced space for variational data assimilation. Journal of Computational Physics **379**, 51–69 (2019)
4. Bertoli, G., Schroeter, K., Arcucci, R., Caporali, E.: A hybrid machine learning framework for improved short-term peak-flow forecasting. arXiv preprint arXiv:2601.09336 (2026)
5. Boucher, M.A., Quilty, J., Adamowski, J.: Data assimilation for streamflow forecasting using extreme learning machines and multilayer perceptrons. Water Resources Research **56**(6), e2019WR026226 (2020)
6. de Bruijn, J.A., de Moel, H., Jongman, B., de Ruiter, M.C., Wagemaker, J., Aerts, J.C.: A global database of historic and real-time flood events based on social media. Scientific data **6**(1), 311 (2019)
7. Byaruhanga, N., Kibirige, D., Gokool, S., Mkhonta, G.: Evolution of flood prediction and forecasting models for flood early warning systems: A scoping review. Water **16**(13), 1763 (2024)
8. Ercolani, G., Castelli, F.: Variational assimilation of streamflow data in distributed flood forecasting. Water Resources Research **53**(1), 158–183 (2017)
9. Fry, M., Swain, O.: Hydrological data management systems within a national river flow archive (2010)
10. Ghorpade, P., Gadge, A., Lende, A., Chordiya, H., Gosavi, G., Mishra, A., Hooli, B., Ingle, Y.S., Shaikh, N.: Flood forecasting using machine learning: a review. In: 2021 8th international conference on smart computing and communications (ICSCC). pp. 32–36. IEEE (2021)
11. Grimaldi, S., Salamon, P., Disperati, J., Zsoter, E., Russo, C., Ramos, A., Carton De Wiart, C., Barnard, C., Hansford, E., Gomes, G., Prudhomme, C.: River discharge and related historical data from the global flood awareness system, v4.0 (2022). https://doi.org/10.24381/cds.a4fdd6b9, accessed on 22-Sep-2025
12. Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., Kanae, S.: Global flood risk under climate change. Nature climate change **3**(9), 816–821 (2013)
13. Jain, S.K., Mani, P., Jain, S.K., Prakash, P., Singh, V.P., Tullos, D., Kumar, S., Agarwal, S.P., Dimri, A.P.: A brief review of flood forecasting techniques and their applications. International journal of river basin management **16**(3), 329–344 (2018)
14. Jeong, M., Kwon, M., Cha, J.H., Kim, D.H.: High flow prediction model integrating physically and deep learning based approaches with quasi real-time watershed data assimilation. Journal of Hydrology **636**, 131304 (2024)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

15. Klingler, C., Schulz, K., Herrnegger, M.: Lamah-ce: Large-sample data for hydrology and environmental sciences for central europe. Earth System Science Data **13**(9), 4529–4565 (2021). https://doi.org/10.5194/essd-13-4529-2021, https://essd.copernicus.org/articles/13/4529/2021/
16. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (lstm) networks. Hydrology and Earth System Sciences **22**(11), 6005–6022 (2018)
17. Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G.: 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)
18. Lai, X., Liang, Q., Yesou, H., Daillet, S.: Variational assimilation of remotely sensed flood extents using a 2-d flood model. Hydrology and Earth System Sciences **18**(11), 4325–4339 (2014)
19. Lever, J., Cheng, S., Casas, C.Q., Liu, C., Fan, H., Platt, R., Rakotoharisoa, A., Johnson, E., Li, S., Shang, Z., et al.: Facing & mitigating common challenges when working with real-world data: The data learning paradigm. Journal of Computational Science **85**, 102523 (2025)
20. Li, X.L., Lü, H., Horton, R., An, T., Yu, Z.: Real-time flood forecast using the coupling support vector machine and data assimilation method. Journal of Hydroinformatics **16**(5), 973–988 (2014)
21. Liu, Y., Liu, J., Li, C., Liu, L., Wang, Y.: A wrf/wrf-hydro coupled forecasting system with real-time precipitation–runoff updating based on 3dvar data assimilation and deep learning. Water **15**(9), 1716 (2023)
22. Liu, Y., Liu, J., Li, C., Yu, F., Wang, W.: Effect of the assimilation frequency of radar reflectivity on rain storm prediction by using wrf-3dvar. Remote Sensing **13**(11), 2103 (2021)
23. Maspo, N.A., Bin Harun, A.N., Goto, M., Cheros, F., Haron, N.A., Mohd Nawi, M.N.: Evaluation of machine learning approach in flood prediction scenarios and its input parameters: A systematic review. In: IOP Conference Series: Earth and Environmental Science. vol. 479, p. 012038. IOP Publishing (2020)
24. Mosavi, A., Ozturk, P., Chau, K.w.: Flood prediction using machine learning models: Literature review. Water **10**(11), 1536 (2018)
25. Plate, E.J.: Flood risk and flood management. Journal of hydrology **267**(1-2), 2–11 (2002)
26. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
27. Smith, P., Pappenberger, F., Wetterhall, F., Del Pozo, J.T., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M., Baugh, C.: On the operational implementation of the european flood awareness system (efas). In: Flood forecasting, pp. 313–348. Elsevier (2016)
28. Thiemig, V., Gomes, G.N., Skøien, J.O., Ziese, M., Rauthe-Schöch, A., Rustemeier, E., Rehfeldt, K., Walawender, J.P., Kolbe, C., Pichon, D., Schweim, C., Salamon, P.: Emo-5: a high-resolution multivariable gridded meteorological dataset for europe. Earth System Science Data **14**(7), 3249–3272 (2022). https://doi.org/10.5194/essd-14-3249-2022, https://essd.copernicus.org/articles/14/3249/2022/
29. Wang, K., Bertoli, G., Cheng, S., Schröter, K., Caporali, E., Piggott, M.D., Wang, Y., Arcucci, R.: Ai-empowered latent four-dimensional variational data assimila-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20

tion for river discharge forecasting. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)
30. Wang, K., Bertoli, G., Schröter, K., Caporali, E., Piggott, M.D., Wang, Y., Arcucci, R.: Latent three-dimensional variational data assimilation with convolutional autoencoder and lstm for flood forecasting. In: Paszynski, M., Barnard, A.S., Zhang, Y.J. (eds.) Computational Science – ICCS 2025 Workshops. pp. 43–56. Springer Nature Switzerland, Cham (2025)
31. Wang, K., Bertoli, G., Schröter, K., Caporali, E., Piggott, M.D., Wang, Y., Arcucci, R.: Latent three-dimensional variational data assimilation with convolutional autoencoder and lstm for flood forecasting. In: International Conference on Computational Science. pp. 43–56. Springer (2025)
32. Wang, K., Cheng, S., Piggott, M.D., Dance, S.L., Wang, Y., Arcucci, R.: Latent data assimilation with non-explicit observation operator in hydrology. Quarterly Journal of the Royal Meteorological Society **151**(772), e5009 (2025)
33. Wang, K., D. Piggott, M., Wang, Y., Arcucci, R.: Neural network as transformation function in data assimilation. In: International Conference on Computational Science. pp. 322–329. Springer (2024)
34. Wang, Y., Min, J., Chen, Y., Huang, X.Y., Zeng, M., Li, X.: Improving precipitation forecast with hybrid 3dvar and time-lagged ensembles in a heavy rainfall event. Atmospheric research **183**, 1–16 (2017)
35. Xu, Q., Shi, Y., Zhao, J., Zhu, X.X.: Floodcastbench: A large-scale dataset and foundation models for flood modeling and forecasting. Scientific Data **12**(1), 431 (2025)
36. Yaseen, Z.M.: A new benchmark on machine learning methodologies for hydrological processes modelling: a comprehensive review for limitations and future research directions. Knowledge-Based Engineering and Sciences **4**(3), 65–103 (2023)
37. Zhao, X., Wang, H., Bai, M., Xu, Y., Dong, S., Rao, H., Ming, W.: A comprehensive review of methods for hydrological forecasting based on deep learning. Water **16**(10), 1407 (2024)
38. Ziliani, M.G., Ghostine, R., Ait-El-Fquih, B., McCabe, M.F., Hoteit, I.: Enhanced flood forecasting through ensemble data assimilation and joint state-parameter estimation. Journal of Hydrology **577**, 123924 (2019)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_20