https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Extending Medium-Range Global Flood Forecasts: The Google Global Flood Forecasting Model Version 2

Deborah Cohen [1], Rony Amira [1], Rom Aschner [1], Yuval Carny [1], Ben Feinstein [1], Hadas Fester [1], Shmulik Fronman [2], Martin Gauch [3], Oren Gilon [1], Rotem Green [1], Avinatan Hassidim [1], Daniel Klotz [4], Frederik Kratzert [4], Dan Korenfeld [1], Gila Loike [1], Amit Markel [1], Yossi Matias [5], Rotem Mayo [1], Asher Metzger [1], Benny Mosheyev [1], Aviel Niego [1], Stephanie Rees [6], Emily Reinstein [7], Amitay Sicherman [1], Guy Shalev [8], Omri Shefi [1], Yuval Shildan [1], Ido Zemach [1], Oleg Zlydenko [1], Grey Nearing [3,*]

[1]Google Research, Electra Tower, Yigal Alon St 98, Tel Aviv-Yafo, 6789141, Israel
[2]Google Research, 6 Pancras Sq, London N1C 4AG, United Kingdom
[3]Google Research, Brandschenkestrasse 110, 8002 Zürich, Switzerland
[4]Google Research, Graben 19, 1010 Wien, Austria
[5]Google Research, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 USA
[6]Google, 1-13 St Giles High St, London WC2H 8AG, United Kingdom
[7]Google, 111 8th Ave, New York, NY 10011 USA
[8]Google DeepMind, Electra Tower, Yigal Alon St 98, Tel Aviv-Yafo, 6789141, Israel

[*]**Corresponding Author:** Grey Nearing <nearing@google.com>

Abstract

This paper evaluates an updated flood forecasting system that significantly extends reliable lead times. We evaluated this updated model (v2) against the prior system (v1) and established third-party benchmarks across 1,223 global test basins. The primary finding is that the v2 system extends the reliable predictive horizon by 6 days in gauged basins and 1 day in ungauged basins relative to the v1 nowcast, as measured by the Nash Sutcliffe Efficiency. Along with this paper, we release an open-source codebase for training both the v1 and v2 forecast models with the open-source Caravan dataset.

1 Introduction

The hydrology research and operations communities are now several years into the adoption cycle for machine learning (ML) based simulation and prediction models. The initial phase of this adoption cycle was characterized by establishing benchmarks relative to traditional conceptual rainfall-runoff models (Kratzert et al., 2019; Mai et al., 2022; Arsenault et al., 2022), particularly in data-scarce and ungauged settings (Kratzert et al., 2019a; Prieto et al., 2019;

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Feng et al., 2023; Nearing et al., 2024). Since then, the field has matured as research diversified into questions related to interpretability (e.g., Kratzert et al., 2019; Lees et al., 2022; Başağaoğlu et al., 2022; De la Fuente et al., 2023; Dai et al., 2023; Hosseini et al., 2025; Rosati et al., 2025), uncertainty quantification (e.g., Klotz et al., 2021; Chaudhary et al., 2022; Nourani et al., 2022; Liu et al., 2023), data assimilation (e.g., Feng et al., 2020; Nearing et al., 2021; Deng et al., 2023; Sabzipour et al., 2023), and hybrid process-based machine learning (e.g., Hoedt et al., 2021; Jia et al., 2021; Tsai et al., 2021; Feng et al., 2022; Höge et al., 2022; Frame et al., 2023; Acuña Espinoza et al., 2025).

We use operational machine learning hydrology models for global-scale riverine flood forecasting (Nevo et al., 2022; Nearing et al., 2024). This paper demonstrates a synthesis of approaches to help mitigate data-related issues and increase the reliability of streamflow forecasts. We specifically evaluate the technical implementation of version 2 (v2) of the Google Global Flood Forecasting system. This system is run operationally and powers the river forecasting component of Google's FloodHub (g.co/floodhub).

Upgrades between v1 (Nevo et al., 2022) and v2 systems primarily focus on improvements around three data-related challenges related to training data availability, temporally limited data records, and Input data distribution shifts.

This paper serves two primary purposes: (1) to provide transparency to the scientific community about operational flood forecasting system progress and challenges, and (2) to facilitate research around ML-based flood forecasting by providing open source resources. The GoogleHydrology open-source code repository (https://github.com/google-research/flood-forecasting) implements a version of the model architectures and training pipeline evaluated in this paper, and we additionally released a multi-decade historical data archive, called the Google Runoff Reanalysis and Reforecast (GRRR) dataset, which contains historical simulations and historical reforecasts at over 1 million locations globally.

2 Model Description

2.1 Data

The river forecast model is trained to predict daily average streamflow at a basin outlet. It uses two types of input data: static catchment attributes, which describe physical characteristics that are implicitly assumed to be time-invariant (e.g., topography, soil type, land cover), and dynamic meteorological forcings, which represent the time-varying weather conditions (e.g., precipitation, temperature) driving the hydrological response. Sources for these three types of data are described in the following subsections.

2.1.1 Static Catchment Attributes

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

We use a set of 92 static catchment attributes to characterize the time-invariant physical properties of each basin. These attributes are spatially averaged over catchment area and derived from the HydroATLAS dataset (Linke et al., 2019) plus hydro-climatic statistics calculated from ERA5-Land reanalysis (Muñoz-Sabater et al., 2021). A detailed description of all attributes is provided in **Table 1**. The Caravan GitHub repository (https://github.com/kratzert/Caravan) contains the code to calculate all ERA5 and climate attributes.

At a conceptual level, static attributes provide information about:

- **Topography:** Mean elevation, slope, and relief characteristics.
- **Climate Indices:** Aridity, seasonality, and precipitation frequency/duration statistics.
- **Land Cover:** Fractions of forest, urban areas, water bodies, snow cover, and various vegetation classes.
- **Soil Properties:** Soil texture (sand, silt, clay content), organic carbon, and hydraulic properties.
- **Anthropogenic Factors:** Human footprint index, GDP, population density, and road density.

Table 1: *Full list of static catchment attributes incorporated into the model as input features.*

Attribute Name	Description
HydroATLAS Attributes	
aet_mm_uyr	Actual evapotranspiration; multi-year annual average (mm/yr).
ari_ix_uav	Aridity index; multi-year average.
crp_pc_use	Cropland extent; percentage of sub-basin area.
ele_mt_uav	Elevation; multi-year average in meters above sea level.
ero_kh_uav	Soil erosion; multi-year average in kg/ha/yr.
for_pc_use	Forest cover; percentage of sub-basin area.
gdp_ud_usu	Gross Domestic Product (GDP); total sum for sub-basin (USD).

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

gla_pc_use	Glacier cover; percentage of sub-basin area.
glc_pc_u01 - glc_pc_u22	Land cover classes (e.g., forests, grasslands, urban, water bodies); percentage of sub-basin area. See full list for specific classes.
hft_ix_u09	Human Footprint index for year 2009 (multi-year average).
hft_ix_u93	Human Footprint index for year 1993 (multi-year average).
inu_pc_ult	Inundation extent (long-term); percentage of sub-basin area.
inu_pc_umn	Inundation extent (minimum); percentage of sub-basin area.
inu_pc_umx	Inundation extent (maximum); percentage of sub-basin area.
ire_pc_use	Irrigated area; percentage of sub-basin area.
kar_pc_use	Karst area; percentage of sub-basin area.
lka_pc_use	Lake area; percentage of sub-basin area.
lkv_mc_usu	Lake volume; total sum in sub-basin (million m^3).
nli_ix_uav	Nighttime lights index; multi-year average.
pac_pc_use	Protected area cover; percentage of sub-basin area.
pet_mm_uyr	Potential evapotranspiration; multi-year annual average (mm/yr).
pnv_pc_u01 - pnv_pc_u15	Potential Natural Vegetation classes; percentage of sub-basin area.
pop_ct_usu	Population count; total sum within sub-basin.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

ppd_pk_uav	Population density; multi-year average (people/km^2).
pre_mm_uyr	Precipitation; multi-year annual average (mm/yr).
prm_pc_use	Permafrost extent; percentage of sub-basin area.
pst_pc_use	Pasture extent; percentage of sub-basin area.
rdd_mk_uav	Road density; multi-year average (m/km^2).
rev_mc_usu	Reservoir volume; total sum in sub-basin (million m^3).
ria_ha_usu	Reservoir area; total sum in sub-basin (hectares).
riv_tc_usu	River volume; total sum in sub-basin (thousand m^3).
snw_pc_uyr	Snow cover; percentage multi-year annual average.
swc_pc_uyr	Soil water content; percentage multi-year annual average.
tmp_dc_uyr	Air temperature; multi-year annual average (degrees Celsius).
urb_pc_use	Urban area; percentage of sub-basin area.
wet_pc_u01 - wet_pc_u09	Wetland classes (e.g., lakes, reservoirs, rivers, marshes); percentage of sub-basin area.
wet_pc_ug1, wet_pc_ug2	Global wetland classes (Lakes/Reservoirs, Floodplains/Marshes).
ERA5-Land Attributes	
aridity	Aridity index derived from ERA5-Land reanalysis.
frac_snow	Fraction of precipitation falling as snow.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

high_prec_dur	Average duration of high precipitation events.
high_prec_freq	Frequency of high precipitation events.
low_prec_dur	Average duration of low precipitation events.
low_prec_freq	Frequency of low precipitation events.
moisture_index	Moisture index derived from ERA5-Land.
p_mean	Long-term mean precipitation.
pet_mean	Long-term mean potential evapotranspiration from ERA5-Land.
seasonality	Seasonality index of precipitation from ERA5-Land.
Derived Attributes	
area	Calculated drainage area.

[1]*Climate variables derived from FAO evapotranspiration used the guidelines outlined by Allen et al., (1998).*

2.1.2 Dynamic Meteorological Inputs

Kratzert et al. (2021) found that using multiple meteorological data sources in a single LSTM model provides performance improvements. Accordingly, our model ingests daily time-series data from multiple global weather data sources. Historical records of our dynamic meteorological forcings are available from the open source Caravan MultiMet extension (Shalev et al., 2024) in sufficient quantity to train global streamflow models with the GoogleHydrology codebase.

The dynamic inputs are:

1. **ECMWF HRES (European Centre for Medium-Range Weather Forecasts):**
    - High-Resolution operational deterministic forecasts:
        - Total precipitation
        - 2-meter temperature
        - Surface net solar radiation
        - Surface net thermal radiation

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

- Surface pressure
- Snow fraction

2. **CPC (NOAA Climate Prediction Center):** Global Unified Gauge-Based Analysis of Daily Precipitation (Chen et al., 2008).
   - Precipitation
3. **GraphCast (Google DeepMind):** AI-based medium-range global weather forecasting mode (Lam et al., 2023)l. Variables include:
   - Total precipitation
   - 2-meter temperature
4. **IMERG (NASA):** Integrated Multi-satellite Retrievals for GPM Early Run (Huffman et al., 2020).
   - Precipitation

All meteorological data are aggregated from their native temporal resolutions to daily. The method of aggregation depends on the physical units of each data variable (rates vs. quantities).

The model uses these inputs in two distinct modes: hindcast and forecast. The hindcast period represents observed historical weather up to the issue time of a forecast and is used to spin up the model's states. For this phase, we use ECMWF HRES and GraphCast weather forecast data products in hindcast mode by extracting the 0-day lead time forecast for each issue time. Only ECMWF HRES and GraphCast products are available as true forecasts, meaning that CPC and IMERG are not used in the forecast phase of the model. The distinction between the hindcast and forecast phases of our model is analogous to the difference between Quantitative Precipitation Estimates (QPE) and Quantitative Precipitation Forecasts (QPF) in operational hydrology (Liu et al., 2012).

As illustrated in **Figure 1**, meteorological input products have varying availability timelines for training. While streamflow labels and reanalysis proxies (ERA5-Land) are available for several decades (we make an arbitrary decision to limit data records to the period 1980 - 2024), operational forecast archives for HRES and GraphCast begin later (approx. 2012 and 2016, respectively).

To take advantage of the full historical streamflow record (1980–2024) for training, we employ a data imputation strategy called feature unioning. For time periods where operational forecasts are unavailable from either HRES or Graphcast, we substitute these with ERA5-Land reanalysis data. Because HRES shares the same underlying physical model as ERA5, and GraphCast is trained on ERA5, the assumption is that statistical distributions of the reanalysis

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

data serve as an effective proxy for the forecast inputs during the earlier training years.

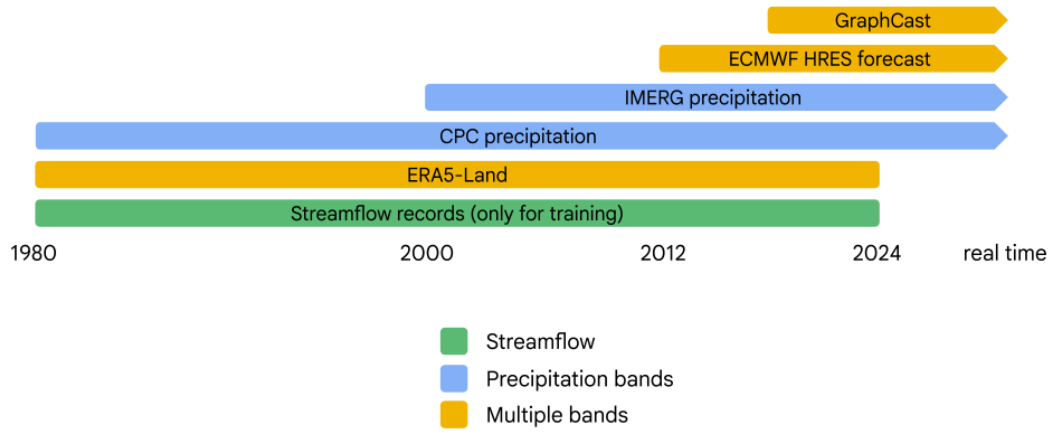


Figure 1: *Temporal availability of the dynamic input datasets and target streamflow data.*

2.1.3 Streamflow Target Data

The target variable is daily streamflow, obtained for training the v2 system from the Caravan dataset (Kratzert et al., 2023), Global Runoff Data Center (GRDC; Färber et al., 2024), and BANDAS (Solís-Alvarado et al., 2015). The v1 benchmark was trained on data from the GRDC. All GRDC test gauges from the v1 evaluation by Nearing et al. (2024) are present in the v2 and third-party benchmark test sets, ensuring a direct comparison. The global spatial distribution of these expanded training and testing gauges is illustrated in **Figure 2**.

It is important to acknowledge that Caravan is an open-community dataset based on data from regional large-sample hydrology datasets that are often referred to as Catchment Attributes and Meteorology for Large Sample studies (CAMELS) datasets (Newman et al., 2015; Addor et al., 2017). Caravan relies fundamentally on contributions like those from Denmark (Liu et al., 2024), Chile (Alvarez-Garreton et al., 2018), Iceland (Helgason et al., 2024), Israel (Morin, 2023), Switzerland (Höge et al., 2023), Spain (Casado Rodríguez, 2023), and the GRDC (Färber et al., 2024).

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

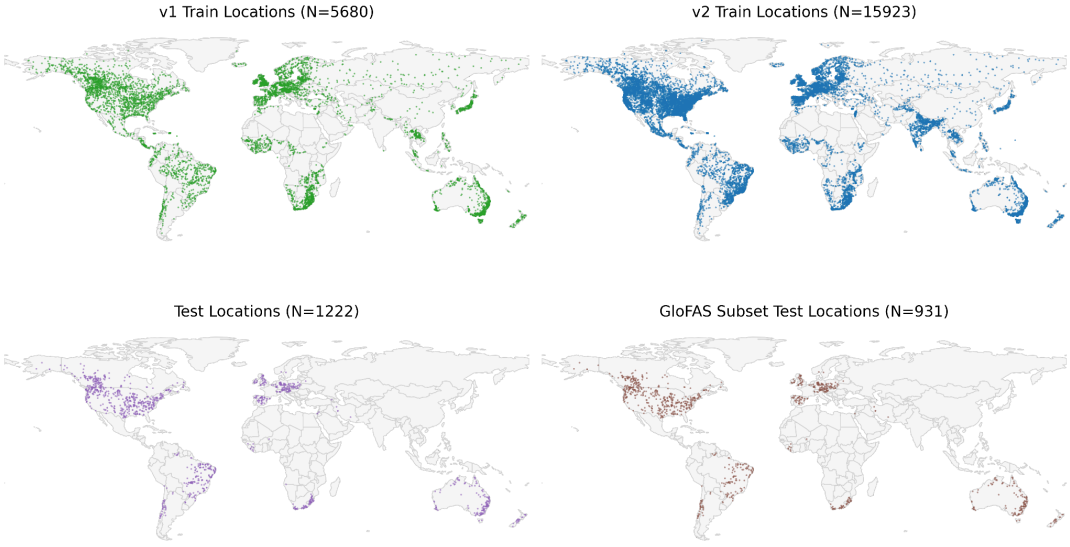


Figure 2: *Spatial distribution of training and testing gauges globally. The top row illustrates the expansion of training data locations from the v1 dataset to the expanded v2 dataset due to Caravan. The bottom row maps the specific holdout test set (left pane), and the fraction of those test locations where we have GloFAS data.*

2.2 Model Architectures

The model used in our v1 system is an encoder-decoder (ED-LSTM) architecture described by Nevo et al. (2022) and evaluated by Nearing et al. (2024). The model used in the v2 system is a Mean Embedding (ME-LSTM) architecture developed by Gauch et al. (2025). The ME-LSTM is more robust to missing data and addresses a forecast initialization challenge described in the following subsection.

2.2.1 Encoder-Decoder LSTM (ED-LSTM)

The ED-LSTM runs separate LSTMs for the hindcast vs. forecast periods. The hidden and cell states of the hindcast LSTM are used to initialize the forecast LSTM through a small neural network. This allows for different weights in each of the LSTM gates for hindcast vs. forecast meteorological inputs, and addresses the challenge of distribution differences between observation-based or analysis-based features vs. numerical weather predictions. This model is described in detail in prior publications (Nevo et al., 2022; Nearing et al., 2024).

One motivation for the development of a new model architecture for the v2 system was a specific failure mode observed in the ED-LSTM (Nearing et al., 2023a). Empirically, the handoff between the states of the hindcast and forecast LSTMs can cause a discontinuity during forecast initialization. A toy example illustrating this forecast artifact is shown in the left panel of **Figure 3**. The cause of this artifact is that the forecast LSTM is sometimes initialized with a cell

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

state that is outside of a steady-state distribution and the model spends the first few time steps of the forecast horizon annealing its internal states to the new input climatology rather than responding strictly to hydrological drivers. This is illustrated in the right panel of **Figure 3**.

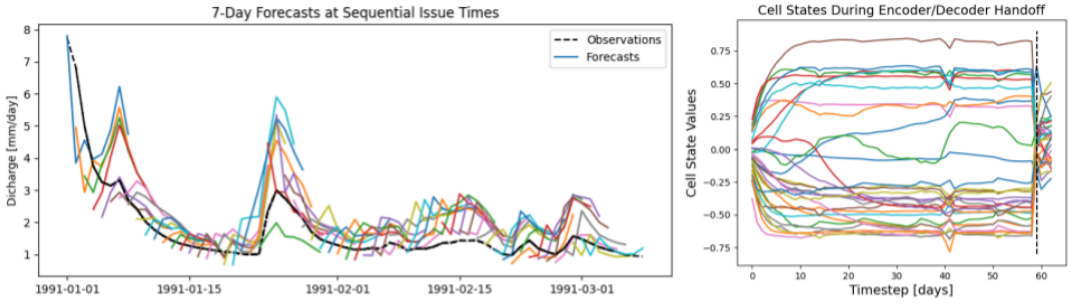


Figure 3. *Forecast initialization artifacts and their underlying cause. The left panel shows an example of an artifact in ED-LSTM forecasts and the right panel shows the cause of this artifact. This is a toy example for illustration purposes that uses a short data record from a randomly selected basin.*

2.2.2 Mean Embedding Forecast LSTM (ME-LSTM)

The ME-LSTM addresses limitations of the ED-LSTM by using a stacked LSTM framework to handle input distribution shifts and embedding layers to handle missing inputs. The model architecture is described in detail by Gauch et al. (2025) and illustrated in **Figure 4**.

Instead of concatenating all meteorological inputs into a single vector at each timestep, which fails if any input is missing, the ME-LSTM treats different meteorological products (e.g., HRES, GraphCast, IMERG) as independent input groups. Each product group is processed by a dedicated embedding network that projects inputs into a shared latent space. Embedding networks are multilayer perceptrons. Static catchment attributes are concatenated to dynamic inputs of each product group before embedding. These embeddings are aggregated using a masked mean operation, which computes the average of the embeddings at each time step, ignoring any product that contains missing data.

The ME-LSTM then employs two LSTM layers arranged in a stack that run over the full time sequence (hindcast + forecast), eliminating the handoff discontinuity. The first layer processes a sequence of aggregated embeddings derived from hindcast data and runs over the full sequence, i.e., hindcast (spinup) + forecast sequence length. The second LSTM layer predicts streamflow. It receives two inputs at every time step: the aggregated forecast embeddings and the hidden state output from the first LSTM layer.

Both v1 and v2 models evaluated in this paper use a mixture density layer to make probabilistic predictions. At each location and timestep, each model predicts the parameters of a countable mixture of asymmetric Laplacians (CMAL) distribution (Klotz et al., 2022). We take the mean of

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

the predicted distribution to be the deterministic model prediction that we evaluate in this paper.

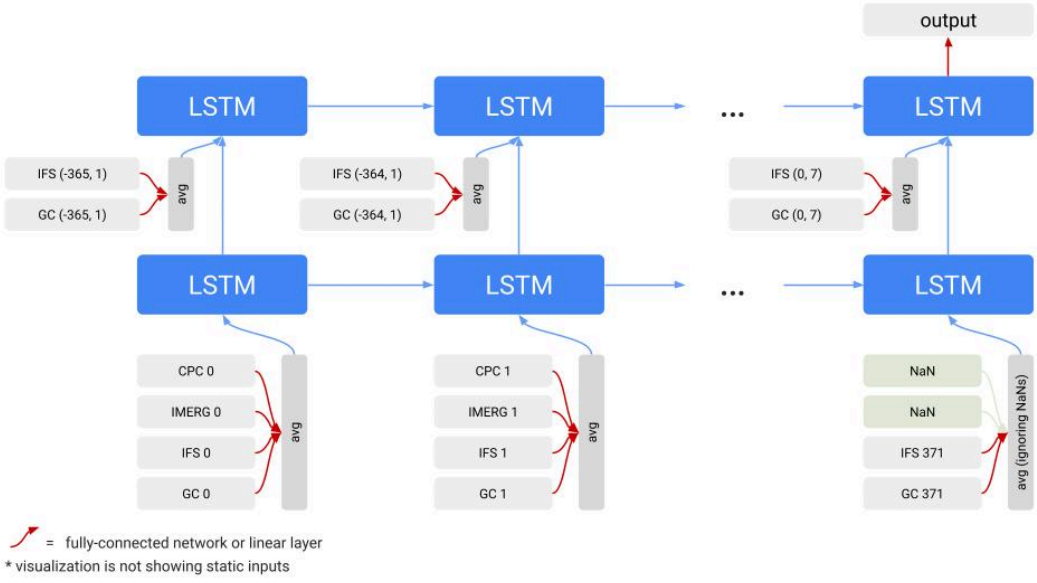


Figure 4: *Architecture of the Mean Embedding Forecast LSTM (ME-LSTM).*

2.2.3 Hyperparameters

Model hyperparameters for both v1 and v2 models are provided in GoogleHydrology configuration files released with this paper. Hyperparameters for the v2 system model were chosen by approximate hyperparameter tuning using a procedure analogous to what was described in Appendix A by Kratzert et al. (2024). We do not report this hypertuning experiment in this paper, however users are free to do their own hyperparameter tuning with the open source GoogleHydrology codebase provided. The most important hyperparameters for the v2 model are detailed in **Table 2**.

Table 2: *Main Hyperparameters of the ME-LSTM Model Architecture.*

Component	Specification
Hindcast LSTM Hidden Size	512
Forecast LSTM Hidden Size	512
Sequence Length (T_{seq})	365 days
Forecast Horizon (T_{lead})	7 days

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Static Embedding Net	Network: Fully-Connected Depth: 3-Layer Neurons: 100, 100, 20 Activations: tanh, tanh, linear
Hindcast Embedding Nets	Network: Fully-Connected Depth: 2-Layer Neurons: 100, 20 Activations: tanh, linear
Forecast Embedding Nets	Network: Fully-Connected Depth: 4-Layer Neurons: 20, 20, 20, 20 Activations: tanh, tanh, tanh, linear

2.3 Training

The v2 system model was trained using an Adam optimizer with a batch size of 512 with the following:

- **Loss Function:** We used the CMAL likelihood loss described by Klotz et al. (2022).
- **Noise Injection:** We injected Gaussian noise with a standard deviation of 0.005 into normalized target variables during training.
- **Learning Rate:** We used 0.0005.
- **Gradient Clipping:** Gradients were clipped to a norm of 1.0.
- **Input feature dropout:** We randomly dropped timeseries input features with probability 0.1 to increase the robustness of our model against missing data, as described in Gauch et al. (2025).

We used 125 epochs with a batch limit of 2000 updates per epoch.

3 Model Evaluation

Our evaluation assesses forecast skill in both gauged and ungauged settings. Gauged prediction means that streamflow observations from the evaluation gauges are used in training (although out-of-sample in time), whereas ungauged prediction assumes no local streamflow data is used for training. For the ungauged setting, the v1 system used random k-fold (k=10) cross-validation, whereas the v2 system used a single holdout test set.

3.1 Confounding Factors in Operational System

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Upgrades

A key limitation in evaluating operational system upgrades is the simultaneous evolution of multiple factors like data and architecture. Consequently, the performance deltas between the v1 and v2 systems are a compound effect from the architectural shift (ME-LSTM) and the expanded geographic training sample from Caravan. We additionally evaluate an ablated version of the v2 system without GraphCast to disaggregate performance gains from the new input meteorological forcing data.

3.2 Evaluation Metrics

Gauch et al. (2023) outlined a set of evaluation metrics for hydrograph prediction that can explain expert-driven model intercomparison. In the interest of tractability, here we used only a subset of those metrics that allow us to highlight particular characteristics of model performance. Since all of the model runs produced for this study are publicly available, readers are welcome to explore these results using different metrics.

In particular, we evaluated performance here using the Nash-Sutcliffe Efficiency (NSE; Nash & Sutcliffe, 1970) and a decomposition of the Kling-Gupta Efficiency (KGE; Gupta et al., 2009):

- **Nash-Sutcliffe Efficiency (NSE):** This is a widely used normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance. An NSE of 1.0 indicates perfect correspondence between predictions and observations. An NSE of 0 indicates that the model predictions are as accurate as the mean of the observed data, and an NSE less than zero indicates that the observed mean is a better predictor than the model.
- **Kling-Gupta Efficiency (KGE) Decomposition:** The KGE metric incorporates correlation, variability bias, and mean bias into a single metric. As with the NSE, a KGE metric of 1.0 indicates a perfect model, and a KGE of 0 indicates that the model predictions are as accurate as the observation mean. We look at these three constituent components independently:
    - **Correlation (r):** The Pearson correlation coefficient evaluates the phase timing and temporal dynamics of the hydrograph.
    - **Bias Ratio (β):** This ratio quantifies the long-term volumetric water balance error (mass conservation). This is calculated as the mean of the simulated predictions divided by the mean of the observed values.
    - **Variability Ratio (γ):** This ratio measures the model's ability to capture the amplitude of flow extremes. It is calculated as the coefficient of variation of the simulated predictions divided by the coefficient of variation of the observed

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

values.

3.3 Time Periods & Cross Validation

Our benchmarking experiments assess model performance in both gauged and ungauged settings. When we say that a model is gauged or makes gauged predictions, we mean that the model is used to make predictions in the same locations where it received training data. When we say that a model is ungauged or makes ungauged predictions, we mean that the test locations are withheld from training. In both cases (gauged and ungauged), we use a temporal cross-validation split.

The test period for all models is January 1, 2016 through December 31, 2023. Both v1 and v2 models were trained with cross validation over individual calendar years in that time period. Since both models use a hindcast or spin-up sequence length of 365 days, we withheld one year of training data before and after each hold-out year for each temporal cross-validation split to prevent data leakage. As an example, for testing on the year 2020, we held out from training all data during the period January 1, 2019 through December 31, 2021, and added predictions from the period January 1, 2020 through December 31, 2020 to the timeseries or test predictions that all metrics were calculated over. By repeating this strategy for each test year, we constructed out-of-sample predictions for the entire time period of 2016 - 2023. The training period for all models is 1980 - 2023, minus the appropriate 2 or 3 year holdout for each cross validation split.

Because we use existing model runs for all of the benchmarks, described in the next section, we constructed sets of test basins by choosing locations that exist in the v1 model runs by Nearing et al (2024), in the locations used for our v2 operational system, and in the GEOGLOWS dataset. This intersection resulted in 1,222 shared test basins, of which 931 also exist in the GloFAS test set. The bottom row of **Figure 2** shows these test locations. For the GloFAS benchmark, geolocation matching between streamflow gauges and the GloFAS river network was performed by ECMWF as part of the Nearing et al. (2024) publication.

The v2 models were trained using a single spatial split (with and without GraphCast) while the v1 models were trained using random k-fold (k=10) cross-validation as described by Nearing et al. (2024).

3.4 Benchmarks

To contextualize the performance of the v2 system, we compare our results against three reference systems: the v1 operational system, GloFAS, and GEOGLOWS.

The first benchmark is the version 1 (v1) operational system of the Google Global Flood Forecasting system. This system was described by Nevo et al. (2022) and evaluated by Nearing et al. (2024). The forecast model in this system is the ED-LSTM model described in **Section**

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

2.2.1. For this evaluation we used the open data release from Nearing et al. (2024).

The second benchmark is the Global Flood Awareness System (GloFAS), which serves as the global flood forecasting system of the Copernicus Emergency Management Service (CEMS). GloFAS couples land surface modeling with global routing networks to produce probabilistic river discharge forecasts (Alfieri et al., 2013; Harrigan et al., 2023). For our evaluations, we use the current operational version, GloFAS version 4. Readers are referred to the corresponding literature and CEMS documentation for detailed descriptions of the model's structure and calibration.

The third benchmark is the GEOGLOWS ECMWF Streamflow Routing model. GEOGLOWS provides global-scale hydrologic forecasting by mapping ECMWF runoff predictions to a high-resolution global river network using the Routing Application for Parallel computation of Discharge (RAPID) framework (Souffront et al., 2019). For details regarding the GEOGLOWS configuration and routing methodology, we refer readers to the associated academic literature and the model's online documentation.

Geolocation matching between streamflow gauges and the GloFAS river network was performed by the GloFAS team as part of the Nearing et al. (2024) publication. To align gauges with GEOGLOWS prediction points, we matched catchment polygons to the lumped upstream area polygons of TDX-Hydro subbasins, the hydrofabric used by GEOGLOWS. For each polygon pair, we calculated the Intersection over Union (IoU) of their areas and selected the match with the highest spatial agreement, enforcing a minimum IoU threshold of 0.8. The bottom row of **Figure 2** shows the locations of these test gauges used for evaluation and benchmarking.

4 Results

4.1 Global Performance Distributions

Figure 5 shows the NSE distributions of all models at all lead times in both gauged and ungauged settings. These figures provide a high-level overview and are difficult to interpret, so we will dissect these results in more detail in following subsections. The main takeaway from **Figure 5** is that v2 represents an improvement in accuracy compared with v1, and both systems perform significantly better than the traditional benchmarks of GloFAS and GeoGlows.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

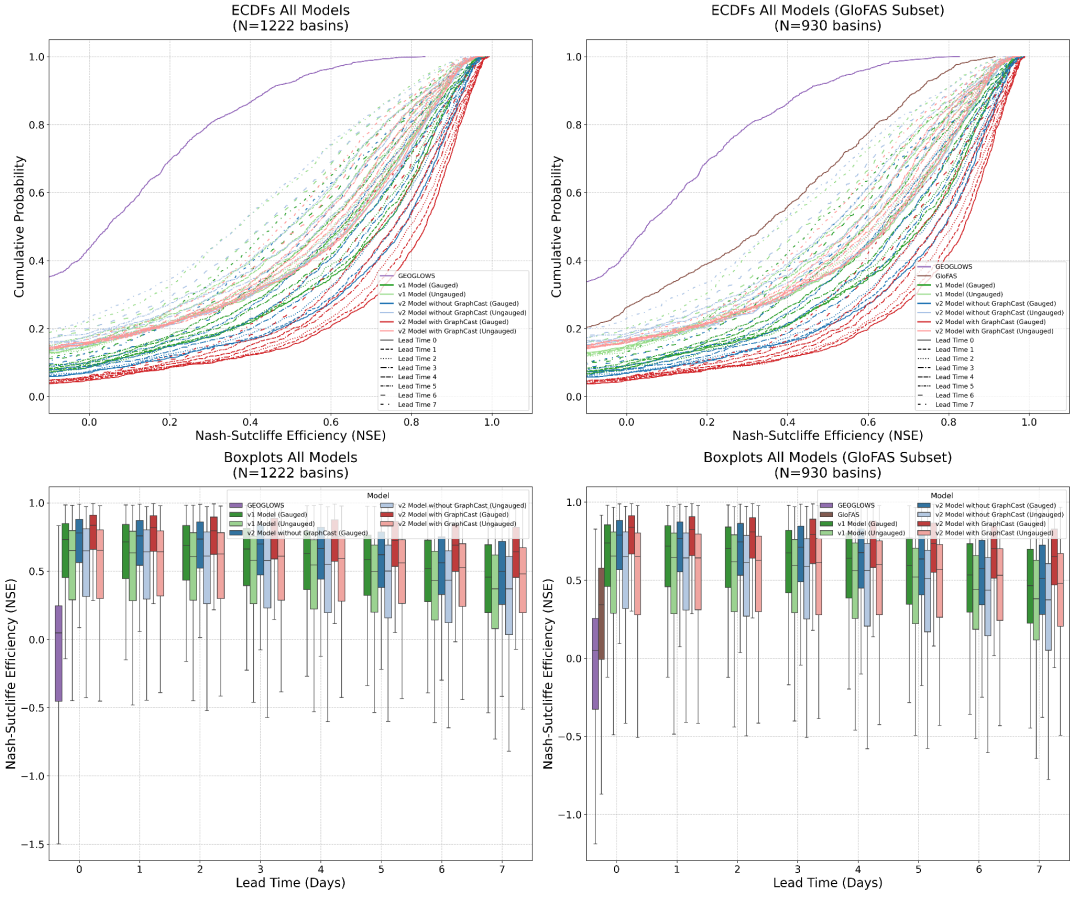


Figure 5: *Global performance distributions of Nash-Sutcliffe Efficiency (NSE) across evaluated models and lead times. The top row shows Empirical Cumulative Distribution Functions (ECDFs) for NSE across all lead times. The bottom row shows the same data as boxplots. The left column shows evaluation results for the full test set, and the right column shows evaluation results for the portion of the test set with GloFAS data.*

4.2 Disaggregating Improvements

Figure 6 disaggregates the improvements provided by the ME-LSTM architecture and expanded training data from the predictive skill injected by the GraphCast meteorological forcings. The core v1-to-v2 upgrade (blue boxes) provides improvement in NSE across all lead times in the gauged setting, but not in the ungauged setting. The addition of GraphCast (green boxes) has a larger overall effect that is smaller at short lead times but grows with a longer forecast horizon. This effect exists in both the gauged and ungauged setting, but is larger for gauged predictions.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

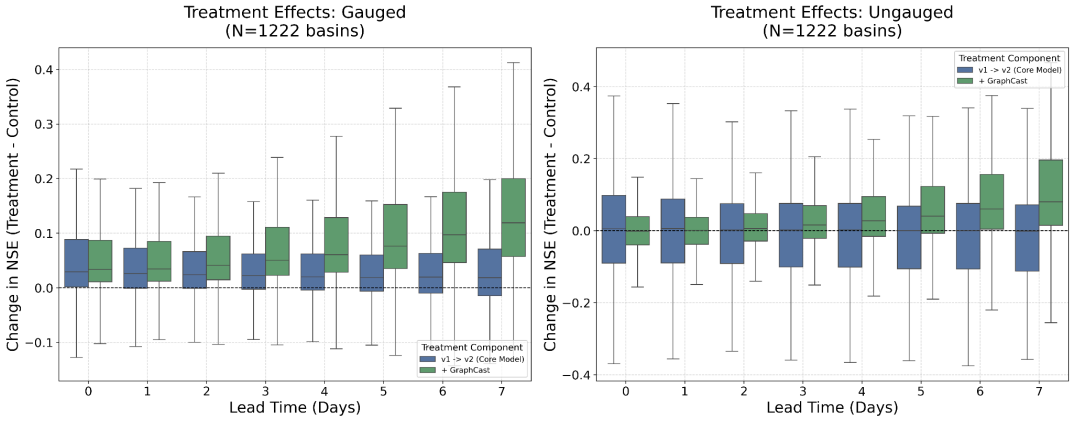


Figure 6: *Paired treatment effects across lead times. Blue boxes represent the ΔNSE gained by transitioning from the v1 to v2 model architecture and expanded Caravan training data. Green boxes represent the additional ΔNSE gained by incorporating GraphCast.*

It is instructive to decompose the KGE metric into three constituent parts as described in **Section 3.2**. **Figure 7** shows this KGE decomposition across lead times. The primary effect is that GraphCast forcings improve correlation but lower forecast variance. This is conceptually consistent with the fact that AI weather forecasts have higher accuracy than forecasts from traditional models, but that they are typically optimized for mean squared error, which inherently leads to spatial and temporal smoothing and an underprediction of variance at longer lead times (Ben Bouallègue et al., 2024).

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

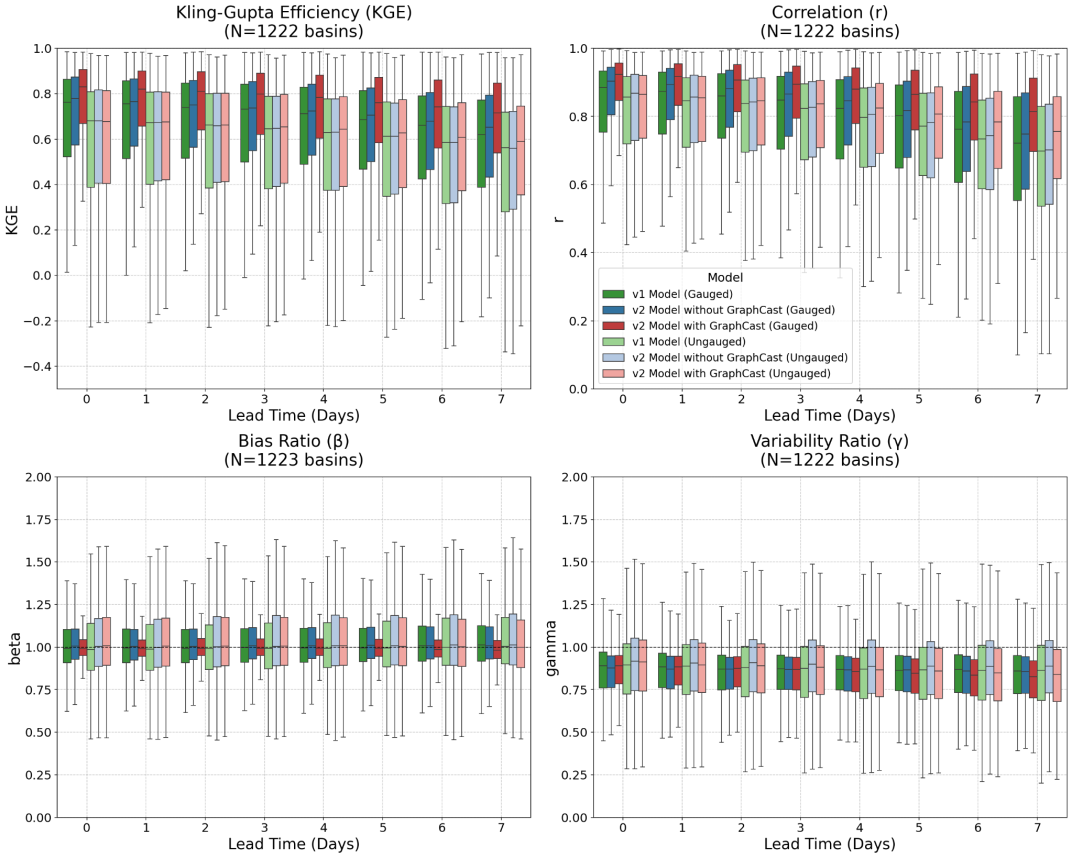


Figure 7: *Component Decomposition of KGE (Correlation, Bias Ratio, Variability Ratio) across lead times.*

4.3 Lead Time Extension

The previous analyses, and especially **Figure 5**, demonstrate that the v2 system offers an extended predictive horizon relative to v1. We used a one-sided Wilcoxon signed-rank test to measure how much extra lead time is gained at a statistically significant level (α=0.05). **Figure 8** visualizes this test. In the gauged setting, the v2 system demonstrates statistically equivalent or superior predictive skill up to a 6-day forecast horizon relative to the v1 benchmark at a 0-day (nowcast) forecast horizon. In the ungauged setting, this statistical equivalence holds out to 1 day of extra lead time.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

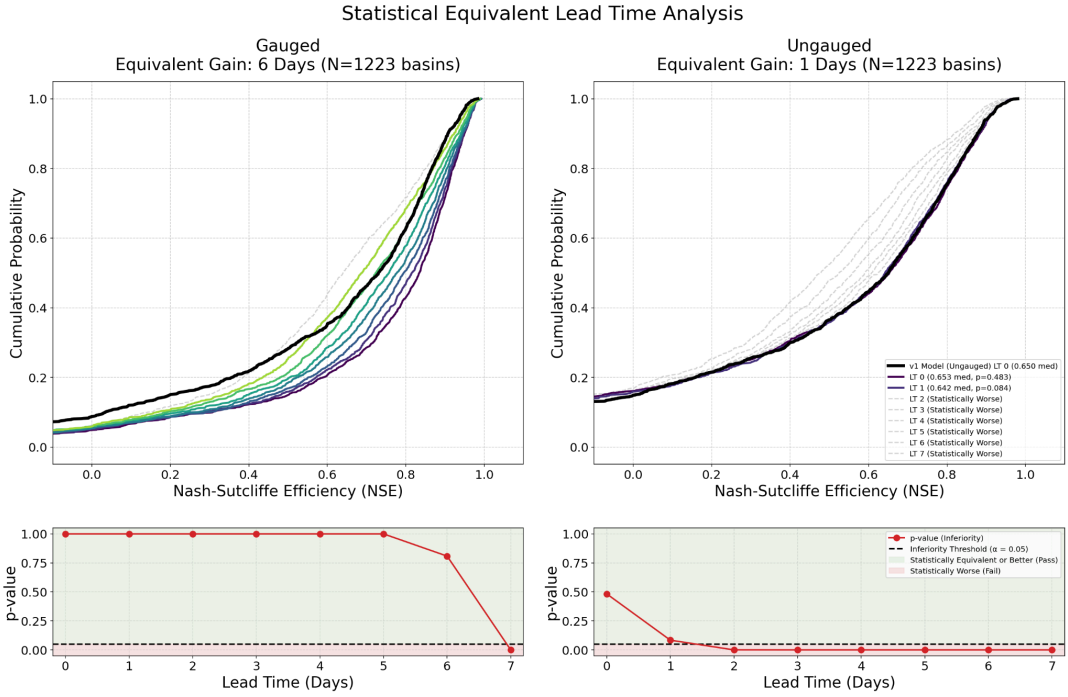


Figure 8: *Statistical equivalent lead time evaluation for the v2 system forecasts (lead times 0 through 7) relative to the v1 benchmark nowcast (lead time 0). The bottom row displays p-values derived from a one-sided Wilcoxon signed-rank test comparing the NSE of the v2 model (with GraphCast) at extending lead times against the v1 model at Day-0. The p-value (red line) being within the green area ($p > \alpha$) indicates lead times where the v2 forecast remains statistically equivalent to, or better than, the v1 nowcast. The horizontal dashed line represents the significance threshold* (α=0.05). *The v2 system achieves up to 6 days of statistically non-different lead time in the gauged setting (left panel) and 1 day in the ungauged setting (right panel).*

4.4 Effect of Hydrological Characteristics

We computed Pearson correlation coefficients between static catchment attributes and both absolute predictive skill (NSE) and the change in skill (ΔNSE). **Figure 9** displays the ten attributes with the highest correlations (for the v2 model) in gauged and ungauged basins.

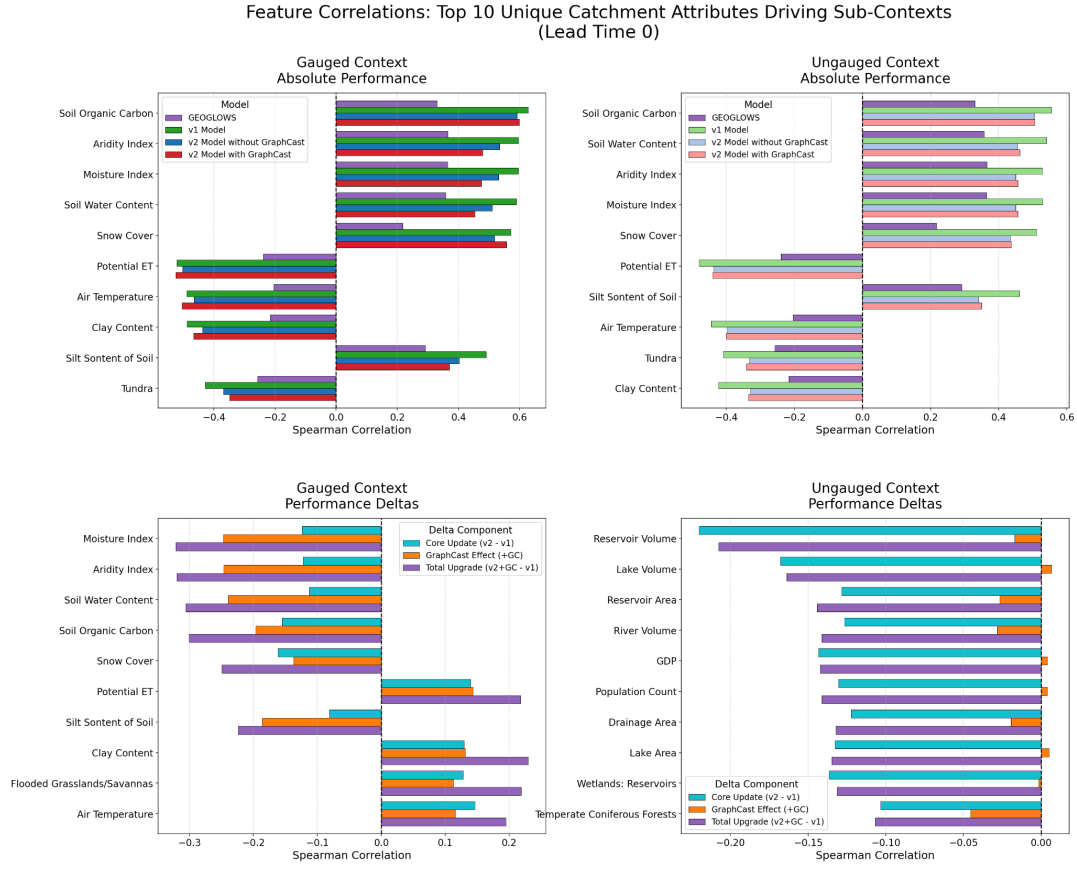
Absolute predictive skill correlates positively with soil organic carbon, aridity index, and moisture indexes, and snow cover fractions across all models. It is important to note that the particular aridity index we are using (from HydroSheds) is lower in dryer basins and higher in wetter basins. Forecasts are most accurate in wet catchments with more snow and more vegetation and less accurate in arid catchments, due to the fact that streamflow in arid basins is flashy and determined mostly by the accuracy of precipitation data. Vegetation, soil water storage, and snow act as low-pass filters on the rainfall-runoff relationship. Prior work

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

demonstrates how LSTM-based models are able to learn seasonal snow accumulation and melt dynamics from precipitation and streamflow data (Kratzert et al., 2019c).

In gauged catchments, skill differences (ΔNSE) between model versions correlates negatively with humidity indexes, meaning that *improvements* are larger in more arid locations. In ungauged catchments, ΔNSE correlates negatively with lake and reservoir area fractions. This indicates less improvement in managed or regulated catchments when direct streamflow observations are unavailable. Notice that this effect in ungauged basins is not related to the introduction of GraphCast forcings, but due to the model itself.

Figure 10 shows the absolute (not delta) performance of each model stratified along deciles of the 9 catchment attributes with the highest correlation with both gauged and ungauged performance. While trends are generally the same between v1 and v2 models in both gauged and ungauged settings, there are certain attribute deciles where the v2 model is not always better, however these differences are small. The attribute-to-performance trends also exist for GEOGLOWS, but the correlations are weaker (also, see **Figure 9**), however the GEOGLOWS scores are lower overall.



https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Figure 9: *Pearson correlation coefficients between static catchment attributes and model performance. The top panels show correlations with absolute NSE in gauged (left) and ungauged (right) basins. The bottom panels show correlations with the ΔNSE between system versions.*

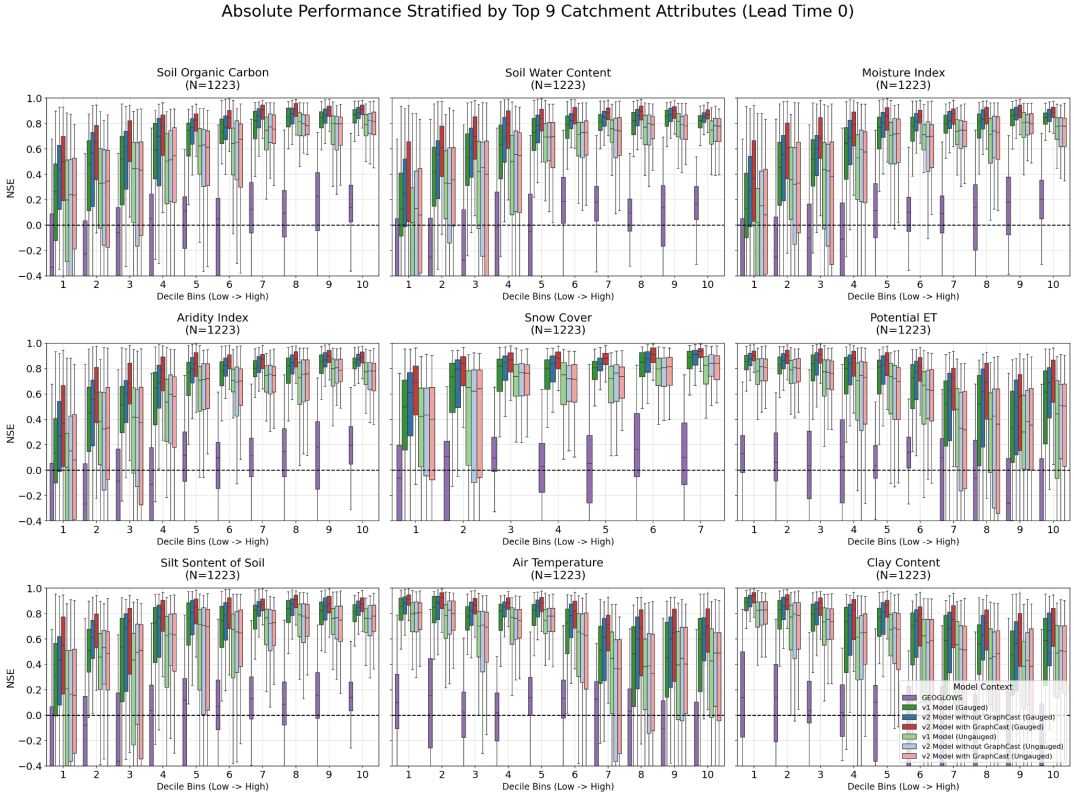


Figure 10: *Absolute model performance stratified by key catchment attributes. Distribution of Nash-Sutcliffe Efficiency (NSE) at 0-day lead time for the benchmark (GEOGLOWS), baseline (v1), and updated (v2) model configurations across gauged and ungauged contexts. Basins are divided into ten equal-sized bins (deciles) ranging from lowest (1) to highest (10) values for the top nine catchment attributes most strongly rank-correlated with performance (**Figure 9**). Boxplots indicate the median and interquartile range of NSE within each decile, illustrating how specific physical and climatological characteristics impact the predictive skill of the varying modeling frameworks.*

4.5 Gauged vs. Ungauged Performance

Figure 11 shows skill differences (*ΔNSE*) between gauged vs. ungauged basins in the v1 system, the v2 system without GraphCast, and the v2 system with GraphCast. The bottom panel of **Figure 11** shows median skill differences and relative percentage NSE penalties across

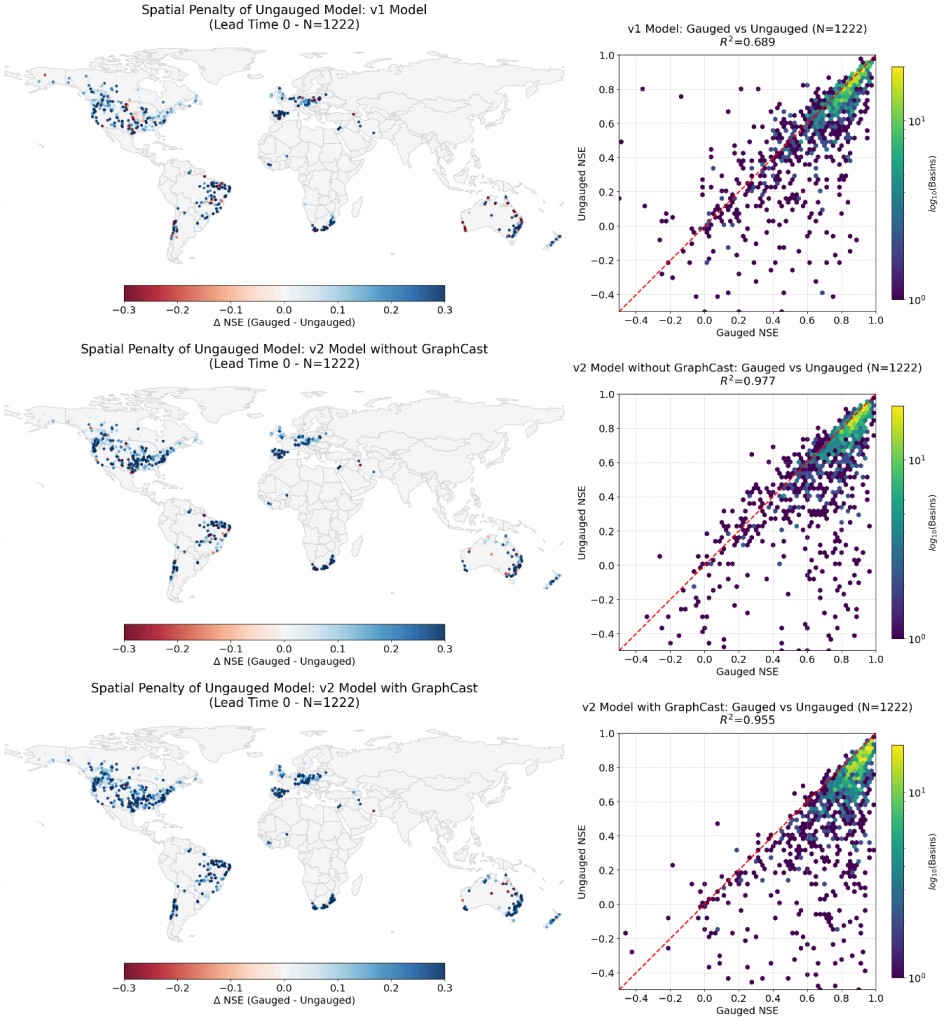
https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

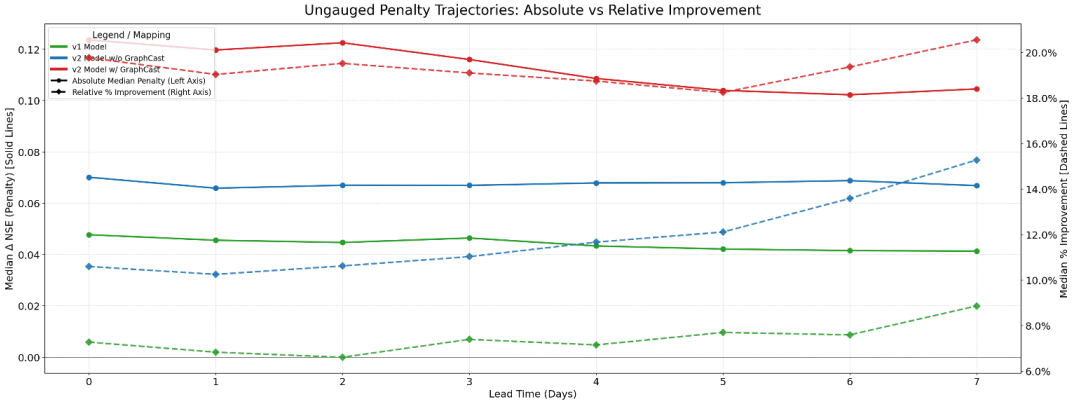
lead times.

At a 0-day lead time, the v2 system without GraphCast has a median gauged NSE of 0.78, with an absolute median penalty of 0.07 (10.6% relative decrease). The v2 system with GraphCast achieves a median gauged NSE of 0.83, with an absolute median penalty of 0.12 (19.8% relative decrease).

The absolute median penalty remains relatively constant over the forecast horizon. At a 7-day lead time, the v2 system without GraphCast yields a median gauged NSE of 0.4989 with an absolute penalty of 0.07 (15.3% relative decrease). The v2 system with GraphCast yields a median gauged NSE of 0.64 with an absolute penalty of 0.10 (20.6% relative decrease). Although the largest skill difference between gauged and ungauged settings in both an absolute and relative sense is between the v2 systems with and without GraphCast at longer lead times, it is important to recognize that the v2 system with GraphCast is the most accurate overall (according to this median NSE metric). The ungauged median NSE at a 7-day lead time is 0.48 for v2 with GraphCast and 0.37 for both v1 and v2 without Graphcast. The second row of **Figure 5** shows the median scores for all models at all lead times. Results indicate that while the v2 system yields higher absolute performance globally, it is proportionally more sensitive to the absence of local streamflow data for training.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026





https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Figure 11: *Gauged versus ungauged performance penalty measured as ΔNSE. The top three panels display the spatial distribution and scatter of gauged versus ungauged NSE scores for the v1 system, the v2 system without GraphCast, and the full v2 system, respectively. The bottom panel illustrates the absolute median and relative percentage NSE penalties across the forecast horizon.*

5 Conclusions

5.1 Summary of System Upgrades

Version 2 of the Google Global Flood Forecasting system replaces the v1 encoder-decoder LSTM with a continuous Mean Embedding (ME-LSTM) architecture. The v2 system incorporates an expanded training dataset from the Caravan community dataset and integrates GraphCast AI-based meteorological forcings alongside traditional numerical weather prediction products. These architectural changes resolve certain forecast initialization artifacts observed in the v1 system (**Section 2.2.1**).

The v2 system improves aggregate predictive skill as compared to the v1 system and two third-party benchmarks (GloFAS and GEOGLOWS). The main takeaways from our analysis are as follows:

- KGE decomposition shows these improvements are primarily driven by enhanced temporal correlation rather than reductions in volumetric bias or variance. GraphCast inputs improve correlation at longer lead times, but reduce forecast variance.
- The v2 system shows statistically significant lead time extension by 6 days in gauged basins and 1 day in ungauged basins relative to the v1 nowcast.
- Performance improvements are correlated with arid environments for gauged basins and with lake and reservoir indicators in ungauged basins, indicating less improvement in managed or regulated catchments .
- The addition of GraphCast appears to be the largest driver of skill increases in ungauged basins, especially at longer lead times.

5.2 Global Data Frameworks & Open Science

The technical advancements described in this paper and the associated open data and open source code releases represent an attempt to support the principles of open science in global hydrology. The future of operational AI hydrology depends on the availability of standardized, historical data archives and real-time data streams. We would like to highlight and endorse ongoing efforts by the World Meteorological Organization (WMO) to establish a unified global data infrastructure. Data collected by National Meteorological and Hydrology Services

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

(NMHSs), and global efforts to collect and standardize hydrology data are critical for the advancement of large sample hydrology (Addor et al., 2020).

Specifically, we highlight three initiatives that warrant the support and contributions of the hydrological science and operations communities:

- **WMO Unified Data Policy (Resolution 1, Cg-Ext(2021)):** This resolution commits Member states to the free and unrestricted exchange of Earth system data. It provides the political and legal foundation necessary for training robust, less-biased AI models (WMO, 2021).
- **WMO Information System 2.0 (WIS 2.0):** By transitioning from legacy message-switching to a modern, pub/sub-based data exchange framework, WIS 2.0 lowers the barrier to entry for NMHSs to share and discover data in real-time (WMO, 2023a).
- **WMO Integrated Processing and Prediction System (WIPPS):** Formerly the GDPFS, WIPPS creates a standardized mechanism for cascading high-quality global model products (such as the ones presented here) down to local services. This ensures that advancements in global AI forecasting can be systematically integrated into national early warning systems (WMO, 2023b).

By releasing open-access datasets like the Google Runoff Reanalysis & Reforecast (GRRR) dataset (see Code and Data Availability Section) and Caravan (Kratzert et al., 2023), we aim to accelerate the collective ability of the scientific community to validate, improve, and deploy flood forecasting systems at both global and local scales.

6 Code and Data Availability

To support transparency, reproducibility, and future research in large-sample hydrology, the code, data, and model outputs associated with this study are made publicly available.

- **Codebase:** The open-source GoogleHydrology framework provides the ability to train and evaluate models that are structurally and functionally identical to the models used in this study. This is available on GitHub at: https://github.com/google-research/flood-forecasting.
- **Archived Release:** A persistent, citable archive containing the v2 model outputs and analysis code used in this paper are published (Nearing et al., 2026) under Zenodo with a CCA-4 license (https://doi.org/10.5281/zenodo.19676842).
- **Benchmark Data:** Data for the v1 and GloFAS benchmarks are published (Nearing et. al., 2023b) under Zenodo with a CCA-2 Non-Commercial license (https://doi.org/10.5281/zenodo.813937).

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

- **Additional Data Release:** The Google Runoff Reanalysis & Reforecast (GRRR) dataset contains global river discharge estimates and hydrologic predictions for approximately 1M locations is publicly accessible via the Google Flood Forecasting resources page at https://sites.research.google/gr/floodforecasting/resources/. This dataset is provided to support open science by allowing validation with local data almost anywhere in the world, but was not used in this publication.

Author Contributions

Grey Nearing wrote the paper with help from Martin Gauch, Frederik Kratzert, and Deborah Cohen.

Martin Gauch and Grey Nearing ran the v2 model runs used in this paper. Frederik Kratzert performed the Geoglows benchmarking. Grey Nearing performed the data analysis.

Guy Shalev, Martin Gauch, Frederik Kratzert, Deborah Cohen, Ben Feinstein, Asher Metzger, Daniel Klotz, Ido Zemach, Oren Gilon, Rotem Green, Rotem Mayo, Oleg Zlydenko, Grey Nearing contributed to developing the v2 model.

Rony Amira, Omri Shefi, Amit Markel, Martin Gauch, Frederik Kratzert, Gila Loike and Grey Nearing contributed to developing the GoogleHydrology open source codebase.

Aviel Niego, Hadas Fester, Rom Aschner, Dan Korenfeld, Yuval Shildan, Benny Mosheyev, Amitay Sicherman, Frederik Kratzert, Martin Gauch, Guy Shalev, Oren Gilon contributed to operationalizing the v2 model.

Stephanie Rees, Emily Reinstein, Gila Loike, Yuval Carny, Grey Nearing, Asher Metzger supported international partnerships that provided feedback to guide model development.

Yossi Matias, Avinatan Hassidim, Deborah Cohen, Oren Gilon, and Shmulik Fronman provided team leadership.

Acknowledgements

During the preparation of this work the authors used Gemini to edit and improve the readability of the manuscript. After using this service, the authors reviewed and edited the entirety of the text and take full responsibility for the content of the publication.

Competing Interests

All authors are employed by Google, the organization that developed and operates the Google

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Global Flood Forecasting system and the associated open-source GoogleHydrology codebase evaluated in this manuscript. Additionally, the authors may hold stock or share ownership in Alphabet Inc., Google's parent company.

Financial Support

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Acuña Espinoza, Eduardo, et al. "Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events." Hydrology and Earth System Sciences 29.5 (2025): 1277-1294.

Addor, Nans, et al. "The CAMELS data set: catchment attributes and meteorology for large-sample studies." Hydrology and Earth System Sciences 21.10 (2017): 5293-5313.

Addor, Nans, et al. "Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges." Hydrological Sciences Journal 65.5 (2020): 712-725.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. "GloFAS – global ensemble streamflow routing and flood early warning." Hydrology and Earth System Sciences 17.3 (2013): 1161-1175.

Allen, Richard G., et al. "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56." Fao, Rome 300.9 (1998): D05109.

Alvarez-Garreton, Camila, et al. "The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies-Chile dataset." Hydrology and Earth System Sciences 22.11 (2018): 5817-5846.

Arsenault, Richard, et al. "Continuous streamflow prediction in ungauged basins: Long Short-Term Memory Neural Networks clearly outperform hydrological models." Hydrology and Earth System Sciences Discussions 2022 (2022): 1-29.

Arsenault, Richard, et al. "A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds." Scientific Data 7.1 (2020): 243.

Başağaoğlu, Hakan, et al. "A review on interpretable and explainable artificial intelligence in hydroclimatic applications." Water 14.8 (2022): 1230.

Ben Bouallègue, Zied, et al. "The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context." *Bulletin of the American Meteorological Society* 105.6 (2024): E864-E883.

Casado Rodriguez, J.: CAMELS-ES: Catchment Attributes and Meteorology for Large-Sample Studies Spain, https://doi.org/10.5281/zenodo.8373020, 2023.

Chagas, Vinicius BP, et al. "CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil." Earth System Science Data Discussions 2020 (2020): 1-41.

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Chaudhary, Priyanka, et al. "Flood uncertainty estimation using deep ensembles." Water 14.19 (2022): 2980.

Chen, Mingyue, et al. "Assessing objective techniques for gauge-based analyses of global daily precipitation." *Journal of Geophysical Research: Atmospheres* 113.D4 (2008).

Coxon, Gemma, et al. "CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain." Earth System Science Data Discussions 2020 (2020): 1-34.

Dai, Zhihui, et al. "A hydrological data prediction model based on LSTM with attention mechanism." Water 15.4 (2023): 670.

De la Fuente, Luis Andres, et al. "Towards interpretable LSTM-based modelling of hydrological systems." Hydrology and Earth System Sciences Discussions 2023 (2023): 1-36.

Deng, Chao, et al. "Catchment runoff simulation by coupling data assimilation and machine learning methods." Advances in Water Science 34.6 (2023): 839-849.

Färber, Claudia, et al. "GRDC-Caravan: extending caravan with data from the global runoff data centre." Earth System Science Data Discussions 2024 (2024): 1-17.

Feng, Dapeng, Kuai Fang, and Chaopeng Shen. "Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales." Water Resources Research 56.9 (2020): e2019WR026793.

Feng. Dapeng, et al. "Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy." Water Resources Research 58.10 (2022): e2022WR032404.

Feng, Dapeng, et al. "The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment." Hydrology and Earth System Sciences 27.12 (2023): 2357-2373.

Fowler, Keirnan JA, et al. "CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia." Earth System Science Data Discussions 2021 (2021): 1-30.

Frame, Jonathan M., et al. "On strictly enforced mass conservation constraints for modelling the Rainfall-Runoff process." Hydrological Processes 37.3 (2023): e14847.

Gauch, Martin, et al. "In defense of metrics: Metrics sufficiently encode typical human preferences regarding hydrological model performance." Water Resources Research 59.6 (2023): e2022WR033918.

Gauch, Martin, et al. "How to deal with missing input data." Hydrology and Earth System Sciences 29.21 (2025): 6221-6235.

Gupta, Hoshin V., et al. "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling." Journal of hydrology 377.1-2 (2009): 80-91.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., ... & Pappenberger, F. "GloFAS v4.0: a global hydrological ensemble forecast system." (2023).

Helgason, Hordur Bragi, and Bart Nijssen. "LamaH-Ice: LArge-SaMple data for hydrology and environmental sciences

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

for Iceland." Earth System Science Data 16.6 (2024): 2741-2771.

Hoedt, Pieter-Jan, et al. "Mc-lstm: Mass-conserving lstm." International conference on machine learning. PMLR, 2021.

Höge, Marvin, et al. "Improving hydrologic models for predictions and process understanding using neural ODES." Hydrology and Earth System Sciences Discussions 2022 (2022): 1-29.

Höge, Marvin, et al. "CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland." Earth System Science Data Discussions 2023 (2023): 1-46.

Hosseini Hossein Abadi, Farzad, Cristina Prieto Sierra, and César Álvarez Diaz. "An explainable AI approach for interpreting regionally optimized deep neural networks in hydrological prediction." (2025).

Huffman, George J., et al. "Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG)." *Satellite precipitation measurement: Volume 1*. Cham: Springer International Publishing, 2020. 343-353.

Jia, Xiaowei, et al. "Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles." ACM/IMS Transactions on Data Science 2.3 (2021): 1-26.

Klingler, Christoph, Karsten Schulz, and Mathew Herrnegger. "Lamahl large-sample data for hydrology and environmental sciences for central europe." Earth System Science Data Discussions 2021 (2021): 1-46.

Klotz, Daniel, et al. "Uncertainty estimation with deep learning for rainfall-runoff modeling." Hydrology and Earth System Sciences 26.6 (2022): 1673-1693.

Kratzert, Frederik, et al. "NeuralHydrology-interpreting LSTMs in hydrology." Explainable AI: Interpreting, explaining and visualizing deep learning. Cham: Springer International Publishing, 2019. 347-362.

Kratzert, Frederik, et al. "Toward improved predictions in ungauged basins: Exploiting the power of machine learning." Water Resources Research 55.12 (2019): 11344-11354.

Kratzert, Frederik, et al. "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets." Hydrology and Earth System Sciences 23.12 (2019): 5089-5110.

Kratzert, Frederik, et al. "A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling." Hydrology and Earth System Sciences Discussions 2021 (2021): 1-26.

Kratzert, Frederik, et al. "Caravan-A global community dataset for large-sample hydrology." Scientific Data 10.1 (2023): 61.

Kratzert, Frederik, et al. "HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin." Hydrology and Earth System Sciences 28.17 (2024): 4187-4201.

Krejči, J. & Nearing, G. (2025). CAMELS-CZ: A catchment attributes and meteorology dataset for large-sample studies in Czechia [Dataset]. Zenodo. https://www.google.com/search?q=https://doi.org/10.5281/zenodo.17593968.

Lam, Remi, et al. "Learning skillful medium-range global weather forecasting." Science 382.6677 (2023): 1416-1421.

Linke, Simon, et al. "Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution."

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

Scientific data 6.1 (2019): 283.

Liu, Jinping. "Progresses and Challenges on QPE/QPF Utilization in Hydrology." Tropical Cyclone Research and Review 1.2 (2012): 194-197.

Liu, Siyan, et al. "Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions." Frontiers in Water 5 (2023): 1150126.

Liu, Jun, et al. "CAMELS-DK: Hydrometeorological Time Series and Landscape Attributes for 3330 Catchments in Denmark." Earth System Science Data Discussions 2024 (2024): 1-30.

Lees, Thomas. Deep learning for hydrological modelling: from benchmarking to concept formation. Diss. University of Oxford, 2022.

Loritz, Ralf, et al. "CAMELS-DE: hydro-meteorological time series and attributes for 1555 catchments in Germany." Earth System Science Data Discussions 2024 (2024): 1-30.

Mai, Juliane, et al. "The great lakes runoff intercomparison project phase 4: the great lakes (GRIP-GL)." Hydrology and Earth System Sciences Discussions 2022 (2022): 1-54.

Morin, E. "Caravan extension Israel-Israel dataset for large-sample hydrology, Zenodo [data set]." 2023,

Muñoz-Sabater, Joaquin, et al. "ERA5-Land: A state-of-the-art global reanalysis dataset for land applications." Earth system science data 13.9 (2021): 4349-4383.

Nash, J. Eamonn, and Jonh V. Sutcliffe. "River flow forecasting through conceptual models part I-A discussion of principles." Journal of hydrology 10.3 (1970): 282-290.

Nearing, Grey, et al. "Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks." Hydrology and earth system sciences discussions 2021 (2021): 1-25.

Nearing, G., et al.: From Hindcast to Forecast with Deep Learning Streamflow Models, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-10020, https://doi.org/10.5194/egusphere-egu23-10020, 2023a.

Nearing, G.: Global prediction of extreme floods in ungauged watersheds, Zenodo [data set], https://doi.org/10.5281/zenodo.8139379, 2023b.

Nearing, Grey, et al. "Global prediction of extreme floods in ungauged watersheds." Nature 627.8004 (2024): 559-563.

Nearing, G., Kratzert, F., and Gauch, M.: Extending Medium-Range Global Flood Forecasts: The Google Global Flood Forecasting Model Version 2, Zenodo [data set], https://doi.org/10.5281/zenodo.19676842, 2026.

Newman, Andrew J., et al. "Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance." Hydrology and Earth System Sciences 19.1 (2015): 209-223.

Nevo, Sella, et al. "Flood forecasting with machine learning models in an operational framework." Hydrology and Earth System Sciences 26.15 (2022): 4013-4032.

Nourani, Vahid, Kasra Khodkar, and Mekonnen Gebremichael. "Uncertainty assessment of LSTM based groundwater

https://doi.org/10.5194/egusphere-2026-2283
Preprint. Discussion started: 29 April 2026

level predictions." Hydrological Sciences Journal 67.5 (2022): 773-790.

Pearson, Karl. "VII. Note on regression and inheritance in the case of two parents." proceedings of the royal society of London 58.347-352 (1895): 240-242.

Prieto, Cristina, et al. "Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests." Water Resources Research 55.5 (2019): 4364-4392.

Rosati, Michael, et al. "Decoding LSTM Memory to Reveal Baseflow Contributions in Fractured and Sedimentary Mountain Basins: A Case Study in the Sangre de Cristo Mountains, Southwestern United States." (2025).

Sabzipour, Behmard, et al. "Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment." Journal of Hydrology 627 (2023): 130380.

Shalev, Guy, and Frederik Kratzert. "Caravan MultiMet: Extending Caravan with Multiple Weather Nowcasts and Forecasts." arXiv preprint arXiv:2411.09459 (2024).

Souffront Alcantara, M. A., Nelson, E. J., Shakya, K., Edwards, C., Roberts, W., Krewson, C., ... & Jones, N. "Hydrologic modeling as a service (HMaaS): a new approach to address hydroinformatic challenges in developing countries." Frontiers in Environmental Science 7 (2019): 158.

Tsai, Wen-Ping, et al. "From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling." Nature communications 12.1 (2021): 5988.

WMO (2021). WMO Unified Policy for the International Exchange of Earth System Data. Resolution 1 (Cg-Ext(2021)). World Meteorological Organization.

WMO (2023a). Manual on the WMO Information System (WIS). WMO-No. 1060. World Meteorological Organization.

WMO (2023b). Manual on the WMO Integrated Processing and Prediction System (WIPPS). WMO-No. 485. World Meteorological Organization.