AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2025WR042606

Special Collection:
Advances in large-scale hydrological modeling and prediction under global change

Reconstructing China's Natural Streamflow at 1-km Resolution

Ningpeng Dong[1], Mingxiang Yang[1], Jianhui Wei[2], Shiqin Xu[3], Yong Zhao[1], Xuejun Zhang[1], Bo Liu[4], Mengqi Wu[5], Hao Wang[1], and Harald Kunstmann[2,6,7]

[1]State Key Laboratory of Water Cycle and Water Security, China Institute of Water Resources and Hydropower Research, Beijing, China, [2]Institute of Meteorology and Climate Research (IMKIFU), Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany, [3]Hydrology, Agriculture and Land Observation (HALO) Laboratory, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [4]School of Earth Sciences and Engineering, Hohai University, Nanjing, China, [5]Hubei Water Saving Research Center, Hubei Water Resources Research Institute, Wuhan, China, [6]Institute of Geography, University of Augsburg, Augsburg, Germany, [7]Centre for Climate Resilience, University of Augsburg, Augsburg, Germany

Abstract Climate change and intensifying human activities are placing unprecedented pressure on China's water resources, necessitating high-resolution, naturalized streamflow records for effective management. We develop a 1-km gridded fully coupled land-surface-hydrologic-hydrodynamic modeling system for China based on the CLHMSv2.0 model. To refine grid-level runoff simulations, we propose a machine learning based grid-scale calibration approach using observed streamflow and climatic and physiographic attributes from thousands of Chinese catchments. To represent China's complex river-lake networks at fine scales, we introduce an optimized hybrid 1D/2D diffusion wave scheme to resolve backwater effects and bidirectional exchanges between rivers and lakes. The modeling system is extensively validated at daily and monthly scales against 1,225 flow gauges and 5 lake water level gauges across the nation. Streamflow evaluations show that the median daily (monthly) Nash-Sutcliffe efficiency is 0.57 (0.77) and the daily (monthly) Kling-Gupta Efficiency is 0.65 (0.73); water level simulations for major freshwater lakes are also satisfactory, with correlation coefficients mostly above 0.80. Applying the system over 1962–2024, we generate the national 1-km gridded daily natural streamflow estimates named CHASE v1.0, which is conditionally available at https://hydrodata.cn/chase.

Key Points:

- A 1-km gridded fully coupled land-surface-hydrologic-hydrodynamic modeling system is developed for China
- A machine-learning-based calibration approach is introduced to refine grid-scale parameters using observations from thousands of gauges
- The newly developed model shows high accuracy and robustness in nationwide streamflow and lake water level simulations

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

N. Dong,
dongnp@iwhr.com

Citation:

Dong, N., Yang, M., Wei, J., Xu, S., Zhao, Y., Zhang, X., et al. (2026). Reconstructing China's natural streamflow at 1-km resolution. *Water Resources Research*, *62*, e2025WR042606. https://doi.org/10.1029/2025WR042606

Received 19 OCT 2025
Accepted 26 MAY 2026

1. Introduction

Among all the countries globally, China's monsoon climate, steep terrain and dam cascade development jointly generate one of the highest spatiotemporal variabilities of runoff (Liu et al., 2020; Xu et al., 2023). Given that anthropogenic regulation has altered a majority of the nation's river system (Han et al., 2024; Wang et al., 2025), observation-based records are increasingly non-stationary and no longer reflect the natural water cycle. High-resolution naturalized streamflow data are therefore urgently required for water related studies and sustainable water resources management (Chen et al., 2016; Vörösmarty et al., 2010; Wada et al., 2017).

Over the past decade, several modeling paradigms have emerged to reconstruct natural streamflow at continental-to-global scales. One prominent paradigm couples a gridded or sub-catchment runoff generator with a vectorized river-network routing module. For instance, Lin et al. (2019) combined runoff from VIC model (Liang et al., 1994) with the RAPID Muskingum routing model to estimate daily discharge for 2.94 million river reaches. Miao et al. (2022) and Ghimire et al. (2023) used the VIC model coupled with routing models to reconstruct flow records over China and CONUS, respectively. Feng et al. (2022, 2024) and Song et al. (2025) developed the differentiable δHBV and δHBV2.0 models to simulate runoff generation and flow routing with unit-hydrograph and Muskingum schemes at global and CONUS scales. Akpoti et al. (2024) coupled the VegET agro-hydrologic model with the mizuRoute routing model (Mizukami et al., 2016, 2021) to produce the daily streamflow product for Africa. More recently, Yang et al. (2025) introduced a Grid-LSTM-RAPID system to generate the GRADES-hydroDL data set. These frameworks achieve computational efficiency by decoupling runoff generation from channel routing, and they also allow the coupling of vectorized routing models that often represent channels more precisely (Alfieri et al., 2020; David et al., 2011). In recent years, such couplings have been increasingly supported by interface standards and coupling frameworks, such as the CSDMS Basic Model Interface (BMI) and GLOFRIM (Hoch et al., 2017; Hutton et al., 2020; Peckham et al., 2013).

Another major paradigm employs fully distributed, grid-based hydrologic models that solve water budgets and routing equations for each grid cell in an integrated way. Operational examples include the Global Flood Awareness System (GloFAS), which uses the grid-based LISFLOOD hydrological model for global ensemble flood forecasting (Alfieri et al., 2013, 2020; Hirpa et al., 2018), and global hydrological models such as PCR-GLOBWB, H08, and WaterGAP (Hanasaki et al., 2022; Hoch et al., 2023; Müller Schmied et al., 2014, 2021). Regional or continental-scale implementations include high-resolution PCR-GLOBWB over Europe (van Jaarsveld et al., 2025), LISFLOOD over Europe (Tilloy et al., 2025), and ParFlow-CLM over Europe and North America (Orth et al., 2016; O'Neill et al., 2021). A potential limitation of this paradigm is that a precise river network representation requires fine spatial discretization, which makes large-domain simulations computationally expensive (Hoch et al., 2023; van Jaarsveld et al., 2025; Shrestha et al., 2025). Despite that, this paradigm can provide a more integrated and physically consistent framework and has the potential for simulating processes such as rapid river-groundwater exchanges and floodplain dynamics, which are critical in China's widespread karst terrains and river-lake systems.

Beyond structural design, accurate parameter estimation is another research hotspot in large-scale hydrologic modeling. One common estimation approach is to calibrate sensitive parameters at gauged basins and then extrapolate to ungauged areas via distance-based, similarity-based, or multi-scale regionalization schemes. Distance-based schemes rely on geographic proximity (Oudin et al., 2008); similarity-based approaches cluster donor catchments using attributes such as aridity index, slope, or soil texture (Beck et al., 2016); while multi-scale schemes establish parameter-attribute relationships via transfer functions (Gou et al., 2020; Samaniego et al., 2010). Recently, the Large-Sample Emulator (LSE) was introduced as a novel calibration framework that jointly trains an emulator across a diverse collection of basins to optimize and regionalize parameters for process-based hydrological models, demonstrating effective prediction in ungauged basins (Farahani et al., 2025; Tang et al., 2025). Alternatively, spatially continuous hydrological products provide an emerging basis for grid-level calibration (Lin et al., 2019; L. Yang et al., 2021, Y. Yang et al., 2021). For example, utilizing gridded runoff curve number, SMAP soil moisture, GRACE terrestrial water storage, or GLEAM evapotranspiration has been shown to help constrain model parameters (Hong et al., 2007; López et al., 2017; Mei et al., 2023; Niu et al., 2020; Rajib et al., 2018). For example, Xie et al. (2021) demonstrated that jointly using runoff and evaporation products can reduce distributed parameter uncertainty in the VIC model. These studies highlight the promise of grid-based calibration using spatially continuous signatures for reducing parameter uncertainties in large-domain simulations, which is particularly relevant for a country as vast and hydrologically diverse as China.

Despite such advances, global models and global signature products remain poorly constrained across China. Due to limited accessibility of Chinese gauge data in public repositories, global models such as GloFAS, PCR-GLOBWB, WaterGAP, H08, GRADES and δHBV incorporate very few Chinese gauges for calibration (Harrigan et al., 2020; L. Yang et al., 2021, 2021, 2025). Likewise, large-scale runoff reconstruction products, such as Global RUNoff reconstruction (GRUN; Ghiggi et al., 2019), Linear Optimal Runoff Aggregate (LORA; Hobeichi et al., 2019), and FLO1K (Barbarossa et al., 2018), and signature products such as Global Streamflow Characteristics Data set (GSCD; Beck et al., 2015), are mostly trained on a handful of Chinese stations, which could limit their ability to capture the rainfall-runoff non-linearity induced by China's complex terrain and the distinct East Asian monsoon. To date, Gou et al. (2021) and Miao et al. (2022) remain two of the few national-scale modeling efforts that use more than a hundred gauges (200–300 gauges) for parameter calibration and regionalization, which provides valuable data sources for China's hydrologic studies.

To address these deficiencies, we develop a 1-km gridded national land-surface-hydrologic-hydrodynamic modeling system based on the Coupled Land surface-Hydrologic Model System Version 2.0 (CLHMSv2.0) model that is specifically tailored for China. To refine runoff simulations at grid scales, 1-km gridded natural runoff-depth and baseflow index maps of China are generated for model calibration by training machine learning models on observed streamflow and climatic and physiographic attributes from thousands of Chinese catchments. Additionally, a hybrid 1D/2D diffusion wave solver is introduced to depict China's complex river-lake systems under 2D hydrodynamic conditions, such as those in the lower Huai, Yangtze, and Tarim river basins. The modeling system is then extensively validated against 1,225 flow gauges and 5 lake water level gauges across the nation, and is finally applied over 1962–2024 to generate the 1-km gridded daily natural streamflow estimates.

2. Model Development and Implementation

China's river system is among the most extensive and hydrologically diverse worldwide. Major basins include the Yangtze, Yellow, Pearl, Huai, Hai, Songhua, and Liao, each characterized by distinct hydroclimatic regimes (Guo et al., 2023; Ma et al., 2022). In particular, river-lake systems such as Dongting and Poyang (An et al., 2022) introduce strong surface-groundwater coupling, floodplain retention, and backwater effects.

To depict the complex hydrologic regimes of China, we adopted the fully distributed, grid-based modeling diagram at 1-km resolution based on the CLHMSv2 model. A gridded framework ensures process consistency and spatial fidelity by integrating water/energy states and fluxes such as runoff generation, soil-groundwater exchange, channel routing, and lake/floodplain storage on a single, physically consistent mesh. These advantages are particularly relevant in China, where shallow water tables and karst landscapes lead to highly localized groundwater fluxes, and where river-lake systems exhibit distributed storage and strong backwater effects. Built on this consideration, in the following sections we introduce the implementation of a 1-km gridded national land-surface-hydrologic-hydrodynamic modeling system based on the CLHMSv2 model.

2.1. Model Structure and Its Modifications

The CLHMS (Dong et al., 2022, 2023; Hao et al., 2024) is a fully coupled model that integrates a land surface scheme (LSX) with a physically based hydrological model (HMS) (Yu et al., 1999, 2006). The coupled architecture allows full depiction of vegetation-atmosphere exchanges, soil hydrodynamics, snow and glacier energy balances, river routing, and groundwater dynamics, thereby providing an integrated framework for land-hydrology simulations (Fersch et al., 2013; Wagner et al., 2016). To be specific, LSX simulates vegetation, soil, snow, and glacier processes, while HMS represents groundwater dynamics and surface routing. The two components are fully coupled primarily through the exchange flux between the bottom soil layer and the groundwater layer computed using the Darcy-Buckingham equation, as opposed to the free-drainage bottom boundary for most traditional land surface models (see Fersch et al., 2013 for more details). All processes are solved synchronously at each time step to form a fully coupled model architecture, as further illustrated in Figure 1.

Vegetation processes are resolved through a two-layer canopy structure consisting of an upper tree layer and a lower grass layer. Radiative fluxes and turbulent exchanges of momentum, sensible heat, and water vapor between the canopy and the soil surface are calculated at every time step, including transpiration and precipitation interception.

Soil is partitioned into six layers over the upper 2.5 m, with prognostic simulation of soil temperature and moisture (liquid and ice) for each layer. Vertical processes include heat diffusion, unsaturated liquid water transport, saturated gravitational drainage, local surface runoff, water exchange with groundwater, uptake of soil moisture by plant roots, and explicit freeze-thaw dynamics of soil ice.

Snow and glacier dynamics are simulated with an explicit energy-balance scheme. The scheme enables simulation of snowfall accumulation, compaction, melt infiltration, refreezing, and retention of liquid water. The glacier is treated as a single thermodynamically active layer, with its mass balance updated each time step by solving the surface energy budget.

Groundwater dynamics are represented by an explicit single-layer groundwater module, which captures the groundwater level and lateral flow flux using 2D Boussinesq equation. The groundwater module allows bi-directional vertical recharge and discharge with rivers/lakes, driven by the head difference between groundwater and surface water levels, which is resolved through the diffusion wave routing process.

In the original implementations of the model, river routing is solved using diffusion wave equations with the Gauss-Seidel iteration method, which are efficient for coarse resolutions (10–50 km) but computationally unaffordable for kilometer-scale domains. To solve the diffusion wave equations on the 1D river channels and 2D river-lake systems simultaneously at a national 1-km scale, in this study, we replaced the legacy iteration method by a mixed implicit strategy that couples (a) a local, Jacobi-type iterative implicit update in 1D subsystems with (b) a global sparse Krylov solver on selected 2D subsystems. Using OpenMP parallelization, the full national 1-km model takes ∼2 hr to simulate 1 year of runtime on a single-node cluster with 96 CPU cores, which is about 10 times faster than the previous implementation.

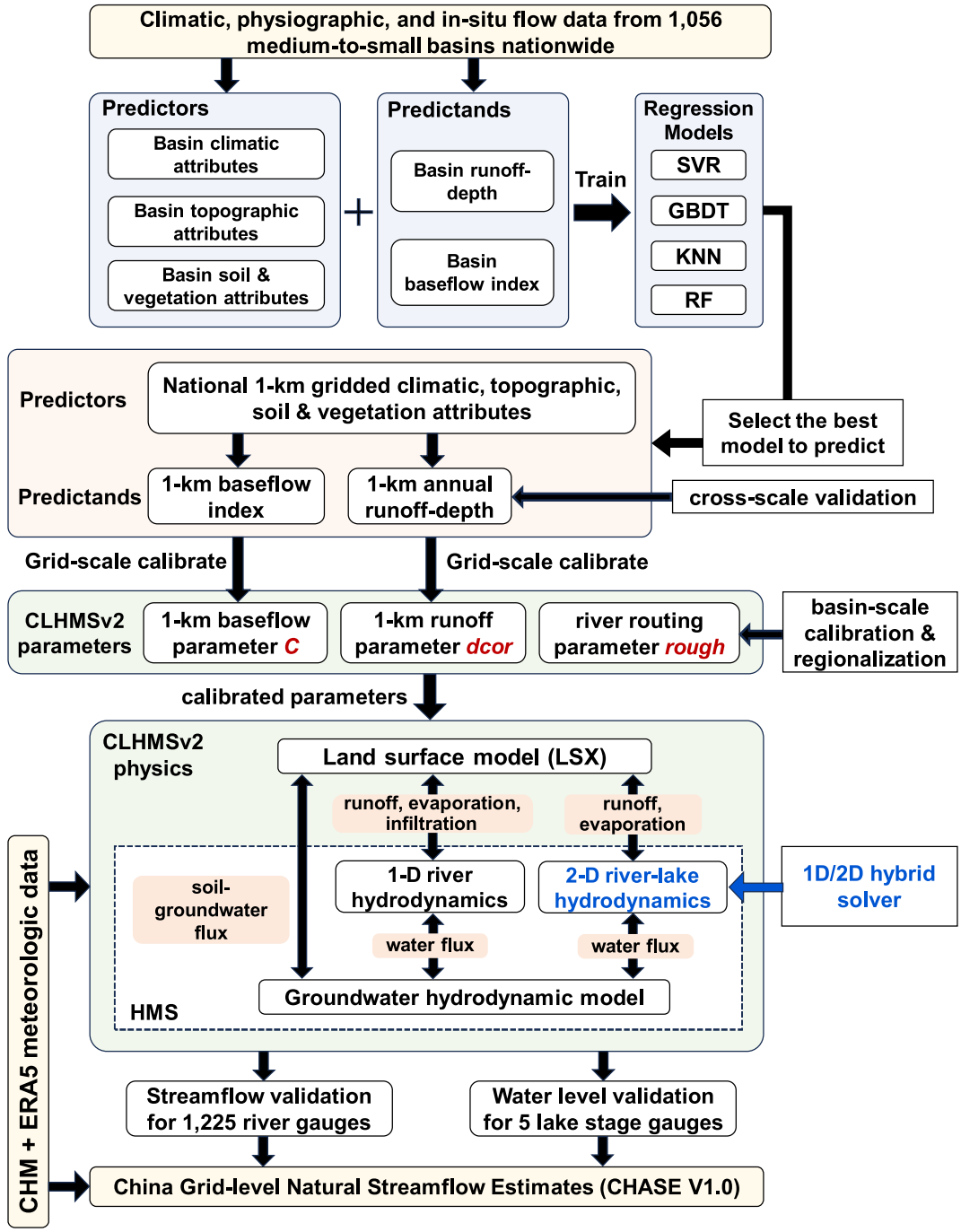

Figure 1. Overall workflow of this study, including the calibration procedure and the CLHMSv2 model parameters and structure.

For 1D subsystems, the new water level of a grid cell in a fully implicit step depends on neighbors via nonlinear Manning fluxes, which leads to a large coupled system. Here, we introduce the diagonal-implicit update as a local implicit method (Anderson et al., 1996), which freezes neighbor values and linearizes each flux as a coefficient times the head difference. This method leaves a single unknown per cell, and a few iterations over the grid can yield a stable approximation to the implicit solution without assembling or factoring large matrices. This Jacobi-type iteration is suitable for shared-memory vectorization due to its strictly local data access.

For 2D subsystems, we form a linearized diffusion wave equation for the local patch, which is stored in a compressed sparse form. This system is solved with a Krylov method of preconditioned conjugate gradient (PCG) plus simple diagonal scaling (Van der Vorst, 2003). PCG solvers work well because they use only sparse matrix-vector multiplies and basic vector operations, so they handle large blocks with much lower memory and bandwidth.

By leveraging the above solvers, we partitioned the entire routing domain into 1D river subsystems and selected 2D river-lake subsystems. Here, the 2D river-lake subsystems are defined for China's largest freshwater lakes that are embedded within river networks, where the 2D diffusive-wave scheme is used to represent backwater effects and bidirectional lake-river exchanges. The 2D domains can also be expanded to include floodplains or wetlands if specified, and the 1D diffusive-wave scheme is applied elsewhere along the river network. Overland flow routing is not activated because the model runs at 1 km^2 grid cell on a daily scale, and localized runoff is assumed to reach the channel well within 1 day. An additional hillslope overland flow routing solver would therefore add complexity with limited benefit for the intended purpose of daily streamflow reconstruction.

2.2. Model Inputs

To drive the model, we compile a consistent input data set spanning meteorology, soils-aquifer properties, hydrography, and vegetation.

Meteorological forcings. Daily precipitation is taken from CHM_PREV2, a 0.1° product generated by merging 3,746 gauges with 11 precipitation-related covariates through a LightGBM-based scheme (Hu et al., 2025). Model also requires air temperature, solar radiation, specific humidity, air pressure, and wind speed, which are derived from ERA5-Land reanalysis at 0.1°resolution (Muñoz-Sabater et al., 2021).

Soil and Aquifer Properties. Soil physical attributes such as clay and sand fractions are derived from the Harmonized World Soil Database (HWSD v1.2) at 1-km resolution (Wieder, 2014). Groundwater parameters, including aquifer thickness and specific yield, are taken from the China National Geological Survey data set (MGMR, 1990) and subsequent refinements by Yang et al. (2010).

River Network, Lake, and Hydrographic Basins. River network, flow directions, and catchment boundaries are derived from the MERIT Hydro data set (Yamazaki et al., 2019), with manual corrections applied. While MERIT Hydro is generally accurate, we found a few local mismatches mainly at small tributary junctions and short reach connections, and we manually corrected a small number of these cases. Lake extents were extracted from HydroLAKES (Messager et al., 2016). To better represent lake bathymetry at 1-km, we lowered the DEM elevation within each mapped lake extent using lake terrain offsets relative to the surrounding land, based on lake terrain information provided by local water authorities.

Land Cover. Land-cover classification is based on the AVHRR 1-km Global Land Cover Characterization (GLCC) data set (Loveland et al., 2000), which provides consistent vegetation type distribution across China.

All gridded inputs are processed onto a national 1-km grid in a Lambert azimuthal equal-area (LAEA) projection. For climate forcing and soil and aquifer properties, we use bilinear interpolation. At coastlines and domain boundaries where a target cell has fewer valid surrounding coarse-grid cells, the nearest available valid coarse-grid value is filled. For land use, we use nearest-neighbor resampling. For flow directions and DEM, we upscale from the fine grid by retaining the value at the location of maximum flow accumulation within each 1-km cell to preserve major river information.

3. A Machine Learning Based Calibration Framework at Grid Scale

3.1. Overall Workflow

Traditional calibration adjusts model parameters within each basin to a single set of spatially uniform values that best reproduces outlet streamflow. This approach overlooks within-basin spatial heterogeneity and can lead to artificial discontinuities in parameters across neighboring basins. Here we propose a grid-based calibration framework that produces spatially consistent national parameter fields. Using 1,056 medium-to-small sized basins with at least 5 years of flow observations during 1962–1979, we first calculate two basin signatures, namely annual runoff depth and baseflow index. For each signature, we train four machine-learning regression models (RF, GBDT, SVR, and KNN) to relate the signature to basin attributes describing climate, topography,

soils, and vegetation. We then apply the best regression model out of these four models to national 1-km gridded attributes to generate 1-km gridded runoff-depth and baseflow index maps. Then, the runoff-depth map is used to calibrate the CLHMS runoff-generation parameter *dcor* at 1-km grid scale, and the baseflow index map is used to calibrate the CLHMS baseflow parameter *C* at 1-km grid scale. Routing-related CLHMS parameter *rough* is calibrated at the basin scale using flow data from all 1,225 river flow gauges and transferred using physically based similarity regionalization (PSR). Details of parameters *dcor* and *C* can be found in Section 3.4. The resulting parameter fields drive the coupled CLHMSv2 model, and simulated streamflow is evaluated against gauge observations. Finally, with climate forcing for 1962–2024 as inputs, the model generates the long-term naturalized flow records for China, named CHinA grid-level natural Streamflow Estimates Version 1.0 (CHASE v1.0). The full workflow is summarized in Figure 1.

3.2. Gauge Data

A key prerequisite for producing naturalized streamflow is to ensure that the calibration and evaluation data are minimally affected by human activities. In China, however, near-natural flow regimes become increasingly rare after 1980s because of high population density and extensive hydraulic construction. To reduce these influences, we restricted streamflow data retrieval to 1962–1979. This choice is consistent with previous assessments showing that less than 9% of ∼1,700 basins can be classified as near-natural in the post-1980 period (L. Yang et al., 2021, Y. Yang et al., 2021). Our choice also aligns with the state-of-the-art China Natural Runoff Data set (Gou et al., 2021), which also adopted a pre-1980 period for model calibration and validation.

Within 1962–1979, we further screened gauges to identify near-natural basins and periods using two human-impact thresholds: (a) an upstream degree of regulation (DOR < 5%) derived from reservoir information in the national database, and (b) a withdrawal-to-flow ratio <10% computed from a gridded water-withdrawal data set from our previous work (Dong et al., 2022). Gauges exceeding either threshold were excluded. After screening, 1,225 flow gauges nationwide were retained. All retained flow gauges have at least 5 years of daily observations, but do not necessarily share a common set of years, as early records are not always continuous. Figure 2 summarizes the data length for each flow gauge.

These 1,225 gauges have a catchment area ranging from 14 km^2 to 1.8×10^6 km^2, in total covering approximately 60% of the national land area and span a wide range of hydrological contexts. The median and mean catchment areas are 2,200 and 18,050 km^2, respectively, and around 15% of gauges have catchment areas less than 500 km^2, which differs from previous national-scale modeling efforts that focus on larger basins (Dong et al., 2022). The density of gauges ensures that diverse hydroclimatic regimes are represented in the calibration process. In contrast, regions without gauge coverage are concentrated in three types of environments: (a) densely populated plains in the North China Plain where river networks are highly artificial and natural runoff data series do not exist, (b) sparsely inhabited areas of the endorheic Tibetan Plateau, where no gauges are built, and (c) Gobi Desert, where perennial rivers are absent (Figure 2).

In addition to flow gauges, we include 5 lake water-level gauges in the model evaluation (Figure 2), because these gauges provide in situ water level for assessing whether the model can reproduce backwater effects and bidirectional river-lake exchanges, which can eventually be used to validate the simulated streamflow in lake-influenced reaches. These include water level records from China's largest freshwater lakes, namely, Xingzi (Poyang Lake, ∼4,000 km^2 water area), Chenglingji (Dongting Lake, ∼2,000 km^2), Dapukou (Tai Lake, ∼2,500 km^2), Huailinzhen (Chao Lake, ∼800 km^2) and Jiangba (Hongze Lake, ∼2,500 km^2) (see Figure 9 for locations of these gauges in the lakes).

3.3. Deriving China's Gridded Natural Runoff Depth and Baseflow Index Using Machine Learning

In this section, we employed a machine learning-based framework to generate a gridded map of China's mean natural runoff depth and baseflow index at 1-km resolution. The procedure involved three key steps: (a) compilation and preprocessing of predictands and predictor variables from these 1,056 medium-to-small basins, (b) model development using four machine learning algorithms, (c) cross-scale model validation, best model selection, and grid-scale prediction.

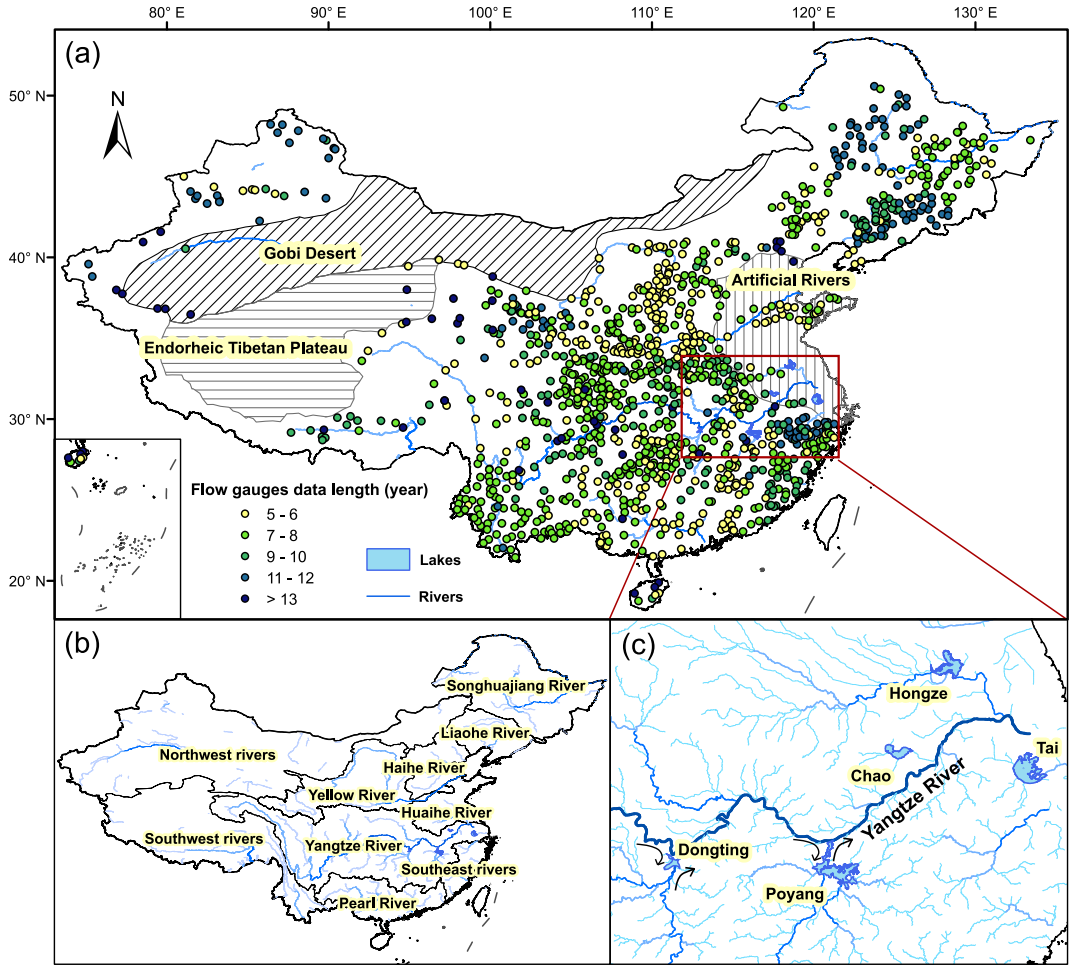

Figure 2. (a) Distribution of 1,225 streamflow gauges used in this study. Line patches represent three large ungauged zones: Gobi Desert, endorheic Tibetan Plateau, and artificial river region in the North China Plain. (b) China's major river basins. (c) China's largest river-lake systems, where black arrows represent bi-directional flux between Poyang Lake, Dongting Lake and the Yangtze River.

3.3.1. Data Preparation

We prepared basin-scale predictands and predictors for machine-learning regression of mean runoff depth and baseflow index using gauged flow data from 1962 to 1979. For each gauged basin, the mean runoff depth was derived by normalizing long-term daily streamflow records with catchment area. Baseflow index was computed as the ratio of mean baseflow depth to mean runoff depth, where mean baseflow depth was separated from long-term daily streamflow series using the HYSEP sliding-interval method (Eckhardt, 2008; Sloto & Crouse, 1996), because it is simple, reproducible and widely used in large-sample applications. Given that strong heterogeneity within very large basins can make basin-mean predictors less representative and reduce predictability, basins with drainage areas greater than 15,000 km^2 were removed. This 15,000 km^2 threshold is pragmatic, as multi-basin studies often use an upper area threshold of about 10,000–20,000 km^2 (Beck et al., 2015; Winter et al., 2024), and a systematic sensitivity test is planned in future work. After screening, a total of 1,056 gauges were retained for machine learning model training and testing. For each gauged basin, we further tested if the observed flow data length is sufficient to represent the average climatology. Specifically, we compared the basin-mean precipitation averaged over the years with available streamflow observations at each gauge with the basin-mean precipitation averaged over the full 1962–1979 period. We found that they differ by less than 10% for most basins (Figure S1 in Supporting Information S1), suggesting that the available streamflow records are generally sufficient to provide consistent estimates of long-term mean conditions across gauges.

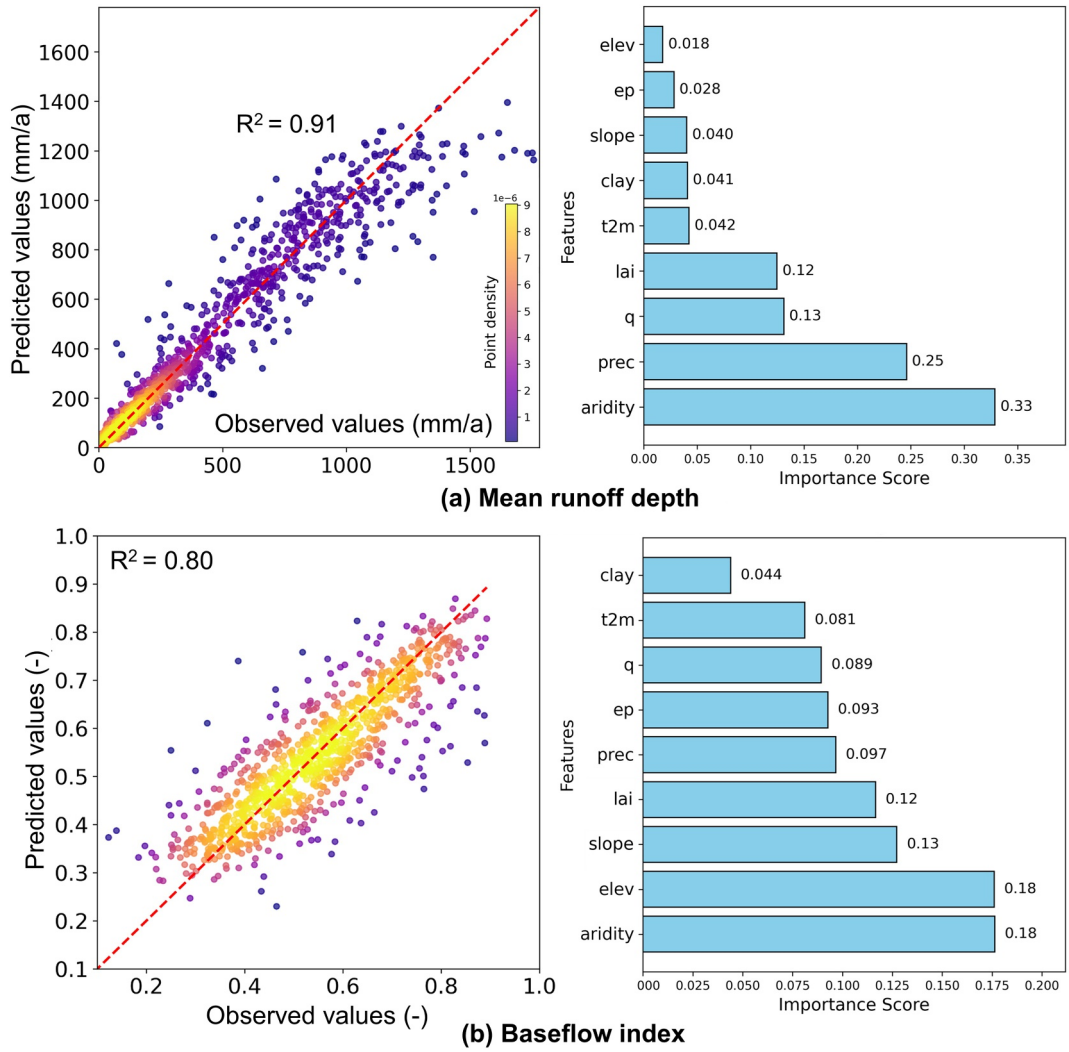

Figure 3. Scatter plots of RF-predicted (a) mean annual runoff depth and (b) baseflow index versus observations for all gauges, and the feature importance.

Following Beck et al. (2015) and Yan et al. (2019), nine predictor attributes of each gauged basin were selected for runoff depth and baseflow index prediction, which were grouped into three categories: (a) climate (mean precipitation, mean temperature, aridity index, mean specific humidity, mean potential evapotranspiration); (b) topography (elevation, slope); and (c) soil and vegetation properties (clay fraction, leaf area index) (Rogger et al., 2017). Climate predictors were derived from CHM_PRE V2 and ERA5-Land data set as stated in Section 2.2, except that the mean potential evapotranspiration is derived from the GLEAM data set. Topography and soil properties are derived from MERIT Hydro and HWSD. Leaf area index (LAI) is taken from the 5-year averaged data (2001–2005) from Global 1-km Land Surface Parameter Data set (Li et al., 2024), which integrates multi-source satellite observations and model inversion to provide high-resolution data for Earth system modeling. For each predictor, we first prepared a consistent 1-km gridded layer, and then aggregated grid values to basin average values.

To understand the correlations among predictors, we performed a Pearson correlation analysis for each pair of the basin attributes used in the ML models and found that some pairs are strongly correlated; for example, specific humidity and precipitation have a PCC of 0.94 (see Figure S2 in Supporting Information S1). However, as Pearson correlation reflects linear dependence only, a predictor may still contribute incremental information through non-linear relationships captured by the ML models.

3.3.2. Machine Learning Model Development

We compared four machine learning algorithms, that is, Support Vector Regression (SVR), Random Forests (RF), Gradient Boosted Decision Trees (GBDT), and k-Nearest Neighbors (KNN). They represent commonly used regression families in hydrology, with SVR as a kernel-based method that can capture non-linear relationships, RF and GBDT as tree-based ensemble methods that handle non-linearity and predictor interactions, and KNN as a simple non-parametric baseline based on similarity. These models are aimed to establish statistical relationships between basin-averaged annual runoff depth/baseflow index, and the corresponding basin attributes in Section 3.3.1 (Breiman, 2001; De'ath & Fabricius, 2000; Friedman, 2002). We trained two parallel sets of the above four models: one with mean runoff depth as the target and another with baseflow index as the target.

We train models with the leave-one-site-out (LOSO) cross-validation approach, where each gauged basin was excluded in turn as the test set while the model was trained on the remaining gauged basins. Compared with a single random 70–30 (or 80–20) split, it avoids sensitivity to one particular partition and provides a more stable estimate of out-of-sample performance. With the test gauged basin excluded, hyperparameters were tuned using randomized search with 30 iterations combined with five-fold cross-validation. The search was guided by minimizing mean squared error, and the best configuration was retained for subsequent evaluation. For RF, optimized hyperparameters include the number of trees, maximum depth, minimum samples per split and leaf, and feature selection strategy. In terms of the mean runoff depth, a square root transformation was applied to the original data prior to training to address its high skewness (Beck et al., 2016). The baseflow index, on the other hand, generally follows uniform distribution and was used for training in its original form without any additional preprocessing.

3.3.3. Machine Learning Model Validation, Selection and Prediction

For each of the four algorithms (Section 3.3.2), we quantified predictive performance using the coefficient of determination (R^2) across the gauged basins. For each target (runoff depth and baseflow index), we selected the best-performing model out of four according to R^2, and used it to produce the 1-km runoff-depth and baseflow-index maps of China, respectively. First, we prepared national 1-km maps of the climate, topography, soil, and vegetation attributes on a common grid, projection, and land mask, so that each 1-km grid cell has a complete set of predictor values. We then applied the selected ML model to each grid cell using these 1-km gridded predictors to generate runoff depth and baseflow index at each 1-km grid cell. These 1-km fields serve as high-resolution calibration targets for deriving key model parameters of CLHMS, as described in the next section.

As machine learning models are trained on basin-aggregated signatures but applied at the grid scale for predictions, there remains a concern whether this basin-to-grid transfer introduces a cross-scale inconsistency problem. To examine this, we conducted an additional check by aggregating the predicted 1-km runoff-depth field over each gauge's upstream area and comparing the resulting basin-mean values with the corresponding observations (Section 4.2; Figure 4). We also use the basin area-PBIAS plot as a simple check because in case of a scale issue, smaller upstream nested basins would tend to show much larger bias than larger downstream basins.

3.4. Calibrating CLHMSv2 Model Parameters at Grid Scales Using the Derived Maps

CLHMSv2 model calibration is focused on three sensitive parameters that strongly influence runoff generation, groundwater exchange, and flow routing (Dong et al., 2022). The first is the runoff generation parameter *dcor* in the land-surface scheme, which governs the partitioning between infiltration and direct surface runoff and constrains flood volumes and total runoff depth. The second is the baseflow parameter C, which is introduced as a multiplicative correction factor that adjusts the effective hydraulic conductivity controlling water exchange between the bottom soil layer and the groundwater layer. The third is the hydraulic roughness (*rough*) in the hydrodynamic module, which controls rate of flow routing and floodplain retention.

We calibrated the runoff generation parameter *dcor* directly at 1-km grid scale using the generated runoff depth map and baseflow index parameter C using the generated baseflow index map (Section 3.3), with the objective of minimizing the differences between simulated and mapped mean runoff depth and baseflow index, respectively. Note that we do not use a single, fixed map for calibration. Instead, for each gauge, we regenerate maps in a leave-one-site-out (LOSO) manner by excluding that gauge and its upstream gauges from the training set. These LOSO-derived maps are then used to calibrate grid parameters for that gauge's catchment. This ensures that no observed

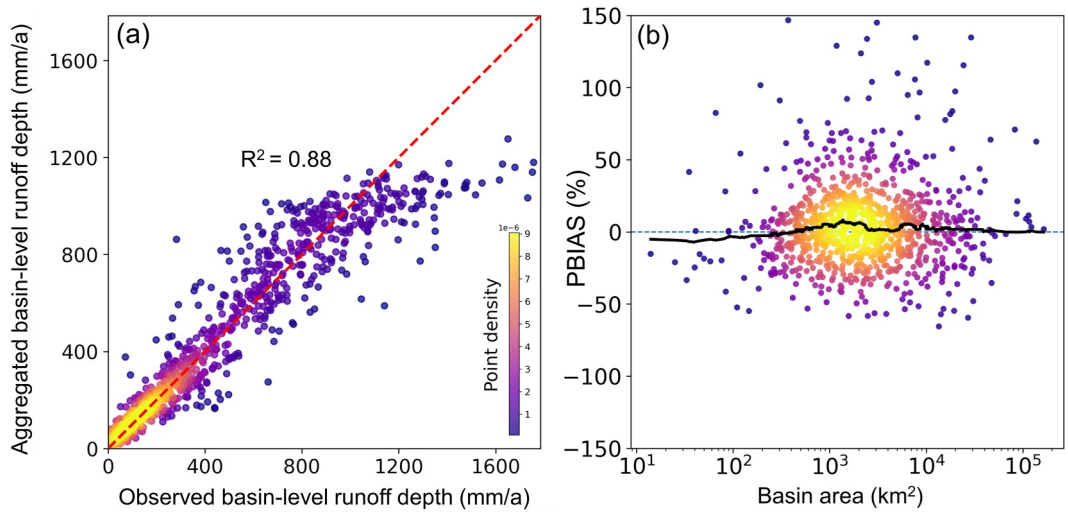

Figure 4. (a) Mean basin-level runoff depths observed from gauges versus those aggregated from grid-scale RF predictions; (b) the percentage bias of aggregated mean runoff depth versus basin area, with black curve being the running median. Scatter colors represent the local concentration of samples estimated by the log-transformed 2-D Gaussian kernel density, with brighter colors indicating denser clusters.

information leaks into the maps used for its own validation. Glacier and lake grid cells are excluded from calibration and are assigned default values, because catchments with significant lake or glacier coverage are limited, and applying model to these grid cells could introduce unpredictable errors. This grid-based calibration provides two major benefits: (a) it reflects local climatic-physiographic controls on runoff and baseflow better than basin-averaged values, and (b) it avoids the scale mismatch that would otherwise arise from applying lumped parameters across heterogeneous landscapes.

Due to lack of runoff-like grid-based characteristics, *rough* was calibrated at the basin scale using daily hydrographs from gauged catchments, with the objective of maximizing the Pearson's correlation coefficient (PCC) between observed and simulated streamflow. PCC is selected as the objective because roughness mainly controls routing speed and timing. Specifically, we divided the gauge data into calibration and validation periods, using the first half for calibration and the second half for validation, with calibration carried out in an upstream-to-downstream sequence. Following Beck et al. (2016), the *rough* parameter was transferred to ungauged areas via physically based similarity regionalization (PSR) using climate, topography, soil, and vegetation predictors (Section 3.3.1). In this study, we did not perform an explicit cross-validation for the PSR, because the approach has already been systematically evaluated and demonstrated as a reliable method for ungauged basin prediction in previous global and continental-scale studies (Bock et al., 2016; Feigl et al., 2022; Pagliero et al., 2019; Song et al., 2022). Moreover, the ungauged regions in our domain are either densely engineered plains or sparsely inhabited deserts and endorheic basins, where the flow routing rate is of limited hydrological relevance.

3.5. Evaluation Metrics of Model Performance

We evaluate model skill using Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE), Pearson correlation coefficient (PCC), and percentage bias (PBIAS). The KGE used here is the original version that includes correlation, mean bias, and variability terms (Gupta et al., 2009). NSE and KGE summarize overall agreement between simulated and observed time series, where higher values indicate better performance and 1 is the optimal value. NSE < 0 and KGE < -0.41 implies that simulations are worse than using the observed mean as a predictor (Knoben et al., 2019). PCC measures flow timing and co-variability, ranging from -1 to 1, with values closer to 1 indicating stronger agreement. PBIAS quantifies systematic overestimation or underestimation of streamflow amount, with 0 as the optimal value and positive values indicating overestimation. In addition, to assess peak-flow behavior, we report two high-flow error measures: R1, the relative error in the mean of the top 1% of daily flows, and R10, the relative error in the mean of the top 10% of daily flows. For both R1 and R10, 0 is the optimal value, positive values indicate overestimation of high flows, and negative values indicate underestimation. We also report the median relative error of the annual maximum 3-day flow in Figure S3 of Supporting Information S1.

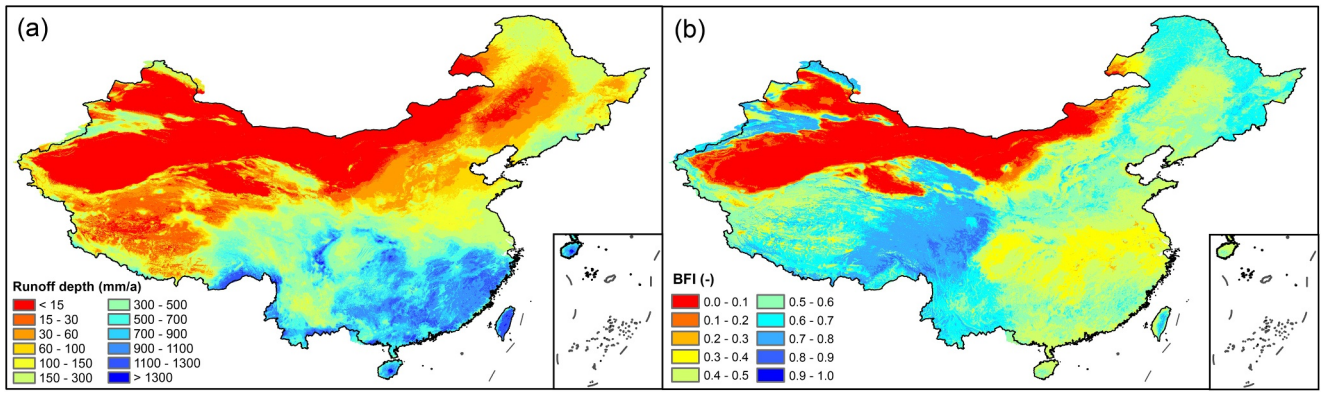

Figure 5. (a) Mean runoff depth (mm/a) and (b) baseflow index estimates (−) of the RF model generated based on flow data from the period of 1962–1979.

4. Results

4.1. Evaluation of Machine Learning Models for Runoff and Baseflow Reconstruction

To derive a 1-km gridded map of runoff depth/baseflow index across China, we first evaluated the performance of four machine learning models by comparing the simulated runoff depth/baseflow index with observations for 1,056 basins nationwide.

Figure 3a shows the observed runoff depth against those predicted by RF, which achieves the highest coefficient of determination (R^2) of 0.91 among the four models. This is followed closely by Gradient Boosted Decision Trees (GBDT; $R^2 = 0.90$) and Support Vector Regression (SVR) ($R^2 = 0.90$); the k-Nearest Neighbors performed least effectively (KNN, $R^2 = 0.89$) (not shown). For RF, results with R^2 of 0.91 are comparable to prior large-sample studies; for example, Beck et al. (2016) used ∼14,000 stations globally to estimate the mean streamflow and reported an R^2 of 0.87. A spatial evaluation of RF predictions is presented in Figure S4 of Supporting Information S1. For mean runoff depth, larger biases occur more frequently in semi-arid northern China, whereas southern China generally sees less biases with most gauges within ±25%. This spatial pattern is consistent with the greater intermittency and event-driven runoff generation in arid and semi-arid regions, which is harder to capture using basin-mean attributes.

We calculate the impurity-based feature importance to interpret the RF model. Aridity serves as the most influential (aridity = 0.33), followed by precipitation (prec = 0.25) and specific humidity (q = 0.13). Leaf area index (lai = 0.12), air temperature (t2m = 0.042), and clay fractions (clay = 0.041) rank moderately, while topographic variables (slope = 0.04, elev = 0.018) have lower importance. Overall, climate variables (aridity, precipitation, specific humidity) account for over 60% of importance, affirming their controlling role in runoff generation across China's diverse hydroclimates.

Figure 3b shows the observed baseflow index against those predicted by RF, which achieves the highest coefficient of determination (R^2) of 0.80 among the four models. This is followed closely by Gradient Boosted Decision Trees, Support Vector Regression, and k-Nearest Neighbors, with R^2 of 0.78–0.79 (not shown). Spatially, the PBIAS of most gauges fall within ±25%, with |PBIAS| > 25% mainly concentrated in the middle Yellow River, parts of the Yangtze River Basin, and some southeast coastal basins (Figure S4 in Supporting Information S1). A possible explanation is that baseflow index can be more sensitive to fine-scale heterogeneity in soil texture and subsurface properties (Bloomfield et al., 2009), whereas current available predictor layers at 1-km resolution may not fully represent this heterogeneity, resulting in BFI predictions with larger uncertainties. In terms of the feature importance for RF, aridity serves as the most influential (0.18), followed by elevation (0.18) and slope (0.13). Leaf area index (0.12), precipitation (0.097), potential evapotranspiration (0.093), specific humidity (0.089) and air temperature (0.081) rank moderately, while clay fractions (0.044) have lowest importance. Overall, topographic and vegetation variables (elevation, slope, and LAI) account for over 40% of importance, suggesting local vegetation and terrain have a strong constraint on the baseflow characteristics.

While the RF model demonstrates strong skills at reproducing basin-level streamflow characteristics, its grid-scale maps are only useful for calibration if basin characteristics can be reliably recovered by spatially

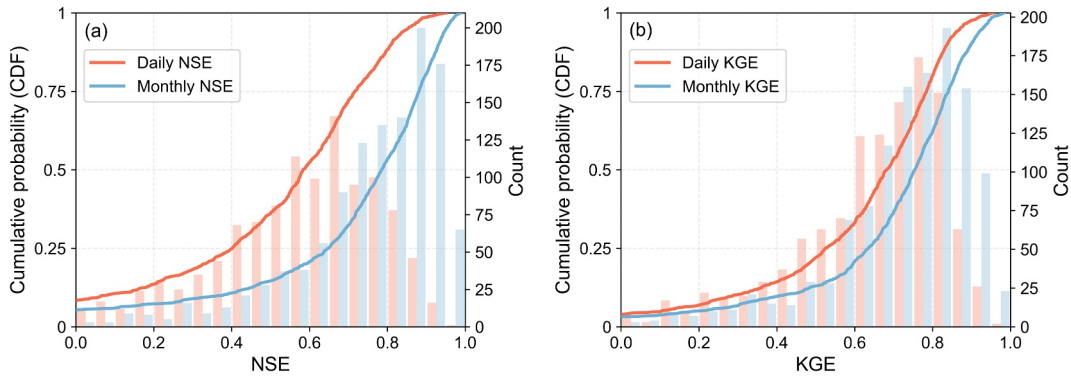

Figure 6. The cumulative probability and histogram of daily and monthly (a) NSE, and (b) KGE of simulated streamflow for all gauges.

aggregating grid predictions. To test this, we averaged the 1-km grid-cell runoff depths across the upstream area of each gauge, and compared the resulting basin-mean values with observations (Figure 4a). The agreement is comparably high ($R^2 = 0.88$), although with a tendency to underestimate basin-averaged runoff depth in some of the wettest basins. This indicates that the grid-scale predictions can generally preserve basin-scale runoff depth when aggregated. We further examined potential scale-conversion effects by plotting the percentage bias (PBIAS) of the aggregated runoff depth against basin area (log scale; Figure 4b). The median PBIAS is close to zero, and more than 80% of basins have |PBIAS| < 25%, consistent with Moriasi et al. (2007) classifying |PBIAS| ≤ 25% as satisfactory for water-balance simulations. Moreover, the running median (black curve) stays near zero across multiple orders of magnitude in basin area. This weak dependence of PBIAS on basin size suggests that neither the basin-to-grid prediction nor the grid-to-basin aggregation introduces a significant scale bias. Together, these results support the use of RF-generated grid maps for grid-scale calibration based on basin-scale data.

4.2. China's 1-km Gridded Runoff Depth and Baseflow Index Map

The two RF models were selected and applied to produce the 1-km gridded runoff depth and baseflow index maps generated based on flow data from the period of 1962–1979 (Figure 5), respectively, for their highest R^2 among the four models. The generated runoff map exhibits pronounced spatial distinctions, with high runoff depths (>1,500 mm/a) in the humid southeastern basins (e.g., Yangtze, Pearl), moderate values (500–1,000 mm/a) in the central and northeastern regions (e.g., Huai, Songhua), and low depths (<200 mm/a) in arid northwestern areas (e.g., Tarim, Gobi Desert) and the Tibetan Plateau. These spatial gradients align with China's monsoonal climatic characteristics, as precipitation declines sharply from the humid southeast toward the arid northwest (Figure S5 in Supporting Information S1).

The spatial distribution of the baseflow index (BFI; Figure 5b) exhibits substantial variability across China. Areas with very high BFI values (0.8–1.0) correspond mainly to regions with steep terrain. The Tian Shan Mountains and the eastern Tibetan Plateau are typical examples, as their rugged terrain and permeable fractures enhance infiltration and subsurface connectivity, allowing groundwater to sustain river flow even during dry periods. Similarly, the Southeast Hills across the Southeast China show relatively high BFI values (0.6–0.8). Arid inland basins, such as the Tarim Basin and Gobi Desert, along with interior Tibetan Plateau, display very low BFI values (0–0.3). In these regions, scarce precipitation and strong evaporative demand restrict groundwater recharge and favor rapid surface runoff following rainfall events. Intermediate BFI values (0.4–0.7) dominate the central and northeastern transition zones, where moderate slopes and mixed climatic conditions yield a more balanced partitioning between surface and subsurface flow.

4.3. Model Performance of Streamflow Simulations

We evaluate the performance of simulated streamflow for each gauge using metrics of NSE, KGE, PCC, PBIAS, R1, and R10, and summarize these metrics at daily and monthly scales. Because the model operates at a daily time

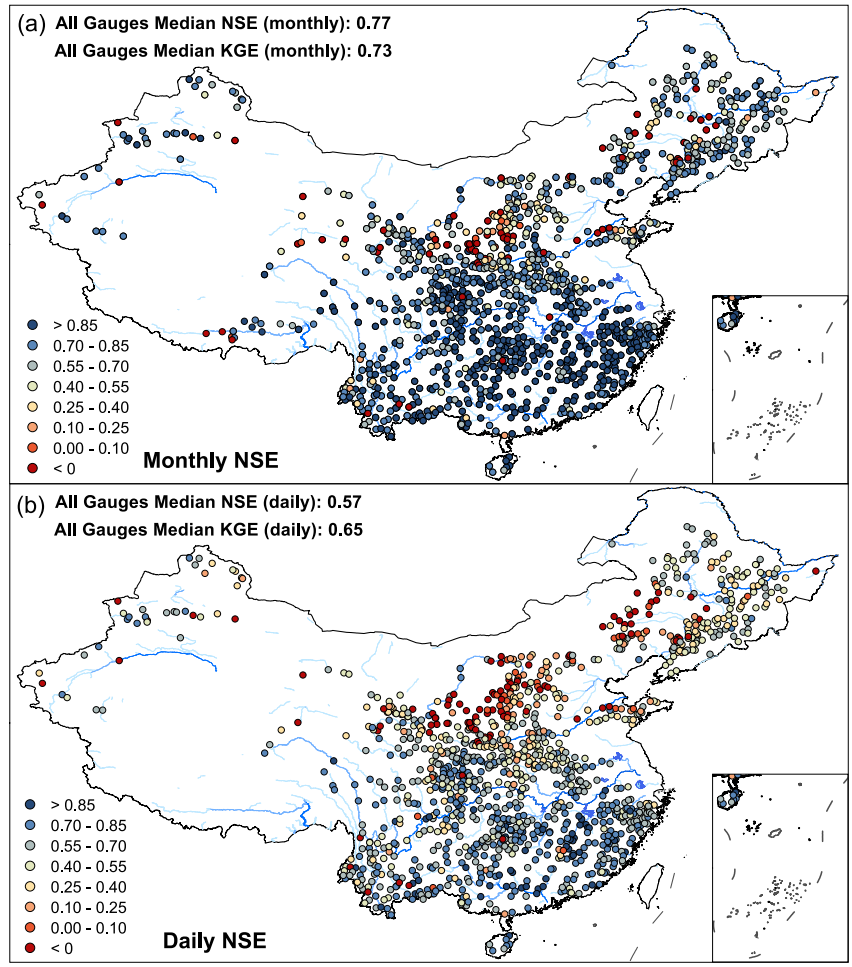

Figure 7. The Nash-Sutcliffe efficiency (NSE) of simulated streamflow at (a) monthly and (b) daily scales for 1,225 streamflow gauges.

step, the daily evaluation provides the most stringent assessment of model skill; the monthly metrics are reported as complementary indicators of aggregated bias and variability.

Figure 6 shows cumulative probability and histogram of NSE and KGE for all gauges. In both panels, the monthly curve is shifted to the right of the daily curve over the entire 0–1 range, which means monthly simulations achieve higher NSE/KGE at any chosen percentile. Compared to daily curves, the monthly cumulative frequency also rises more slowly at low NSE/KGE values and much more steeply near the upper tail, indicating fewer low-skill gauges and a greater concentration of high-skill gauges.

Figure 7 presents the spatial distribution of model performance across China. Specifically, the NSE/KGE values reveal notable temporal and regional variations. On a daily timescale, the median NSE and KGE values are 0.57 and 0.65, respectively, with high-performance gauges (NSE > 0.5) predominantly clustered in the humid eastern and southeastern river basins, such as the Yangtze and Pearl River Basins. In contrast, lower NSE values (<0) are observed in arid northwestern and northeast regions potentially attributable to challenges in simulating intensive glacier and snowmelt dynamics in cold environments. Gauges with low skills are also observed in the Loess Plateau regions of the middle Yellow River Basin, possibly due to the unique and complex runoff characteristics of the loess, which are difficult to accurately capture. In terms of the monthly timescale, the median NSE and KGE are 0.77 and 0.73, both higher than daily scores. This is an expected result, because temporal aggregation smooths sub-monthly timing errors, random forcing errors, and high-frequency runoff variability, thereby increasing the correlation and often improving the variability component of KGE (Gebrechorkos et al., 2024; Lin et al., 2019).

10.1029/2025WR042606

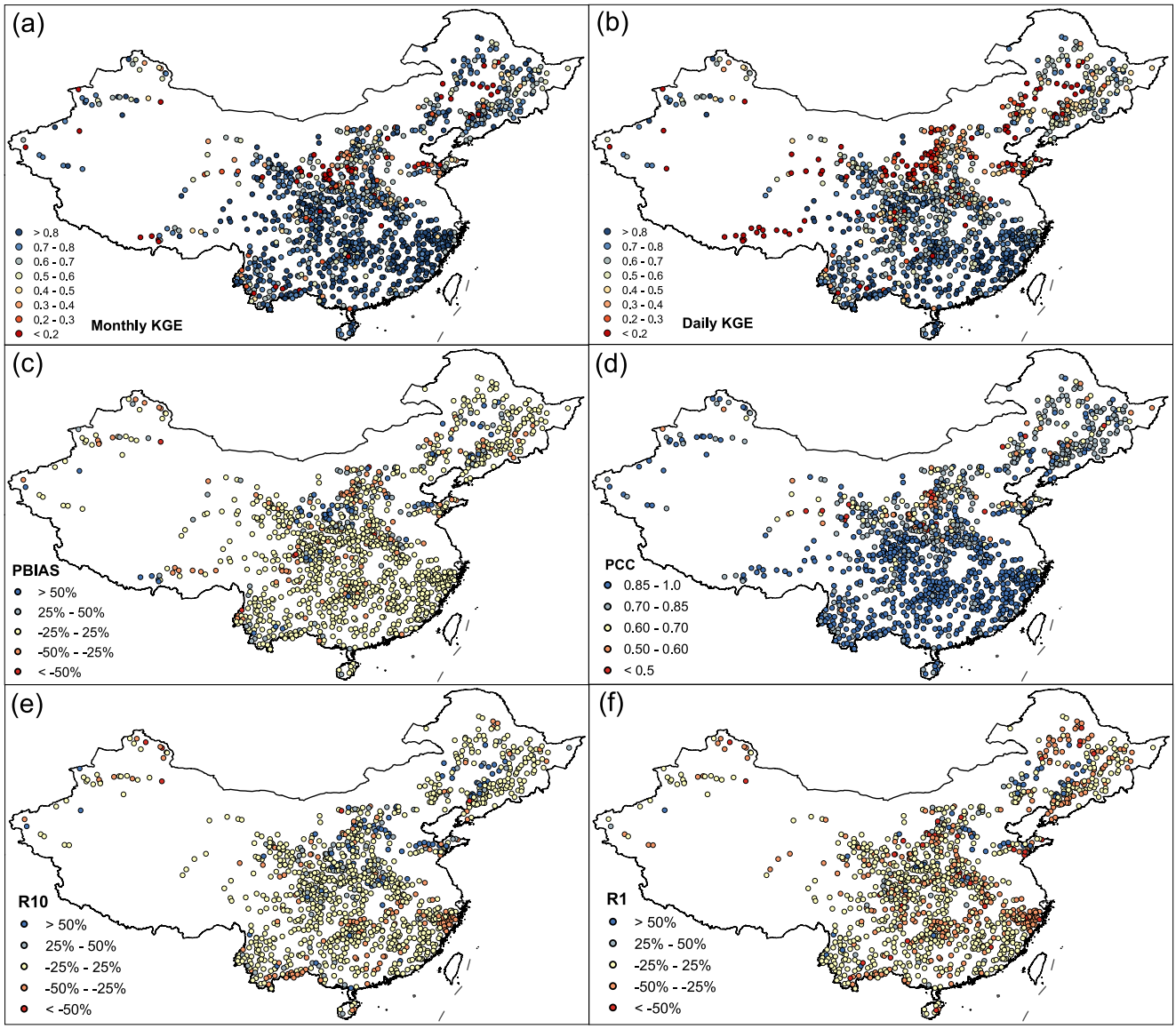

Figure 8. The (a) daily KGE, (b) monthly KGE, (c) PBIAS, (d) PCC, (e) R10, and (f) R1 of simulated streamflow for all gauges.

Figure 8 depicts the distribution of evaluation metrics other than NSE across China, and we analyze these metrics from a basin point of view. For the Yangtze River Basin, the streamflow skills are generally high. Most gauges show blue classes (>0.5) for daily KGE and even darker blues (>0.85) for monthly KGE, and PCC is high and spatially coherent along the middle and lower mainstream and major tributaries. PBIAS is mostly within ±25% range, suggesting acceptable overall bias according to Moriasi et al. (2007). We also report the variability term α of KGE in Figure S6 in Supporting Information S1, with α mostly within 0.8–1.2 range, suggesting that the model generally reproduces the observed flow variability reasonably well. For the extreme-flood metrics, R1 is occasionally underestimated, whereas R10 is generally within the ±25% range, with sporadic overestimation for the Yangtze River Basin.

The Pearl River Basin shows some of the strongest skill at the monthly scale, with widespread high KGE and high PCC. Daily KGE is above 0.5 in general. PBIAS are generally within the ±25 range. As in the Yangtze, R1 is often negative while R10 is mostly within the ±25% range, implying that the model captures the majority of high flows but underpredicts the sharpest flood peaks. Likely causes include the daily temporal resolution of precipitation forcing that damps sub-daily rainfall peaks, spatial smoothing when interpolating precipitation from the

10.1029/2025WR042606

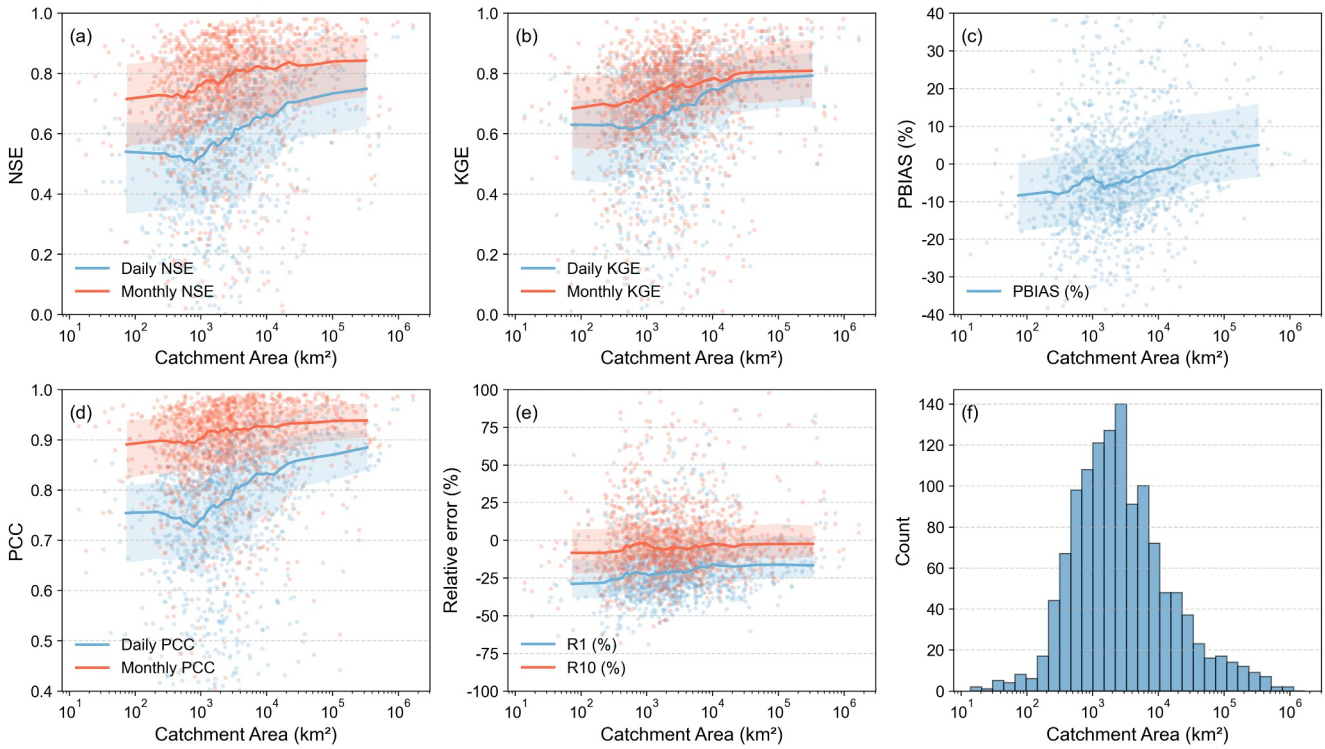

Figure 9. The scatter plot of (a) NSE, (b) KGE, (c) PBIAS, (d) PCC, and (e) R1 and R10 of simulated streamflow across the (f) distribution of catchment area. The colored curve represents running median, and the shaded areas represent the interquartile range.

10-km CHM grid to the 1-km model grid, and limited flexibility of the runoff parameterization to capture nonlinear runoff amplification during extreme events.

Performances across the Hai, Huai, and Yellow River Basins become more varied. Daily KGE and PCC are comparably low in several reaches, especially in the Loess Plateau. PBIAS is sporadically over 25% and even 50% for some gauges. α is occasionally larger than 1.2 and can reach over 1.5 across these basins, suggesting that the simulated variability is inflated relative to observations. This points to challenges in representing event runoff production in semi-arid and arid basins, which we discuss in detail in Section 5.3. For the extreme-flood metrics, R10 is overestimated in some of the gauges, whereas R1 exhibits occasional underestimation across the Hai, Huai, and Yellow River Basins.

In the Songhua and Liao River basins of Northeast China, daily performance is generally moderate and improves at the monthly scale. PCC is typically above 0.7, although it is less spatially uniform than in humid southern China. PBIAS is commonly within ±25% and is often slightly negative. For extreme-flood metrics, R1 shows a mild negative bias, whereas R10 is generally within ±25% or mildly positive. Mixed snow-rain regimes likely contribute to the relatively lower skill in this region, as uncertainties in precipitation phase and melt timing pose added challenges for model parameterization.

Northwest interior basins are mostly arid and strongly influenced by glacier and snow dynamics, which show a wide range of simulation skills. Daily and monthly KGE include more mid/low classes with values less than 0.5, and monthly KGE yields only modest improvement relative to the daily values. PBIAS are often negative, suggesting underestimation of mean flow.

Figure 9 depicts the scatter plots of evaluation metrics for gauges versus catchment areas. In general, the NSE increases with basin size and is higher at the monthly scale. At daily scales, for example, NSE medians rise from 0.52 for basin area <500 km^2 to 0.70 for basin area >10,000 km^2, and the interquartile range from about 0.28 for basin area <500 km^2 to about 0.31 for basin area >10,000 km^2. Monthly averaging lifts performance across all area classes. Monthly medians increase from 0.72 for basin area <500 km^2 to 0.83 for basin area >10,000 km^2, respectively, which is an improvement of ~0.1–0.2 over the daily medians, with the largest gains in smaller and

mid-sized basins. The patterns of KGE are similar to those of NSE, but with slightly lower magnitudes for monthly scales and with slightly higher magnitudes for daily scales. Daily medians rise from 0.58 to 0.74 from the smallest to largest basins, and monthly medians from 0.66 to 0.80, with monthly values higher across all basin sizes.

Relative error shows a tendency toward underestimation in basins smaller than 10,000 km^2 and a shift toward overestimation in basins larger than 10,000 km^2, while the median bias remains within ±20% across all basin sizes. Similarly, the interquartile range generally remains stable at 20% across all basin sizes, suggesting there is no considerable bias scale-dependency. The PCC improves with the increase of basin size and becomes more uniform in the largest basins. Specifically, PCC medians move from 0.75 in the smallest basins to 0.86 in the largest basins.

In terms of the error of extreme flows, the model performance differs between the very largest floods (R1) and the broader high-flow condition (R10). Median R1 is negative across most basin sizes, and the median bias changes from around −25% in the smallest basins to −15% in the largest, suggesting larger basins see more accurate extreme flood simulations. Similarly, median R10 is negative on average in all basin sizes and generally becomes less biased with basin size, as the median drops from 11% in the smallest basins to −2% in the largest basins.

4.4. Model Performance of Water Level Simulations

Figure 10 evaluates the model skill over coupled river-lake systems by presenting two-dimensional flow fields and time series of simulated versus in situ lake water levels at representative gauges for Poyang (Xingzi), Dongting (Chenglingji), Tai (Dapukou), Hongze (Jiangba), and Chao (Hualinzhen) Lakes. The flow maps reproduce the channelized inflow/outflow, and multi-directional spreading within water bodies.

Overall, the modeling system can well resolve lake hydrodynamics with strong backwater effects or multidirectional water flux for Poyang, Dongting, Tai, and Chao Lakes. At Poyang Lake (Xingzi) and Dongting Lake (Chenglingji), the model exhibits high PCC (0.98), NSE (0.96), and KGE (0.96) of water level, suggesting the model can well capture seasonal amplitude and interannual variability of lake levels. This result indicates that the rapidly changing bi-directional flux between Yangtze mainstream channels and the Poyang and Dongting Lakes is well resolved. For Tai Lake (Dapukou) and Chao Lake (Hualinzhen), the model accuracy of lake water level is also satisfactory (PCC = 0.83–0.87; NSE = 0.57–0.73; KGE = 0.46–0.66). The seasonal cycle is generally well reproduced, though several peaks are underestimated and water level recessions show small phase lags. A source of errors is likely due to human interventions in the lakeside metropolises of Suzhou (Tai Lake) and Hefei City (Chao Lake). On the contrary, Hongze Lake (Jiangba) shows low accuracies (PCC = 0.30; NSE < 0; KGE = 0.07), as the simulations overreact to short-term variability and misrepresent multiple peaks. We attribute this degradation primarily to anthropogenic regulation, for example, reservoirs, sluices, and pumps built on the lake since early 1950s, and during the validation period the hydraulic dynamics has been strongly modulated.

5. Discussion

5.1. Derivation and Applications of Grid-Scale Hydrologic Signatures

An effective calibration approach of distributed hydrologic models in continental applications is to perform parameter tuning at gauge levels in a lumped way (i.e., a single parameter set applied to all cells upstream of a gauge), followed by regionalization to ungauged areas (Bock et al., 2016; Feigl et al., 2022; Pagliero et al., 2019; Samaniego et al., 2010). More recently, the Large-Sample Emulator (LSE) provides a scalable pathway to jointly calibrate and regionalize process-based land-hydrology models and has shown consistent runoff improvements (Farahani et al., 2025; Tang et al., 2025). Differentiable modeling has also been developed to estimate hydrologic model parameters with gradient-based optimization (Song et al., 2022). Here we explore a complementary path to leverage spatial parameter constraints available from modern data sets, as recent studies have advocated incorporating spatially continuous signatures (soil moisture, ET, TWS, flow characteristics) into calibration to improve parameter tuning and reduce equifinality (Lin et al., 2019; López et al., 2017; Mei et al., 2023; Rajib et al., 2018; Xie et al., 2021).

Built on this idea, our approach follows the general logic of prior global efforts that learn basin-level attribute-signature relationships and then apply them at grid scales to produce gridded signature maps (Beck et al., 2015). A key question, however, is whether relationships learned at basin scale can transfer coherently to grid scale without

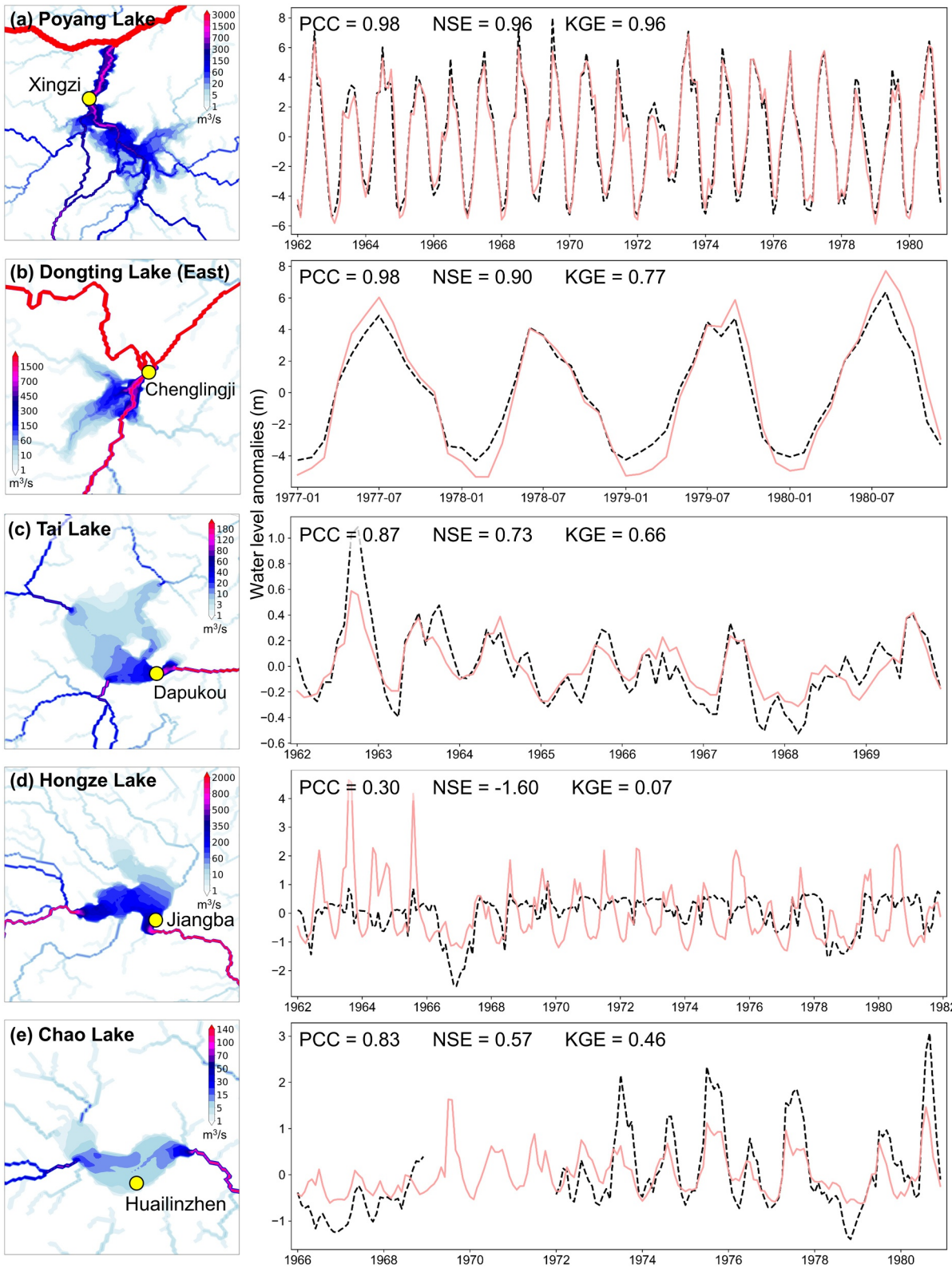

Figure 10. Simulated mean 2D flow fields across the river-lake systems of (a) Poyang Lake, (b) Dongting Lake, (c) Tai Lake, (d) Hongze Lake, and (e) Chao Lake; and the simulated monthly lake water level versus in situ observations at Xingzi, Chenglingji, Dapukou, Jiangba, and Hualinzhen gauges over the same period.

a scale issue, and vice versa. We investigated this cross-scale consistency in multiple ways. First, Figure 4 validates the gridded signature maps after aggregation back to basin scale: when we average grid-cell runoff depths over each gauge basin, the resulting basin means closely match observations ($R^2 = 0.90$), indicating that the grid-scale products preserve basin-level magnitudes. Moreover, the weak dependence of PBIAS on basin size suggests that neither the basin-to-grid prediction nor the grid-to-basin aggregation introduces a systematic scale issue. To further investigate the equifinality in our streamflow simulations, we plot the PBIAS at 1,225 gauges versus catchment area as an independent cross-scale indicator. We hypothesize that equifinality in the 1-km maps will be evident if smaller upstream sub-basins consistently exhibit larger PBIAS values than larger downstream basins in our simulations. We do not observe such a pattern, suggesting that upstream and downstream performance is broadly consistent and that cross-scale equifinality has negligible impacts on the conclusions in our current study. A fully quantitative 1-km equifinality assessment, however, remains challenging due to the lack of 1-km in situ observations for the mapped hydrologic signatures.

Large-sample studies have shown that a small set of climatic and physiographic attributes can explain a substantial fraction of the spatial variability in streamflow signatures such as mean runoff, baseflow index, and flow quantiles (Beck et al., 2015; Gudmundsson & Seneviratne, 2015; Hobeichi et al., 2019; Povak et al., 2014). In consistent with these studies, our machine-learning reconstruction indicates that climate variables account for more than 60% of the predictive importance for runoff depth, whereas terrain and vegetation together account for more than 40% of the importance for BFI (Section 4.1; Figure 3). The dominance of climate controls matches hydrologic understanding that the mean runoff magnitude is primarily constrained by water and energy availability across China's hydroclimatic gradients. For baseflow index, the shift in importance toward elevation, slope and LAI is also explainable, as topography and vegetation can influence subsurface storage, drainage efficiency, and the partitioning between quick flow and groundwater runoff (Price, 2011).

5.2. Comparison With the State-Of-The-Art Natural Runoff Data Set CNRD v1.0

We conducted a quantitative evaluation of CHASE versus China Natural Runoff Data set (CNRD) against the same gauge observations at the monthly scale. CNRD v1.0 (Gou et al., 2021) provides 0.25° gridded monthly natural runoff for 1961–2018 produced with the VIC v4.2 model and calibrated against 200 natural or near-natural catchments. This product represents the state-of-the-art for China's national-scale applications, and has been widely used in China's hydrologic studies (Zhan et al., 2025). To ensure a fair comparison given CNRD is a runoff product at 0.25° resolution rather than flow product, we (a) excluded very large basins (area >50,000 km^2) to reduce the influence of routing travel time when converting monthly runoff to monthly flow, and (b) excluded very small basins (area <600 km^2), which are smaller than a typical 0.25° grid cell and therefore not well represented by CNRD runoff conditions. This left 884 gauges with at least 5 years of observations. We then spatially average the CNRD monthly runoff of these basins to derive the monthly flow of 884 gauges, and compared those with CHASE results.

Across these 884 gauges, CHASE shows consistently higher skill than CNRD in terms of KGE, NSE, and PBIAS at the monthly scale (Figure 11). For example, the median NSE is 0.77 for CHASE versus 0.54 for CNRD, the median KGE is 0.74 for CHASE versus 0.45 for CNRD; and the median PBIAS is 11% for CHASE versus 35% for CNRD. The median PCC is similar for CHASE and CNRD, both standing at 0.92. For reference, Moriasi et al. (2007) provide commonly used performance guidelines for hydrologic simulations, such as NSE > 0.5 and |PBIAS| ≤ 25% as indicative of satisfactory performance. Under these criteria, 87% (55%) of gauges meet the NSE threshold for CHASE (CNRD), and 84% (38%) meet the |PBIAS| threshold for CHASE (CNRD). Specifically, CHASE shows widespread higher NSE/KGE (by ∼0.2–0.5) than CNRD particularly across northern China (e.g., Yellow River and Songhuajiang River basins) and in several inland basins in the northwest China (Figures 11a and 11b). The PBIAS comparison indicates that water-balance error is one of the primary drivers of these NSE/KGE differences. CHASE generally exhibits smaller PBIAS than CNRD nationwide, with the largest PBIAS reductions in the Yellow River, Songhuajiang River, and inland northwest basins, where bias of CHASE is often more than 20% closer to zero than CNRD. In contrast, differences in PCC are relatively smaller, with most gauges falling within ±0.05. A notable exception is the northwest China, where CHASE exhibits considerably higher PCC (by >0.1) than CNRD.

The spatial patterns of these improvements can be explained by following reasons. First, CHASE is constrained by a much higher resolution (1 km) and much denser gauge network (1,225 gauges) in China, including ∼400

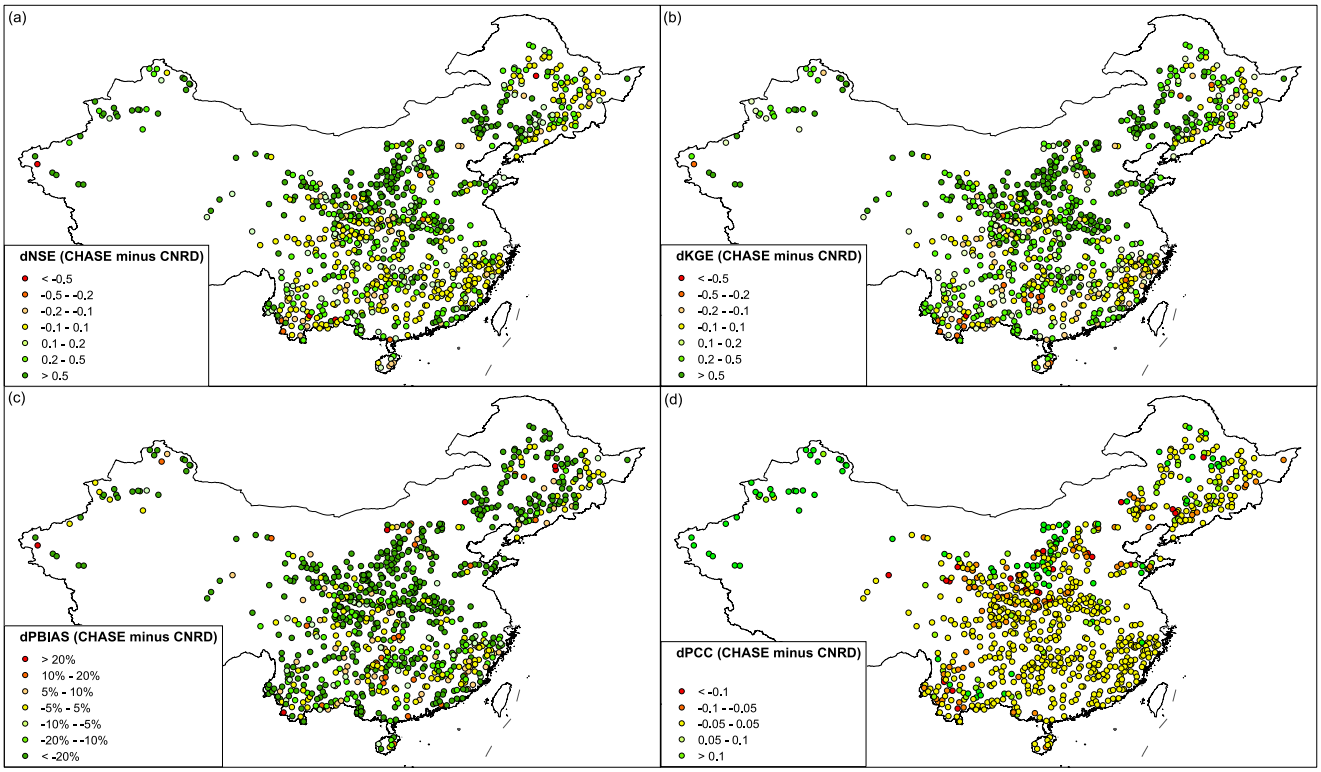

Figure 11. Difference of (a) NSE, (b) KGE, (c) PBIAS, and (d) PCC between CHASE and CNRD for 884 gauges nationwide. Green (red) markers indicate CHASE has a better (worse) performance than CNRD.

medium-to-small basins with an area below 1,000 km^2 (Figure 9f). In contrast, CNRD was calibrated at 0.25° (~25 km) resolution using roughly 200 basins nationwide, with only a few dozen gauges in key northern basins such as the Yellow and Songhuajiang Rivers. Moreover, many of these gauges are located along the mainstreams, which provides weaker constraints on sub-basin heterogeneity and limiting parameter transfer to smaller catchments. This likely contributes to the larger water balance biases of CNRD across the Yellow and Songhuajiang Rivers of northern China relative to CHASE. Second, the improvements of CHASE over CNRD in northwest rivers are also consistent with differences in process representation between CLHMSv2.0 and VIC v4.2. The VIC v4.2 configurations used for CNRD do not explicitly represent glacier accumulation and melt, whereas CLHMSv2.0 includes an energy-based glacier module. This gives CLHMS an advantage as glacier melt can contribute substantially to runoff in high-mountain headwaters of northwest China and strongly influences seasonal runoff volume and timing (Su et al., 2023; Xu et al., 2025).

Overall, this comparison provides a baseline for interpreting the CHASE v1.0 results. The skill achieved here is not only acceptable under commonly used performance guidelines, but is also consistently higher than a widely used China-scale natural runoff product, CNRD v1.0, across 884 gauges nationwide. The largest gains occur in northern and inland arid to semi-arid basins, where large-scale hydrologic modeling and parameter transfer are typically most challenging. In humid southern China, differences are smaller because both approaches already perform relatively well.

5.3. Potential Applications, Caveats and Limitations

To begin with this section, we emphasize that our generated flow data record is not intended to reproduce observed modern regulated flows. Instead, it aims to provide long-term naturalized streamflow driven by climate variability alone, with human influences (e.g., dams, withdrawals, irrigation, and land-use change) excluded as far as possible. This enables applications such as (a) attributing observed streamflow changes to climate versus human activities by comparing naturalized flow records with in situ observed flow records, and (b) assessing

climate-driven streamflow variability and change. Below are several other caveats and limitations for readers' information.

1. A key premise of CHASE v1.0 is that model calibration and evaluation rely on streamflow records that are minimally affected by human regulation. Although we restricted the calibration data to the pre-1980 period and screened near-natural basins, human activities cannot be fully excluded due to the high population density of China. Near-natural basins may still contain residual impacts from small reservoirs, diversions, or irrigation return flows, which can introduce non-climatic signals into the basin signatures, parameter fields, and the resulting naturalized simulations.
2. Model performance is also relatively more limited in the Yellow and Hai River Basins, with the Loess Plateau in the middle Yellow River showing occasional negative NSE at daily and monthly scales. We attribute this to the following reasons. First, parts of these regions exhibit strong non-linear soil behaviors and thus complex runoff response, including soil structure change, crusting and cracking, and wetting-induced collapse in loess soils (Li et al., 2016). Second, under semi-arid to arid climates, runoff generation is more intermittent and event-driven, making simulations more sensitive to small errors in forcing and parameters. Third, flows are generally lower in the Yellow and Hai, and their flow regimes can be more susceptible to occasional human disturbances compared to those in the humid south.
3. Results for smaller catchments ($<$1,000 km^2) should be interpreted more cautiously than for larger basins, as NSE/KGE performance is generally lower at small basins. A likely reason is that the resolution of the meteorological forcing is comparable to, or coarser than, many small basins, which can smooth out mesoscale convective systems. This limitation calls for continued development of higher-resolution meteorological forcing at large scales (Hoch et al., 2023; van Jaarsveld et al., 2025). As an attempt, van Jaarsveld et al. (2025) downscaled coarse-scale forcing by bilinear interpolation to a fine grid and then applying day-of-year correction factors based on high-resolution climatologies.
4. A distinct feature of our model and data set is the inclusion of river-lake 2D flow fields over China's largest freshwater lakes. Across major lakes, simulated water levels typically show high correlation with observations (often PCC $>$ 0.8), suggesting that the model can broadly represent coupled river-lake dynamics at large scales. When interpreting the results, it should be noted that recent studies have reported changes in lake-bed morphology (e.g., from cumulative sand mining) (Yao et al., 2019), which can modify hydrodynamics and stage relationships over decadal time scales, leading to shifted water-level baseline and additional bias in long-term simulations.
5. CLHMSv2 includes major cold-region hydrologic processes, such as snow accumulation/melt, glacier accumulation/melt, and soil freeze-thaw. However, there remains limited representations in the model about how soil thermodynamics translate into groundwater discharge changes under permafrost degradation, as fully coupled representations of hydro-thermal dynamics across surface and subsurface are still challenging and data-limited. This means that, while CHASE v1.0 can be informative for present-day hydrology, its application to climate-change impact studies in permafrost-affected regions (e.g., the Tibetan Plateau) should be interpreted with caution, and would benefit from targeted evaluation using localized observations where available.

6. Conclusions

We developed a 1-km gridded, national land-surface-hydrologic-hydrodynamic modeling framework with machine learning based calibration strategy applied at grid scale. Using flow data from 1,225 gauges nationwide, mean runoff depth and baseflow index maps were generated as spatial calibration targets to constrain runoff generation and subsurface exchange, and 1D/2D hybrid diffusive wave routing was introduced for 2D lake-river network to resolve backwater and bidirectional exchanges. Applied over 1962–2024, the system aims to provide long-term naturalized streamflow driven by climate variability alone, with human influences excluded as far as possible. The resulting data product is named CHinA grid-level natural Streamflow Estimates Version 1.0 (CHASE v1.0).

We highlight several conclusions during model and data development as follows:

1. A small set of climate, topography, soil, and vegetation attributes can explain a large fraction of the nationwide variability in mean runoff depth and baseflow index, but the dominant controls differ between the two

signatures. Runoff depth is mainly climate-driven, whereas baseflow index shows stronger sensitivity to terrain and vegetation-related attributes.

2. The attribute-signature relationships learned by ML models at the basin scale can be transferred to the 1-km grid scale to produce spatially coherent gridded signature maps, which can serve as effective spatial constraints to calibrate runoff-generation and baseflow parameters of CLHMS at grid scale.
3. CLHMSv2 streamflow skill evaluated at 1,225 flow gauges is consistently positive (median NSE = 0.57, KGE = 0.65) at daily scales and improves at monthly scales (NSE = 0.77, KGE = 0.73). Humid, larger basins outperform arid and snow-dominated headwaters, with median daily NSE higher by ∼0.20 in humid basins compared to arid basins and by ∼0.13 in large basins (>10,000 km^2) compared to small (<1,000 km^2) basins. High-flow metrics indicate generally satisfactory performances of extreme flood peaks, albeit with a tendency toward moderate underestimation in several regions. Coupled lake-river systems are generally well reproduced, with temporal water level correlations PCC > 0.8 at Poyang, Dongting, Tai, and Chao Lakes.
4. A comparison of CHASE v1.0 against CNRD v1.0 data set at 884 gauges shows that the median NSE of CHASE is ∼0.23 higher (0.77 vs. 0.54) than that of CNRD, which can be primarily attributed to a much higher resolution and a much denser calibration gauge network for CHASE (1 km; 1,225 gauges) than for CNRD (0.25°; ∼200 gauges).

Future directions include extending the calibration framework to (a) more streamflow percentiles (e.g., 10/90 percentiles) for better capture of flow distribution and (b) spatiotemporally varying parameters within the differentiable framework (Tounsi et al., 2023; Wang et al., 2024). To better represent hydrologic variations in modern times, reservoir operation, irrigation, and water withdrawal could be implemented into the 1-km model based on our previous work (Dong et al., 2023; Hao et al., 2024).

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

The CHASE v1.0 data set, along with the evaluation metrics and gauge information, is deposited at its official website https://hydrodata.cn/chase (CHASE, 2026) and a backup Zenodo link at https://doi.org/10.5281/zenodo.17386507 (Dong, 2026). The generated runoff depth and BFI map is deposited at its official website https://hydrodata.cn/sigmap (SigMap, 2026) and a backup Zenodo link at https://doi.org/10.5281/zenodo.17364737 (Dong, 2025a). Machine learning model training, test and prediction codes, along with the data set evaluation codes, are available in Dong (2025b) at https://doi.org/10.5281/zenodo.18470664. Due to government regulations on the distribution of high-resolution streamflow data, access to CHASE v1.0 is restricted to researchers for research purposes. Access can be obtained via the *Request Access* portal on the above official website (recommended) or the Zenodo repository. Requests are reviewed by the corresponding author in accordance with the government policy.

Acknowledgments
We thank the editor and three reviewers for their insightful comments to help improve this paper. Ningpeng Dong received funding for this study through the National Key Research and Development Program of China (2023YFC3081000), the National Natural Science Foundation of China (42401053), and the Young Elite Scientist Sponsorship Program by CSHE (CSHE-YESS-2026006). Jianhui Wei is supported financially by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) through funding of the KARE_II project (01LR2006D1). This study is also funded by the Research Project of the State Key Laboratory of Water Cycle and Water Security (SKL2024YJZD03).

References

Akpoti, K., Velpuri, N. M., Mizukami, N., Kagone, S., Leh, M., Mekonnen, K., et al. (2024). Advancing water security in Africa with new high-resolution discharge data. *Scientific Data*, *11*(1), 1195. https://doi.org/10.1038/s41597-024-04034-0

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., & Pappenberger, F. (2013). GloFAS—Global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, *17*(3), 1161–1175. https://doi.org/10.5194/hess-17-1161-2013

Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., & Salamon, P. (2020). A global streamflow reanalysis for 1980–2018. *Journal of Hydrology X*, *6*, 100049. https://doi.org/10.1016/j.hydroa.2019.100049

An, C., Fang, H., Zhang, L., Su, X., Fu, X., Huang, H. Q., et al. (2022). Poyang and Dongting Lakes, Yangtze River: Tributary lakes blocked by main-stem aggradation. *Proceedings of the National Academy of Sciences*, *119*(30), e2101384119. https://doi.org/10.1073/pnas.2101384119

Anderson, W. K., Rausch, R. D., & Bonhaus, D. L. (1996). Implicit/multigrid algorithms for incompressible turbulent flows on unstructured grids. *Journal of Computational Physics*, *128*(2), 391–408. https://doi.org/10.1006/jcph.1996.0219

Barbarossa, V., Huijbregts, M. A., Beusen, A. H., Beck, H. E., King, H., & Schipper, A. M. (2018). FLO1K, global maps of mean, maximum and minimum annual streamflow at 1-km resolution from 1960 through 2015. *Scientific Data*, *5*(1), 1–11. https://doi.org/10.1038/sdata.2018.52

Beck, H. E., de Roo, A., & van Dijk, A. I. J. M. (2015). Global maps of streamflow characteristics based on observations from several thousand catchments. *Journal of Hydrometeorology*, *16*(4), 1478–1501. https://doi.org/10.1175/JHM-D-14-0155.1

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599–3622. https://doi.org/10.1002/2015wr018247

Bloomfield, J. P., Allen, D. J., & Griffiths, K. J. (2009). Examining geological controls on baseflow index (BFI) using regression analysis: An illustration from the Thames Basin, UK. *Journal of Hydrology*, *373*(1–2), 164–176. https://doi.org/10.1016/j.jhydrol.2009.04.025

Bock, A. R., Hay, L. E., McCabe, G. J., Markstrom, S. L., & Atkinson, R. D. (2016). Parameter regionalization of a monthly water balance model for the conterminous United States. *Hydrology and Earth System Sciences*, *20*(7), 2861–2876. https://doi.org/10.5194/hess-20-2861-2016

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

CHASE. (2026). China grid-level natural streamflow estimates (CHASE) [Dataset]. Retrieved from https://hydrodata.cn/chase

Chen, J., Shi, H., Sivakumar, B., & Peart, M. R. (2016). Population, water, food, energy and dams. *Renewable and Sustainable Energy Reviews*, *56*, 18–28. https://doi.org/10.1016/j.rser.2015.11.043

David, C. H., Maidment, D. R., Niu, G.-Y., Yang, Z.-L., Habets, F., & Eijkhout, V. (2011). River network routing on the NHDPlus dataset. *Journal of Hydrometeorology*, *12*(5), 913–934. https://doi.org/10.1175/2011JHM1345.1

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:cartap]2.0.co;2

Dong, N. (2025). Runoff depth and baseflow index maps [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17364737

Dong, N. (2025). Analysis code [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.18470664

Dong, N. (2026). China grid-level natural streamflow estimates (CHASE v1.0) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17386507

Dong, N., Wei, J., Yang, M., Yan, D., Yang, C., Gao, H., et al. (2022). Model estimates of China's terrestrial water storage variation due to reservoir operation. *Water Resources Research*, *58*(6), e2021WR031787. https://doi.org/10.1029/2021wr031787

Dong, N., Yang, M., Wei, J., Arnault, J., Laux, P., Xu, S., et al. (2023). Toward improved parameterizations of reservoir operation in ungauged basins: A synergistic framework coupling satellite remote sensing, hydrologic modeling, and conceptual operation schemes. *Water Resources Research*, *59*(3), e2022WR033026. https://doi.org/10.1029/2022wr033026

Eckhardt, K. (2008). A comparison of baseflow indices, which were calculated with seven different baseflow separation methods. *Journal of Hydrology*, *352*(1–2), 168–173. https://doi.org/10.1016/j.jhydrol.2008.01.005

Farahani, M. A., Wood, A. W., Tang, G., & Mizukami, N. (2025). Calibrating a large-domain land/hydrology process model in the age of AI: The SUMMA CAMELS emulator experiments. *Hydrology and Earth System Sciences*, *29*(18), 4515–4537. https://doi.org/10.5194/hess-29-4515-2025

Feigl, M., Thober, S., Schweppe, R., Herrnegger, M., Samaniego, L., & Schulz, K. (2022). Automatic regionalization of model parameters for hydrological models. *Water Resources Research*, *58*(12), e2022WR031966. https://doi.org/10.1029/2022wr031966

Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., et al. (2024). Deep dive into hydrologic simulations at global scale: Harnessing the power of deep learning and physics-informed differentiable models (δHBV-globe1.0-hydroDL). *Geoscientific Model Development*, *17*(18), 7181–7198. https://doi.org/10.5194/gmd-17-7181-2024

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research*, *58*(10), e2022WR032404. https://doi.org/10.1029/2022wr032404

Fersch, B., Wagner, S., Rummler, T., Gochis, D., & Kunstmann, H. (2013). *The impact of groundwater dynamics and soil-type for modeling coupled water exchange processes between land and atmosphere* (Vol. 359, pp. 140–145). IAHS Publication.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/s0167-9473(01)00065-2

Gebrechorkos, S. H., Leyland, J., Dadson, S. J., Cohen, S., Slater, L., Wortmann, M., et al. (2024). Global-scale evaluation of precipitation datasets for hydrological modelling. *Hydrology and Earth System Sciences*, *28*(14), 3099–3118. https://doi.org/10.5194/hess-28-3099-2024

Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2019). GRUN: An observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, *11*(4), 1655–1674. https://doi.org/10.5194/essd-11-1655-2019

Ghimire, G. R., Hansen, C., Gangrade, S., Kao, S. C., Thornton, P. E., & Singh, D. (2023). Insights from Dayflow: A historical streamflow reanalysis dataset for the conterminous United States. *Water Resources Research*, *59*(2), e2022WR032312. https://doi.org/10.1029/2022wr032312

Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., et al. (2020). Sensitivity analysis-based automatic parameter calibration of the variable infiltration capacity (VIC) model for streamflow simulations over China. *Water Resources Research*, *56*(1), e2019WR025968. https://doi.org/10.1029/2019wr025968

Gou, J., Miao, C., Samaniego, L., Xiao, M., Wu, J., & Guo, X. (2021). CNRD v1.0: A high-quality natural runoff dataset for hydrological and climate studies in China. *Bulletin of the American Meteorological Society*, *102*(5), 1–57. https://doi.org/10.1175/bams-d-20-0094.1

Gudmundsson, L., & Seneviratne, S. I. (2015). Towards observation-based gridded runoff estimates for Europe. *Hydrology and Earth System Sciences*, *19*(6), 2859–2879. https://doi.org/10.5194/hess-19-2859-2015

Guo, Y., Huang, F., Sun, P. A., Zhang, C., Xiao, Q., Wen, Z., & Yang, H. (2023). Hydrogeological functioning of a karst underground river basin in Southwest China. *Groundwater Series*, *61*(6), 895–913. https://doi.org/10.1111/gwat.13361

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Han, D., Liu, P., Zhang, L., Cheng, L., Cheng, Q., Zhang, X., et al. (2024). Quantifying the impact of dams on global streamflow over the period of 1985–2014. *Environmental Research Letters*, *19*(10), 104036. https://doi.org/10.1088/1748-9326/ad6a70

Hanasaki, N., Matsuda, H., Fujiwara, M., Hirabayashi, Y., Seto, S., Kanae, S., & Oki, T. (2022). Toward hyper-resolution global hydrological models including human activities: Application to Kyushu Island, Japan. *Hydrology and Earth System Sciences*, *26*(8), 1953–1975. https://doi.org/10.5194/hess-26-1953-2022

Hao, H., Dong, N., Yang, M., Wei, J., Zhang, X., Xu, S., et al. (2024). The changing hydrology of an irrigated and dammed Yangtze River: Streamflow, extremes, and lake hydrodynamics. *Water Resources Research*, *60*(10), e2024WR037841. https://doi.org/10.1029/2024wr037841

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., et al. (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979-present. *Earth System Science Data*, *12*(3), 2043–2060. https://doi.org/10.5194/essd-12-2043-2020

Hirpa, F. A., Salamon, P., Alfieri, L., Thielen-del Pozo, J., Zsoter, E., & Pappenberger, F. (2018). The effect of reference climatology on global flood forecasting. *Journal of Hydrometeorology*, *17*(4), 1131–1145. https://doi.org/10.1175/JHM-D-15-0044.1

Hobeichi, S., Abramowitz, G., Evans, J., & Beck, H. E. (2019). Linear optimal runoff aggregate (LORA): A global gridded synthesis runoff product. *Hydrology and Earth System Sciences*, *23*(2), 851–870. https://doi.org/10.5194/hess-23-851-2019

Hoch, J. M., Neal, J. C., Baart, F., van Beek, R., Winsemius, H. C., Bates, P. D., & Bierkens, M. F. P. (2017). GLOFRIM v1.0—A global flood modeling framework for coupling hydrological and hydrodynamic models. *Geoscientific Model Development*, *10*, 3913–3929. https://doi.org/10.5194/gmd-10-3913-2017

Hoch, J. M., Sutanudjaja, E. H., Wanders, N., van Beek, L. P. H., & Bierkens, M. F. P. (2023). Hyper-resolution PCR-GLOBWB: Opportunities and challenges from refining model spatial resolution to 1-km over the European continent. *Hydrology and Earth System Sciences*, *27*(6), 1383–1401. https://doi.org/10.5194/hess-27-1383-2023

Hong, Y., Adler, R. F., Hossain, F., Curtis, S., & Huffman, G. J. (2007). A first approach to global runoff simulation using satellite rainfall estimation. *Water Resources Research*, *43*(8). https://doi.org/10.1029/2006wr005739

Hu, J., Miao, C., Su, J., Zhang, Q., Gou, J., & Sun, Q. (2025). An upgraded high-precision gridded precipitation dataset for the Chinese mainland considering spatial autocorrelation and covariates. *Earth System Science Data*, *17*(8), 3987–4004. https://doi.org/10.5194/essd-17-3987-2025

Hutton, E. W. H., Piper, M. D., & Tucker, G. E. (2020). The Basic Model Interface 2.0: A standard interface for coupling numerical models in the geosciences. *Journal of Open Source Software*, *8*(1).

Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Li, L., Bisht, G., Hao, D., & Leung, L. R. (2024). Global 1 km land surface parameters for kilometer-scale Earth system modeling. *Earth System Science Data*, *16*(4), 2007–2032. https://doi.org/10.5194/essd-16-2007-2024

Li, P., Li, T., & Vanapalli, S. K. (2016). Influence of environmental factors on the wetting front depth: A case study in the Loess Plateau. *Engineering Geology*, *214*, 1–10. https://doi.org/10.1016/j.enggeo.2016.09.008

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., et al. (2019). Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resources Research*, *55*(8), 6499–6516. https://doi.org/10.1029/2019WR025287

Liu, S., Shi, H., & Sivakumar, B. (2020). Socioeconomic drought under growing population and changing climate: A new index considering the resilience of a regional water resources system. *Journal of Geophysical Research: Atmospheres*, *125*(15), e2020JD033005. https://doi.org/10.1029/2020JD033005

López, P., Sutanudjaja, E. H., Schellekens, J., Sterk, G., & Bierkens, M. F. (2017). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences*, *21*(6), 3125–3144. https://doi.org/10.5194/hess-21-3125-2017

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., & Merchant, J. W. (2000). Development of a global land cover characteristics database and IGBP DISCover from 1-km AVHRR data. *International Journal of Remote Sensing*, *21*(6–7), 1303–1330. https://doi.org/10.1080/014311600210191

Ma, T., Li, X., Zhang, C., Fu, C., Wang, Z., & Bai, Z. (2022). Identification of origin and runoff of karst groundwater in the glacial lake area of the Jinsha River fault zone, China. *Scientific Reports*, *12*(1), 14661. https://doi.org/10.1038/s41598-022-18960-9

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., et al. (2023). Can hydrological models benefit from using global soil moisture, evapotranspiration, and runoff products as calibration targets? *Water Resources Research*, *59*(2), e2022WR032064. https://doi.org/10.1029/2022wr032064

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), 13603. https://doi.org/10.1038/ncomms13603

MGMR. (1990). *China national geologic survey dataset*. The Geological Publishing House. (in Chinese).

Miao, C., Gou, J., Fu, B., Tang, Q., Duan, Q., Chen, Z., et al. (2022). High-quality reconstruction of China's natural streamflow. *Science Bulletin*, *67*(5), 547–556. https://doi.org/10.1016/j.scib.2021.09.022

Mizukami, N., Clark, M. P., Gharari, S., Kluzek, E., Pan, M., Lin, P., et al. (2021). A vector-based river routing model for Earth system models: Parallelization and global applications. *Journal of Advances in Modeling Earth Systems*, *13*(6), 1–20. https://doi.org/10.1029/2020ms002434

Mizukami, N., Clark, M. P., Sampson, K., Nijssen, B., Mao, Y., McMillan, H., et al. (2016). MizuRoute version 1: A river network routing tool for a continental domain water resources applications. *Geoscientific Model Development*, *9*(6), 2223–2238. https://doi.org/10.5194/gmd-9-2223-2016

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., et al. (2021). The global water resources and use model WaterGAP v2.2d: Model description and evaluation. *Geoscientific Model Development*, *14*(2), 1037–1079. https://doi.org/10.5194/gmd-14-1037-2021

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., & Döll, P. (2014). Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrology and Earth System Sciences*, *18*(9), 3511–3538. https://doi.org/10.5194/hess-18-3511-2014

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Niu, Z., He, H., Zhu, G., Ren, X., Zhang, L., & Zhang, K. (2020). A spatial-temporal continuous dataset of the transpiration to evapotranspiration ratio in China from 1981–2015. *Scientific Data*, *7*(1), 369. https://doi.org/10.1038/s41597-020-00693-x

O'Neill, M. M., Tijerina, D. T., Condon, L. E., & Maxwell, R. M. (2021). Assessment of the ParFlow-CLM CONUS 1.0 integrated hydrologic model: Evaluation of hyper-resolution water balance components across the contiguous United States. *Geoscientific Model Development*, *14*(12), 7223–7254. https://doi.org/10.5194/gmd-14-7223-2021

Orth, R., Dutra, E., & Pappenberger, F. (2016). Improving weather predictability by including land surface model parameter uncertainty. *Monthly Weather Review*, *144*(4), 1551–1569. https://doi.org/10.1175/MWR-D-15-0283.1

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research*, *44*(3), W03413. https://doi.org/10.1029/2007wr006240

Pagliero, L., Bouraoui, F., D'Haeyer, T., Bidoglio, G., & De Roo, A. (2019). Investigating regionalization techniques for large-scale hydrological modelling. *Journal of Hydrology*, *570*, 220–235. https://doi.org/10.1016/j.jhydrol.2018.12.071

Peckham, S. D., Hutton, E. W. H., & Norris, B. (2013). A component-based approach to integrated modeling in the geosciences: The design of CSDMS. *Computers & Geosciences*, *53*, 3–12. https://doi.org/10.1016/j.cageo.2012.04.002

Povak, N. A., Hessburg, P. F., McDonnell, T. C., Reynolds, K. M., Sullivan, T. J., Salter, R. B., & Cosby, B. J. (2014). Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. *Water Resources Research*, *50*(4), 2798–2814. https://doi.org/10.1002/2013wr014203

Price, K. (2011). Effects of watershed topography, soils, land use, and climate on baseflow hydrology in humid regions: A review. *Progress in Physical Geography*, *35*(4), 465–492. https://doi.org/10.1177/0309133311402714

Rajib, A., Evenson, G. R., Golden, H. E., & Lane, C. R. (2018). Hydrologic model predictability improves with spatially explicit calibration using remotely sensed evapotranspiration and biophysical parameters. *Journal of Hydrology*, *567*, 668–683. https://doi.org/10.1016/j.jhydrol.2018.10.024

Rogger, M., Agnoletti, M., Alaoui, A., Bathurst, J. C., Bodner, G., Borga, M., et al. (2017). Land use change impacts on floods at the catchment scale: Challenges and opportunities for future research. *Water Resources Research*, *53*(7), 5209–5219. https://doi.org/10.1002/2017wr020723

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5), W05523. https://doi.org/10.1029/2008WR007327

Shrestha, P. K., Samaniego, L., Rakovec, O., Kumar, R., & Thober, S. (2025). A novel stream network upscaling scheme for accurate local streamflow simulations in gridded global hydrological models. *Water Resources Research*, *61*(6), e2024WR038183. https://doi.org/10.1029/2024wr038183

SigMap. (2026). China hydrologic signature maps (SigMap v1.0) [Dataset]. Retrieved from https://hydrodata.cn/sigmap

Sloto, R. A., & Crouse, M. Y. (1996). HYSEP: A computer program for streamflow hydrograph separation and analysis. *USGS Water-Resources Investigations Report 96-4040*, 46. http://water.usgs.gov/software/HYSEP/code/doc/hysep.pdf

Song, Y., Bindas, T., Shen, C., Ji, H., Knoben, W. J., Lonzarich, L., et al. (2025). High-resolution national-scale water modeling is enhanced by multiscale differentiable physics-informed machine learning. *Water Resources Research*, *61*(4), e2024WR038928. https://doi.org/10.1029/2024wr038928

Song, Z., Xia, J., Wang, G., She, D., Hu, C., & Hong, S. (2022). Regionalization of hydrological model parameters using gradient boosting machine. *Hydrology and Earth System Sciences*, *26*(2), 505–524. https://doi.org/10.5194/hess-26-505-2022

Su, T., Miao, C., Duan, Q., Gou, J., Guo, X., & Zhao, X. (2023). Hydrological response to climate change and human activities in the three-river source region. *Hydrology and Earth System Sciences*, *27*(7), 1477–1492. https://doi.org/10.5194/hess-27-1477-2023

Tang, G., Wood, A. W., & Swenson, S. (2025). On using AI-based large-sample emulators for land/hydrology model calibration and regionalization. *Water Resources Research*, *61*(7), e2024WR039525. https://doi.org/10.1029/2024wr039525

Tilloy, A., Paprotny, D., Grimaldi, S., Gomes, G., Bianchi, A., Lange, S., et al. (2025). HERA: A high-resolution pan-European hydrological reanalysis (1951–2020). *Earth System Science Data*, *17*(1), 293–316. https://doi.org/10.5194/essd-17-293-2025

Tounsi, A., Abdelkader, M., & Temimi, M. (2023). Assessing the simulation of streamflow with the LSTM model across the continental United States using the MOPEX dataset. *Neural Computing & Applications*, *35*(30), 22469–22486. https://doi.org/10.1007/s00521-023-08922-1

Van der Vorst, H. A. (2003). *Iterative Krylov methods for large linear systems (No. 13)*. Cambridge University Press.

van Jaarsveld, B., Wanders, N., Sutanudjaja, E. H., Hoch, J., Droppers, B., Janzing, J., et al. (2025). A first attempt to model global hydrology at hyper-resolution. *Earth System Dynamics*, *16*(1), 29–54. https://doi.org/10.5194/esd-16-29-2025

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., et al. (2010). Global threats to human water security and river biodiversity. *Nature*, *467*(7315), 555–561. https://doi.org/10.1038/nature09440

Wada, Y., Reager, J. T., Chao, B. F., Wang, J., Lo, M.-H., Song, C., et al. (2017). Recent changes in land water storage and its contribution to sea level variations. *Surveys in Geophysics*, *38*(1), 131–152. https://doi.org/10.1007/s10712-016-9399-6

Wagner, S., Fersch, B., Yuan, F., Yu, Z., & Kunstmann, H. (2016). Fully coupled atmospheric-hydrological modeling at regional and long-term scales: Development, application, and analysis of WRF-HMS. *Water Resources Research*, *52*(4), 3187–3211. https://doi.org/10.1002/2015WR018185

Wang, C., Jiang, S., Zheng, Y., Han, F., Kumar, R., Rakovec, O., & Li, S. (2024). Distributed hydrological modeling with physics-encoded deep learning: A general framework and its application in the Amazon. *Water Resources Research*, *60*(4), e2023WR036170. https://doi.org/10.1029/2023WR036170

Wang, K., Liu, X., Cui, P., Zhang, Y., Xie, J., Liu, C., & Gosling, S. N. (2025). China's nationwide streamflow decline driven by landscape changes and human interventions. *Science Advances*, *11*(32), eadu8032. https://doi.org/10.1126/sciadv.adu8032

Wieder, W. (2014). Regridded harmonized world soil database v1.2 [Dataset]. *ORNL Distributed Active Archive Center (DAAC)*. https://doi.org/10.3334/ORNLDAAC/1247

Winter, C., Jawitz, J. W., Ebeling, P., Cohen, M. J., & Musolff, A. (2024). Divergence between long-term and event-scale nitrate export patterns. *Geophysical Research Letters*, *51*(10), e2024GL108437. https://doi.org/10.1029/2024gl108437

Xie, K., Liu, P., Zhang, J., Wang, G., Zhang, X., & Zhou, L. (2021). Identification of spatially distributed parameters of hydrological models using the dimension-adaptive key grid calibration strategy. *Journal of Hydrology*, *598*, 125772. https://doi.org/10.1016/j.jhydrol.2020.125772

Xu, M., Wang, P., Zhang, X., Ma, T., Jin, J., Kang, S., et al. (2025). Impacts of glacier shrinkage on peak melt runoff at the sub-basin scale of Northwest China. *Journal of Hydrology*, *654*, 132953. https://doi.org/10.1016/j.jhydrol.2025.132953

Xu, R., Zeng, Z., Pan, M., Ziegler, A. D., Holden, J., Spracklen, D. V., et al. (2023). A global-scale framework for hydropower development incorporating strict environmental constraints. *Nature Water*, *1*(1), 113–122. https://doi.org/10.1038/s44221-022-00004-1

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research*, *55*(6), 5053–5073. https://doi.org/10.1029/2019WR024873

Yan, J., Jia, S., Lv, A., & Zhu, W. (2019). Water resources assessment of China's transboundary river basins using a machine learning approach. *Water Resources Research*, *55*(1), 632–655. https://doi.org/10.1029/2018WR023044

Yang, C., Lin, Z., Yu, Z., Hao, Z., & Liu, S. (2010). Analysis and simulation of human activity impact on streamflow in the Huaihe River basin with a large-scale hydrologic model. *Journal of Hydrometeorology*, *11*(3), 810–821. https://doi.org/10.1175/2009JHM1145.1

Yang, L., Yang, Y., Villarini, G., Li, X., Hu, H., Wang, L., et al. (2021). Climate more important for Chinese flood changes than reservoirs and land use. *Geophysical Research Letters*, *48*(11), e2021GL093061. https://doi.org/10.1029/2021GL093061

Yang, Y., Feng, D., Beck, H. E., Hu, W., Abbas, A., Sengupta, A., et al. (2025). Global daily discharge estimation based on grid long short-term memory (LSTM) model and river routing. *Water Resources Research*, *61*(6), e2024WR039764. https://doi.org/10.1029/2024WR039764

Yang, Y., Pan, M., Lin, P., Beck, H. E., Zeng, Z., Yamazaki, D., et al. (2021). Global reach-level 3-hourly river flood reanalysis (1980–2019). *Bulletin of the American Meteorological Society*, *102*(11), E2086–E2105. https://doi.org/10.1175/bams-d-20-0057.1

Yao, J., Zhang, D., Li, Y., Zhang, Q., & Gao, J. (2019). Quantifying the hydrodynamic impacts of cumulative sand mining on a large river-connected floodplain lake: Poyang Lake. *Journal of Hydrology*, *579*, 124156. https://doi.org/10.1016/j.jhydrol.2019.124156

Yu, Z., Lakhtakia, M. N., Yarnal, B., White, R. A., Miller, D. A., Frakes, B., et al. (1999). Simulating the river-basin response to atmospheric forcing by linking a mesoscale meteorological model and hydrologic model system. *Journal of Hydrology*, *218*(1–2), 72–91. https://doi.org/10.1016/S0022-1694(99)00022-0

Yu, Z., Pollard, D., & Cheng, L. (2006). On continental-scale hydrologic simulations with a coupled hydrologic model. *Journal of Hydrology*, *331*(1–2), 110–124. https://doi.org/10.1016/j.jhydrol.2006.05.021

Zhan, Y., Xu, G., Wang, B., Wu, G., Wu, J., Zhao, T., & Wu, X. (2025). Exacerbated variability and extremes in streamflow across half of China from 1961 to 2018. *Water Resources Research*, *61*(12), e2025WR041968. https://doi.org/10.1029/2025wr041968