Benchmark Framework for Global River Model (Version 1.0)

Xudong Zhou[1,2], Dai Yamazaki[2], Menaka Revel[2,3], Gang Zhao[2], Prakat Modi[4,2]

[1]Institute of Hydraulic and Ocean Engineering, Ningbo University, Ningbo 315211, China

[2]Global Hydrological Prediction Center, Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

[3]Department of Civil and Environmental Engineering, Faculty of Engineering, University of Waterloo, Waterloo, Canada

[4]SIT Research Laboratories, Shibaura Institute of Technology, Tokyo, Japan

Corresponding author: Xudong Zhou (zhouxudong@nbu.edu.cn)

Key Points:

- We developed a benchmark framework for global river models, ensuring quick and comprehensive performance analysis.
- Remote sensing data for water surface elevation and inundation extent helps address the lack of extensive in-situ discharge observations.
- The benchmark model is highly adaptable, allowing for evaluation of model development and intercomparison across multiple models.

Abstract

Global River Models (GRMs), which simulate river flow and flood processes, have rapidly developed in recent decades. However, these advancements necessitate meaningful and standardized quality assessments and comparisons against a suitable set of observational variables using appropriate metrics, a requirement currently lacking within GRM communities. This study proposes the implementation of a benchmark system designed to facilitate the assessment of river models and enables comparisons against established benchmarks. The benchmark system incorporates satellite remote sensing data complementing *in-situ* data, including water surface elevation and inundation extent information, with necessary preprocessing. Consequently, this evaluation system encompasses a larger geographical area compared to traditional methods relying solely on in-situ river discharge measurements for GRMs. A set of evaluation and comparison metrics has been developed, including a quantile-based comparison metric that allows for a comprehensive analysis of multiple simulation outputs. The test application of this benchmark system to a global river model (CaMa-Flood), utilizing diverse runoff inputs, illustrates that the incorporation of bias-corrected runoff data leads to improved model performance across various observational variables and performance metrics. The current iteration of the benchmark system is suitable for global-scale assessments and can effectively evaluate the impact of model development as well as facilitate intercomparisons among different models. The source codes are accessiable from https://doi.org/10.5281/zenodo.10903211.

Plain Language Summary

River models, which help us understand how rivers flow and flood, have gotten a lot better over the years. But, there isn't an agreed-upon way to check if these models are doing a good job by comparing them to real-world data. This study suggests creating a system to test and compare river models more effectively. This new system uses both satellite images and ground measurements to get a full picture of model abilities in simulating different flow characteristics. This study tried this system on a specific global river model, called CaMa-Flood, with different types of data to see how well it works. Results found that using corrected data makes the model predictions better match what we see in the real world across various tests and measurements. This testing system is ready to be used worldwide and can help see how changes to the models improve their predictions. It also makes it easier to compare different models to see which ones work best.

1 Introduction

The global river model (GRM) is one crucial approach for reproducing river water dynamics on a large scale (Lehner & Grill, 2013). No matter whether the river models stand along or are integrated with runoff generation models, the simulated flow process can be used for applications including water resources assessment, flood forecasting, environmental and socio-economic development (Hanasaki et al., 2008; Tharme, 2003; Winsemius et al., 2013). The evaluation of the river models is also a typical way to help calibrate and validate land surface processes, of which accurate observations are often infeasible (Hou et al., 2023; Zaitchik et al., 2010). Ensuring the performance of GRMs is therefore essential for the reliability of the revealed

physics relevant to climate forcing, rainfall-runoff models, and routing models (Chen et al., 2021; Hou et al., 2023; Zhao et al., 2017).

However, it is important to acknowledge that GRM simulations remain imperfect, exhibiting persisting biases linked to factors such as model parameterization and structural aspects (Bernhofen et al., 2018a; Hirpa et al., 2021; Zhou, Ma, et al., 2021). Biases in model forcing inputs also lead to consequent biases in river models (Hou et al., 2023). Despite the progress made to model improvement, the evaluation of GRM implementations against reliable observations remains an essential step (Hoch & Trigg, 2019). Although important, standardized guidelines or instructions for conducting this process, especially for GRMs is still needed. This lack of standardized procedures, including study area, evaluation periods, and evaluation metrics, makes comparing and contrasting findings across different studies challenging (Bernhofen et al., 2018b; Trigg et al., 2016).

Moreover, a notable gap exists in the evaluation of river models as the current implementation of river model evaluation relies heavily on river discharge observations. Other crucial measurements, such as water surface elevation (WSE) and inundation area (WSA), albeit received increasing attention on its own dynamics, no standardized method has been developed for evaluation GRMs using satellite WSE and WSA on a large scale (Eilander et al., 2018; Wu et al., 2019; Yamazaki et al., 2012). This lack of comparison is attributed to two primary factors: inadequate establishment and assessment of observations for these variables, and the limited modeling capability to accurately simulate WSE and WSA. These variables play pivotal roles in routing processes and flood dynamics, exerting substantial influence on hazard and risk analyses (Mason et al., 2007; Tellman et al., 2021). Furthermore, the distinct physical characteristics of WSE and WSA warrant separate evaluation approaches, offering additional insights into the capabilities of GRMs (Musa et al., 2015; Zhou, Prigent, et al., 2021).

In the field of Earth System Models, model benchmark serves as a protocol for establishing standardized evaluations and comparisons using a reference, which could be a baseline or a previous version before model development. An example of such a benchmark system is the International Land Model Benchmarking (ILAMB) framework (Collier et al., 2018), which offers a thorough and multifaceted assessment of land model projections. This system provides readily accessible data, codes, and resources for standard model performance evaluation and model-data intercomparison. Similarly, the Earth System Model Evaluation Tool (ESMValTool) (Eyring et al., 2016; Lauer et al., 2020) offers web services, coding packages, and community engagement for Earth System Model evaluations. The Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER) is another project-oriented approach with more flexibility, albeit with a less structured organization (Best et al., 2015).

However, these existing benchmark systems are designed for land surface models and Earth System Models, predominantly relying on gridded reference data but overlook horizontal hydrological processes which have additional issues related to data scarcity and representativity as follows:

Data scarcity. River discharge is the most convincing data for model calibration and validation, while its accessibility still hampers the utilizations in various scales (Burek & Smilovic, 2023).

Current frequently-used data from Global Runoff Database Center (GRDC), Global Streamflow Indices and Metadata Archive (GSIM (Do et al., 2018; Gudmundsson et al., 2018)) and some national databases, still lack easy access to their daily raw records. Data coverage is especially limited in regions of Asia and Africa, where water resource scarcity and flood hazards are more pronounced (Kettner et al., 2021). The lack of recent data records after 1990s seriously undermines the reliability of model evaluations for the new environment with climate change and dense human activities (Elmi et al., 2024). Satellite remote sensing data has become increasingly popular for deriving water dynamic variables in recent years (Musa et al., 2015). While it offers improved spatial coverage, its accuracy may not yet be satisfactory for direct use as reference data (Zhou, Prigent, et al., 2021). The incorporation of remote sensing variables like WSA and WSE in model calibration and validation are gradually seen at much smaller scales (Jiang et al., 2019; Wood et al., 2016). However, their application on a global scale necessitates careful consideration about variabilities in climate and geography conditions.

Data representativity. Unlike land surface variables which are often confined to specific grid points without negligible horizontal movement or exchanges, hydrological processes exhibit a higher degree of complexity (Dingman, 2015). For instance, measurements of river discharge at a given location reflect not only the immediate physical processes but also the cumulative effects of water generation and flow from the entire upstream catchment area. Variables such as WSA and WSE are also influenced not solely by upstream flow but also by local topographical features, including channel slopes and floodplain terrain (Zhou, Prigent, et al., 2021). River-status variables can change drastically along river longitudinal direction. Moreover, errors in location in the river network (i.e., mainstem or tributary) will cause more critical errors in evaluated variables compared to longitudinal location difference. The point location of Q and WSE thus requires careful allocation to represent the correct measurement (Krabbenhoft et al., 2022; Revel et al., 2023). The placement errors, complexities of river channels (e.g., river confluence, bifurcation) and aligning the measurement point and calculation nodes arise challenging in the allocation process.

Evaluation metrics. Addressing the data scarcity challenge outlined earlier, a recommendation is to incorporate new variables, such as WSA and WSE, to better reflect the underlying flow regimes and spatial coverage. However, a twofold complexity emerges (Modi et al., 2022). Firstly, the evaluation of different variables (i.e., discharge, WSE and WSA) should inherently differ in magnitudes. Secondly, distinct model performance metrics capture varying flow processes. For instance, mean bias reveals systematic deviations, correlation encapsulates process representation, and Nash-Sutcliffe Efficiency (NSE) emphasizes high values. With each metric having its own unique range and magnitude (for instance, correlation values often surpass NSE values), the integration of these disparate metrics into a coherent framework for comprehensive model performance assessment becomes an intricate challenge.

This study proposes to construct the first benchmark system for GRMs. The primary step is the preparation and careful preprocessing of diverse observed data for evaluation, including river discharge, water level, and water area. Furthermore, this study has devised an array of metrics, specially tailored to assess model performance, along with a refined methodology for discerning performance discrepancies between two model experiments. As a test case for applying the developed benchmark system, the study quantifies the extent and nature of enhancements due to model development of runoff inputs. This comprehensive framework enables an exploration of

how these improvements impact model performance across various dimensions. Importantly, the benchmark system possesses the potential for future applications in the evaluation of other global flood models. The structure of this study is organized as follows: Section 2 introduces the general methods and data; Section 3 introduces the sample implementation used in this study; Section 4 presents the results for the sample case; Sections 5 and 6 are the discussions and conclusions.

2 Methodology and Data

Here is the overall structure of the benchmark system, including data preparation (including simulation data and observational data), analysis (including designing of evaluation metrics and comparison metric), and visualization based on evaluations (see Fig. 1). Note that the models and datasets (e.g., CaMa-Flood, GRDC, HydroWeb, GIEMS shown in Fig. 1) are used as examples, but the framework is extendable for other models and/or observation data. Users can replace or extend these sample data as needed, and users can still use the benchmark system even if part of the variables or functions are not used. Details will be separately introduced in the following subsections.

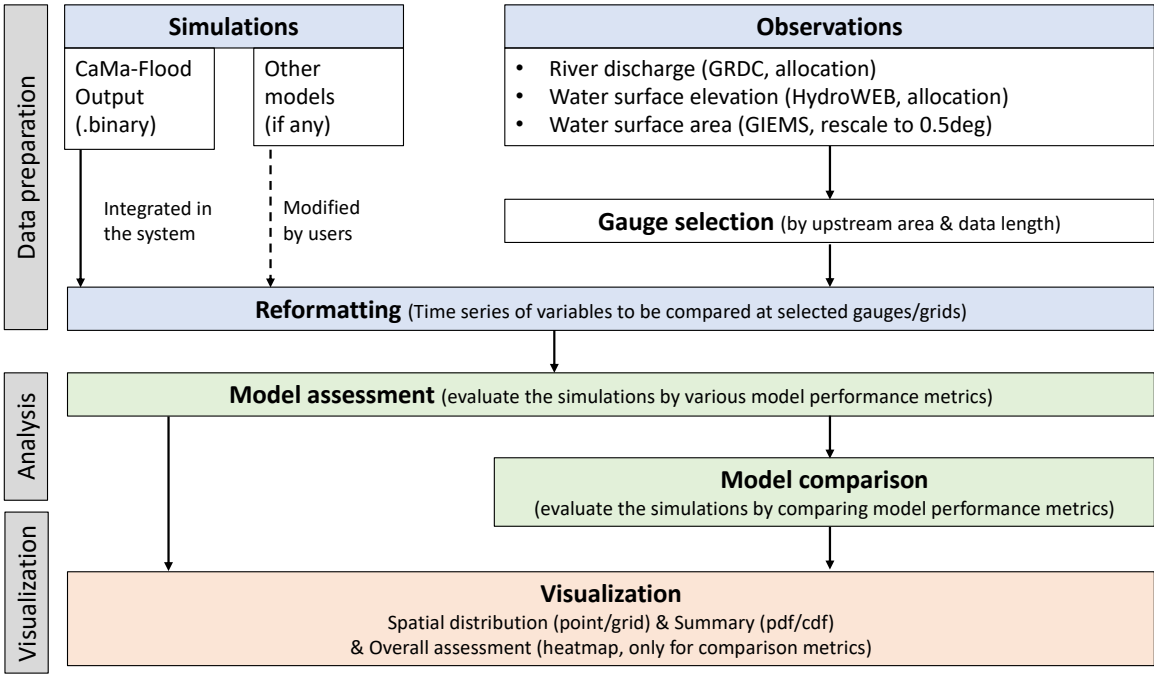

Figure 1. Structures of the benchmark system.

1.1 Data preparation

2.1.1 Model simulations

The benchmark system is designed to assess the general performance of global river models. Depending on the model capability of simulations, three key variables: river discharge (Q), water surface elevation (WSE), and water surface area (WSA), can be provided and evaluated. These variables are essential for a comprehensive evaluation, while the nature of the investigated

variables allows for independent evaluation. The standard input format mandates a three-dimensional model output structure (time, latitude, longitude). The system also accommodates post-processed input data, with data format consisting of time series linked to unique identifiers (IDs) of river gauges or virtual stations, following the format of two-dimensions (time, gauges).

2.1.2 Observations

The default observations to be used in this system include Q, WSE and WSA. The discharge data are expected to be time series, although missing data are allowed within an acceptable ratio. As point data, the coordinates of the gauge are mandatory. Other attribute such as upstream area is recommended, while if it is missing, users can extract this information from other topography data such as MERIT Hydro or HydroSHEDS. WSE and WSA on the large scale are extracted from remote sensing missions. In addition to the coordinates of the WSE virtual station, the datum (the reference elevation system) of it is needed, which is provided with the satellite attributes. The grid information of the surface area is also needed for reasonable comparisons. Thresholds can be set up to fill out gauge/grids without data long enough. The following paragraphs showcase the data we prepared for the initial version of the benchmark system (V1.0). Users can also update or replace the prepared observations for their convenience. Moreover, users can update the system for other observed variables such as river widths for an extendable usage.

a) River discharge (Q)

The daily discharge records were collected from Global Runoff and Data Center (GRDC, http://www.bafg.de/GRDC/, last access: Dec 2, 2023), since GRDC is currently the most used dataset of river discharge. All records are accessible and downloadable from the website https://portal.grdc.bafg.de/. However, a statement of research purpose and consulting with the GRDC team are needed if a large amount of data is requested. All gauges were included in the benchmark system, while gauges can be automatically excluded with user-defined thresholds in a portion of missing data or upstream area according to implementation purpose. In this study, gauges with time series of fewer than two years of records in the sample investigation period (2001-2012) and with an upstream area of less than 10000 km^2 were excluded from the analysis. In total, 1009 gauges remained for evaluation.

b) Water altimetry (WSE)

Satellite altimetry data were obtained from HydroWeb (http://hydroweb.theia-land.fr/, last access: Dec 2, 2023) because it integrates the largest number of satellite missions and provides processed data to users. The altimetry data are recorded at virtual stations (VSs) where satellite ground tracks cross the river network. The temporal interval and time coverage of the altimetry are different at each VS depending on the different altimetry data sources (e.g., ENVISAT from 2002-2012, Jason-1 from 2001-2013 and Jason-2 from 2008). The data length threshold is also applied to the VSs to exclude those gauges with very few available records (less than 60 in this study). In total, 1392 VSs worldwide were used during the period of interest.

c) Water surface area (WSA)

We prepared GIEMS-2 as observations for WSA, which is a multi-source product based on microwaves (Prigent et al., 2020). It is the accumulative water surface area at a 25km sampling density with a monthly interval from 1992 to 2015 and we converted it to 0.25deg resolution for easier comparison. Although there are products of the finer spatio-temporal resolution, e.g., Landsat and MODIS, we decided not to use them in the developed version (V1.0) because a first glace on the model capability on modeling surface area is needed before investigating complicated information at local scales. However, we need to bear in mind that previous studies confirmed GIEMS often overestimates the water area in tropical areas or rainforests because of the backscattered effect on saturated soil moisture (Aires et al., 2017; Zhou, Prigent, et al., 2021).

2.1.3 Gauge allocation

a) Allocation of river discharge gauges

River models are often discrete to cascaded routing units to reduce the calculation computation. Therefore, to accurately evaluate the model simulations with observed measurements, it is crucial to exercise caution of allocation, adapting observations meticulously to the specific characteristics of the calculation node. The measurement of Q and WSE is at a point location, which can match well with a particular simulation node in a high-resolution model (e.g., LISFLOOD-FP). While for coarse resolution with more than a few kilometers (which is normal of large-scale simulations), the measurement becomes more difficult to match with the simulations. This will cause a large misrepresentation of Q, especially around river confluence points and systematic bias in the WSE. Therefore, allocation of the river discharge gauge is generally to ensure tributaries are accurately represented by simulations and observations. Besides, the allocation of the WSE is to capture the offset of elevation in the VS and simulation node.

With a basic strategy of the allocation, users can search the nearby calculation nodes within a certain distance (e.g., 3 by 3 grids) from the investigated gauge and find the one with the minimal error in upstream area compared to the recorded value of the observational gauge. We also allow advanced allocation method which is designed for specific settings (i.e., MERIT-Hydro river network for CaMa-Flood). The advanced allocation considers in a better way the relationship between gauge and river maps in high- and low-resolution. Details of the optional advanced allocation can be found in section 3.3.1.

b) Allocation of virtual stations (VSs)

The allocation of the VSs is to find the simulation Grid ID (ix,iy) corresponding to the VSs. While allocating VSs is more complicated compared to the allocation of river discharge gauges, the water surface elevation is much more varied within a short distance. Also, normally models report WSE for a representative location for river reach (model grid) whereas the WSE slope is continuous. It thus requires an additional offset of the elevation value to match the model simulation nodes. Moreover, discharge gauge is usually located in river channel where observation is relatively easy (e.g., narrow river segments) after careful in-situ investigation. While the satellite VSs are available anywhere along the ground track and they do not have the information of upstream area. Thus, allocation of gauges in some section (e.g. braided channels) needs attention. Simple allocation is still possible based on coordinates information, to find the

simulation Grid ID with the shortest distance. Offset is then calculated as the difference in elevation of the VS and the outlet corresponding to grid. While offset is not always necessary if hydrological model cannot simulate absolute water level dynamics. In case of that the benchmark system still works when offset is undefined (i.e., only relative water level dynamics is compared). Moreover, advanced option which considers the sub-grid river networks (e.g., bias-correction over river channels, the bifurcation rivers) is implemented in this benchmark system for the MERIT-Hydro river network (see section 3.3.2 for details).

c) Comparison of surface area

Comparing WSA is relatively simple as long as we assume the sum of water surface in calculation is corresponding to the range of observation. In case the spatial resolution of the simulation and observation is different, we first unify them before comparison. If needed, advanced option is possible that the simulation is first downscaled to a super-high spatial resolution (which is not restricted to the shape of the simulation catchment) and then upscaled to the same unit of observations grid for a precise comparison (see section 3.3.3 for details).

2.2 Analysis

2.2.1 Model efficiency metrics

Model performance is assessed with various efficiency metrics. All efficiency metrics are calculated in the framework for all variables, including Q, WSE and WSA. However, specific metrics are more important to particular variables because they reflect different features. For instance, evaluation on the high peaks is more important for flood management, while evaluation on the low peaks is more important for drought assessment. The purpose of V1.0 is to make up a generally applicable framework so we included as many metrics as possible. Further studies are needed on which metrics should be focused on for further improve the model.

These metrics are categorized into two groups (Table 1): the first category is the state evaluation (e.g., the bias of the mean value, maximum value, minimum value, amplitude, standard deviation, RMSE), with the optimal value as 0; the second category evaluates the overall accuracy of variables (e.g., correlation, Kling-Gupta efficiency, Nash-Sutcliffe efficiency), with the optimal value as 1. In addition, we also assessed model performance by subtracting mean values independently from the simulated and observed series (i.e., r_RM, NS_RM and kge_RM) to exclude the error due to systematic bias, which is especially useful for WSE because WSE is defined for different datums.

281 **Table 1. Efficiency metrics used in this study.**

	Abbreviation	Assessment	Equation
Category 1	pbias (max)	extreme	$pbias = \frac{\sum(y_i - x_i)}{\sum x_i}$
	pbias (min)	extreme	
	pbias (mean)	mean	
	pbias(ampli)	amplitude	
	pbias(std)	variation	
	RMSE_nor	Deviations (normalized)	$RMSE\_nor = \sqrt{\sum \frac{(y_i - x_i)^2}{n}} \Big/ \bar{x}$
Category 2	r	Pearson's correlation coefficient	$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$
	NSE	Nash-Sutcliffe efficiency, high values	$NSE = 1 - \frac{\sum(y_i - x_i)^2}{\sqrt{\sum(y_i - \bar{y})^2}}$
	kge	Kling-Gupta efficiency, overall performance	$kge = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_y}{\sigma_x} - 1)^2 + (\frac{\mu_y}{\mu_x} - 1)^2}$
	r_RM NS_RM kge_RM	Process evaluation without systematic bias	Same with the equations while observation and simulation are processed by subtracting the mean value

Where x_i, y_i are observed and simulated values for different variable pairs.

282

283 ### 2.2.2 Model comparison evaluation metrics

284 In addition to the performance metrics which is designed for each location, comparison metric is
285 an aggregated score to decide which model version is better. It evaluates the model's
286 improvement or deterioration compared to the reference simulation (named as baseline model in
287 this study). The direct comparison works conventionally. For instance, a higher correlation or
288 low bias represents better model simulations. However, because the abovementioned metrics
289 vary in the optimal value (i.e., 0 or 1) and the ranges (e.g., $-1 \leq r \leq 1$, $0 \leq RMSE \leq +\infty$,
290 $-\infty \leq NSE \leq 1$), these metrics are difficult to be directly compared or to be integrated.
291 Therefore, we introduce two more comparison metrics besides the direct comparison metric (see
292 Fig.2).

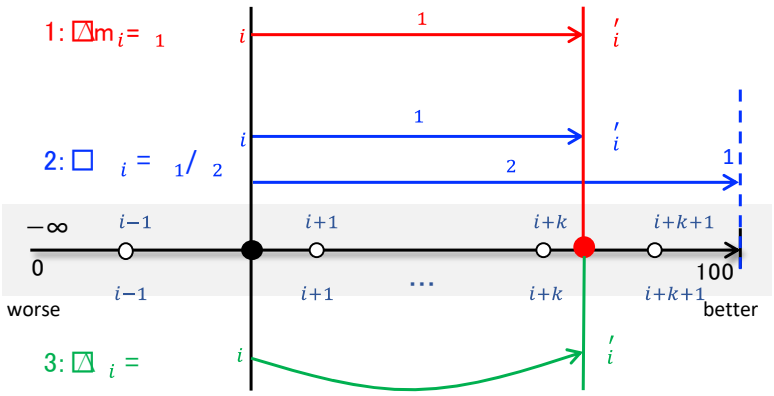

Figure 2. Illustration of different comparison metrics. The three colors represent the three different ways of evaluating the changes in model evaluation.

a) C1: Delta index

The difference in evaluation metric for the new model (NM) experiment compared to the baseline model (BM) for a specific sample location i is represented as the change in metrics (Δm_i) for a new simulation from the baseline simulation (illustrated as red in Fig. 2).

$$\Delta m_i = m'_i - m_i$$

For category 2, a positive Δm_i represents a model improvement. While for category 1, we evaluated it with absolute values for the evaluation metric, and a negative value (approaching the optimal value 0) represents a model improvement.

$$\Delta m_i = |m'_i| - |m_i|$$

b) C2: Improvement index

Δm_i has limitations in gauging the comparative enhancements over the baseline. To illustrate, a rise of 0.1 for NSE from 0 to 0.1 may be relatively less challenging than a progression from 0.8 to 0.9. Consequently, we additionally assessed the improvement index in relation to the optimal value. This approach captures alterations relative to the optimal value (i.e., 0 or 1), as outlined by (Seibert et al., 2018) and illustrated in blue in Fig.2.

$$IMI_i = \frac{m'_i - m_i}{m_{opt} - m_i}$$

m_{opt} is the optimal value for the investigated metrics (0 for metrics in Category 1 and 1 for metrics in Category 2). The maximum *IMI* value is 1, indicating that the New Model (NW) is perfect; 0 means no changes. The higher value indicates a more noticeable improvement in model performance for NW than the baseline (BM). A negative value shows deteriorated model performance.

c) C3: Quantile index

When assessing the overall model performance across various metrics and variables, it is common practice to aggregate metric values with certain weights. However, the lack of clear guidelines for assigning these weights has led to the use of uniform weights in previous studies (Modi et al., 2022). Meanwhile, the magnitudes of direct change index and improvement index values differ across metrics (e.g., correlation coefficient, NSE) and variables (e.g., discharge, water surface elevation, water surface area). To address this, we propose a novel evaluation approach utilizing quantile changes (illustrated in green in Fig. 2).

We first assess the model performance at each specific location for the baseline model, whether it is a gauge or grid. We sorted all samples and calculated the quantile values (denoted as q_i). When evaluating the performance of a new model (NM) for the same specific location, we check the location of the new metrics among the baseline samples (q'_i). The change in the ranking is therefore the evaluation of the specific location (Δq_i).

$$\Delta q_i = q'_i - q_i$$

A positive Δq_i denotes an enhancement in model performance at a specific location (or overall model performance). Importantly, this quantile change approach renders the quantifiable changes in a consistent format, facilitating comparisons across diverse metrics and variables. To ensure the reliability and meaningfulness of comparisons, we recommend utilizing a comparison metric with a minimum of 100 gauges. This threshold ensures a robust statistical foundation for the ranking methodology and guarantees the statistical significance of the outcomes.

2.3 Output & Visualization

2.3.1 Reformatting output

The system provides three levels of outputs for different purposes. The level 1 (L1) is for restructured data from model space to observation space and evaluation to be conducted in the Reformatting step illustrated in Fig. 1. The output is in a format of time series data for each location. The level 2 (L2) output is estimated evaluation metrics based on L1 data, with a format of multiple metrics for each location. For WSA, the L2 output is the 2D map with multiple evaluation metrics layers at each spatial grid. As mentioned in the Input, L1 data can also be prepared independently by other models. Level 3 (L3) is the evaluation of the improvement or deterioration of the model simulated based on L2 data and equations in 2.2.2, with data format as different evaluation metrics for each location. Only the identical gauges (with the same ID) or locations in grids are compared, so that the system allows comparisons if inputs are in different spatial extents. Evaluations over the three variables (Q, WSE, WSA) are independent. For instance, the land surface model or hydrological model, which only simulates river discharge, can be assessed and intercompared as well, ensuring the broad usage of this benchmark system. The intermediate data are very useful, especially for further analysis and plots by avoiding repeating massive I/O and calculation.

2.3.2 Visualization

The system provides some built-in visualization styles, from point maps to overall evaluation matrix maps. A point map is designed for evaluating model performance over space, showing the hotspots where the model performs well or bad and where obvious improvement or deterioration occurs. The overall matrix map shows how the model changes over the baseline model and how overall performance varies across different evaluation metrics. Users can easily extend applications with other visualization styles with the three levels of data mentioned above.

2.4 Structures of the benchmark system

The benchmark toolkit has six main steps, that all can be executed with a very single execution command. For example, execute s01-initial.sh by running ‘./s01-initial.sh’ to initializae the benchmarking tool. And execute s03-reformat.sh by running ‘./s03-reformat.sh $var’ to reformat the model outputs and observations datasets for comparsions. $var can be any one or more of Q (‘dis’ or ‘discharge’), WSE (‘wse’ or ‘sfcelv’) and WSA (‘wsa’ or ‘fldare’), in case of different naming in various outputs. The detailed manual about how to prepare the system and the execution can be found in Supplemetary Text S2.

a) Initialize (creates the output directory structure)
b) visual_pre (visuals the model simulation outputs)
c) reformat (reformats model outputs and observations for comparison)
d) statis (calculated model evaluation statistics)
e) visual (visualizes the comparison between model outputs and observations)
f) summary (summarizes and plots the performance of each model in a heatmap using various evaluation metrics)

3 Sample implementation within CaMa-Flood

Although the benchmark system is designed for all global river models, we showcase the CaMa-Flood outputs to illustrate the system, since CaMa-Flood is a large-scale hydrodynamic models, providing all three variables and we are easily set up different driving conditions to showcase the discrepancies among settings. For river models that cannot simulate WSE or WSA, they still can use the benchmark system to evaluate river discharge.

3.1 Global Flood Model (CaMa-Flood)

CaMa-Flood is designed for large-scale river hydrodynamics simulations. Using a local iteral form of shallow water momentum equation, it can simulate the river discharge, water level and water storage along the river network (i.e., MERIT Hydro). The river channel is simplified as a rectangular shape, and the floodplain topography is aggregated within a unit catchment (the basic calculation node in CaMa-Flood). The water extent is then post-processed using the simulated profile of the water level and water extent for each unit catchment. The simulation is at the acquis-resolution of MERIT Hydro (0.1-arcdegree or 0.25-arcdegree), while the water extent and water level can be downscaled to high-resolution (e.g., ~90m, ~1km) with a simple water balance method. CaMa-Flood has been in continuous development from its initial publication in 2011 (Yamazaki et al., 2011), including in model structures of bifurcation (Yamazaki et al., 2014), dynamic seawater level (Eilander et al., 2020; Ikeuchi et al., 2015, 2017), dam operation (Hanazaki et al., 2022), and in model parameters of river channels (Liang & Zhou, 2022).

Therefore, there has been a significant demand for CaMa-Flood for the benchmark system to evaluate implementations and model developments. On the other hand, as an independent global hydrodynamic model, CaMa-Flood is driven by runoff from other sources, and it is adaptable to runoff inputs at different spatial and temporal resolutions. There are many other implementations to investigate sensitivity to various runoff inputs using CaMa-Flood (Zhao et al., 2017; Zhou, Ma, et al., 2021), while only discharge is evaluated. In this study, we will evaluate CaMa-Flood for all three variables, i.e., Q, WSE and WSA, among simulations driven by various runoff inputs.

3.2 Model settings (sample benchmarking)

In principle, the benchmark system can be applied to compare any different scenarios (e.g., model inputs, model structures, model parameters). Here, we show the implementation with different driven runoff inputs to CaMa-Flood since runoff inputs are the primary source of uncertainties (Zhou, Ma, et al., 2021). Three different runoff inputs are prepared and compared. The E2O_ECMWF is from the eartH2Observe (e2o) wrr2 project (Schellekens et al., 2017). The runoff was driven by the WATCH Forcing Data methodology applied to ERA-Interim data (WFDEI; (Weedon et al., 2014)) with the Tiled ECMWF Scheme for Surface Exchanges over Land incorporating land surface hydrology (H-TESSEL). ERA5 uses the same hydrological model H-TESSEL with an updated version but driven by ERA5 climate reanalysis (Muñoz-Sabater et al., 2021). The third, VIC-BC, is driven by MSWEP precipitation input with the Variable Infiltration Capacity (VIC) model (Yang et al., 2021). However, additional bias correction was applied with a quantile correction approach to the runoff generated by the machine learning method (Beck et al., 2015). Among the three model settings, simulation driven by e2o_ecmwf is set as the baseline model (m0), ERA5 as m1 and VIC_BC as m2 to generalize the descriptions and visualization.

Despite the differences in the runoff inputs, all other model settings in CaMa-Flood (e.g., spatial resolution, river maps, channel parameters) remain the same. Note that the spatial resolution of CaMa-Flood is 0.1 degrees, which is finer than the three inputs; linear interpolation was applied to all inputs. The evaluation period is set as 2001-2012 to better use the overlapped period of different observations.

3.3 Model-observation mapping in CaMa-Flood

3.3.1 Advanced allocation strategy for Q

A two-step advanced allocation is designed for CaMa-Flood which is built on MERIT-Hydro river network. The first step is to allocate it at a high-resolution map (i.e., 1min MERIT Hydro). This is to move large errors in reported gauge attributes, or discrepancy of the digitized river maps (e.g., treatment of bifurcation or depressions in MERIT Hydro). We applied the search strategy in this step using the information of upstream and the distance from the original location. In short, the grid with the minimal error of the upstream area and the least shift in distance will be selected as the allocated location. This step is “semi-automated” because the errors are probably in the reported data or in the river network map, and we recommend users to carefully check the automatic allocation results and examine the suspicious data. Users can decide to correct the reported information or not to use the erroneous river gauges.

Then, the second-step allocation of the "mismatch-resolved" gauges is conducted to the global river network at a coarse resolution (i.e., 0.1deg. This step is to ensure the observations are correctly compared with simulations at specific node. Depending on the correspondence between the coarse-resolution grid and the gauging station, the secondary allocation of these gauges is performed in the following three types (Fig. 3).

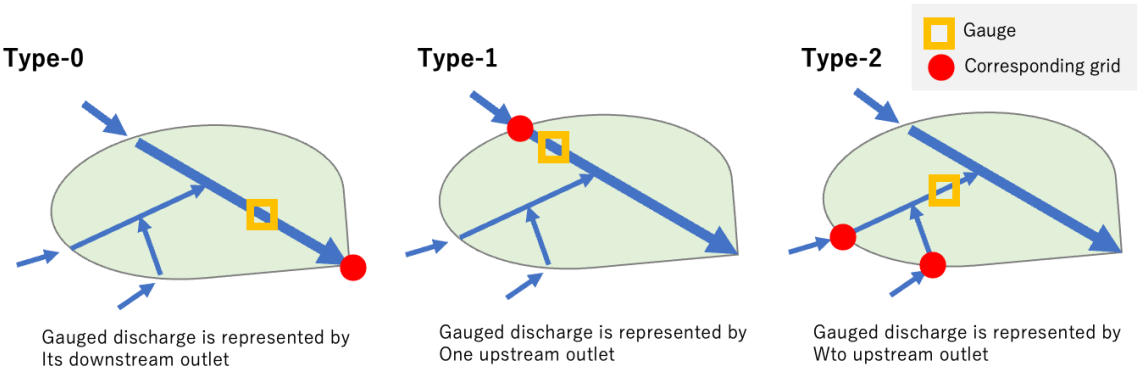

Figure 3. Illustration of allocating river discharge gauges to the unit catchment. The yellow square is the river gauge location. The red dots are outlets of the unit catchment.

- Type-0: Gauge is on the mainstem nearby the catchment outlet. Gauged flow can be reasonably compared to the modelled discharge at the catchment outlet.
- Type-1: Gauge is on the mainstem or one tributary, but nearby the upstream edge. One tributary is merging downstream of the gauge. In this case, Gauged flow is better to be compared against the modelled discharge of the upstream grid.
- Type-2: Gauge is on the tributary, and the tributary has two major upstream tributaries. In this case, gauged flow is better to be compared to the summation of the simulated discharge at two upstream grids.

The second step of allocation outputs the type of gauge-grid correspondence and the grid-coordination of the corresponding grids. The primary corresponding grid coordinate is saved as [ix1,iy1], and secondary corresponding grid coordinate is saved as [ix2,iy2] if any (otherwise marked as [-9999,-9999]).

3.3.2 Advanced allocation strategy for WSE

Advanced option can be applied as well to the allocation of VSs, by considering sub-grid river structure (e.g., correction for bifurcation section) which has been explained in detail as AltiMap (Revel et al., 2024). In short, the relationship of specific VS and the river network is checked for all VSs. And the principle idea of AltiMap is to allocate the VSs to the nearest largest river section. Among all VSs in the raw HydroWeb data, 71.7% VSs are located correctly in the river channel. While 26.88% VSs are modified slightly from nearby land to river channels and 1.34% VSs are moved across river channels in a multi-channel river system. This step matters the preciseness of VS elevation and the offset calculated in the next step.

CaMa-Flood records the river water level at the outlet for the entire river segment, while VSs are located where satellite ground track crosses the river. This leads to an elevation difference of the

two location (i.e., VS and outlet) and in the water level records. Therefore, the elevation difference is recorded as the offset which should be applied to the time series in later comparison.

Users please refer to Supplementary Text S1 for details about the allocation algorithms.

3.3.3 WSA comparison

WSA simulation in CaMa-Flood is still based on unit-catchment, which is slightly different from regular grids. However, the system aggregates results to 0.5 degree, which will largely eliminate the impact. Moreover, the WSA results are compared on a monthly scale, which is different from the other two variables.

3.3.4 Utilization

We integrated the allocation codes in the benchmark system, while we need note that the advanced allocation strategies are specific for CaMa-Flood (or MERIT-Hydro river network). The allocation results also change if using different spatial simulation resolution. Therefore, we provided the sample results at the spatial resolution of 0.1 degree. Users with other demands can follow the instructions and do the allocation by themselves. For river maps other than MERIT-Hydro (e.g., HydroSHEDS), users need to allocate the VSs map accordingly with the provided VSs coordinates.

4 Results

The first sample implementation of the benchmark system, which investigates the impact of various runoffs, will be illustrated here, as well as how the benchmark model works and what information the benchmark system delivers.

4.1 Intermediate data

4.1.1 Reformatted data

The intermediate data documents the full records of different variables at investigating points/grids. The records in time series can be visualized if specifying a particular location. For instance, the variability of Q (Fig. 4a), WSE (Fig. 4b) and WSA (Fig. 4c) around the Obidos in the Amazon River basin are displayed, respectively, as examples. These intermediate data are then used for further evaluation, saving time to read original large simulation outputs.

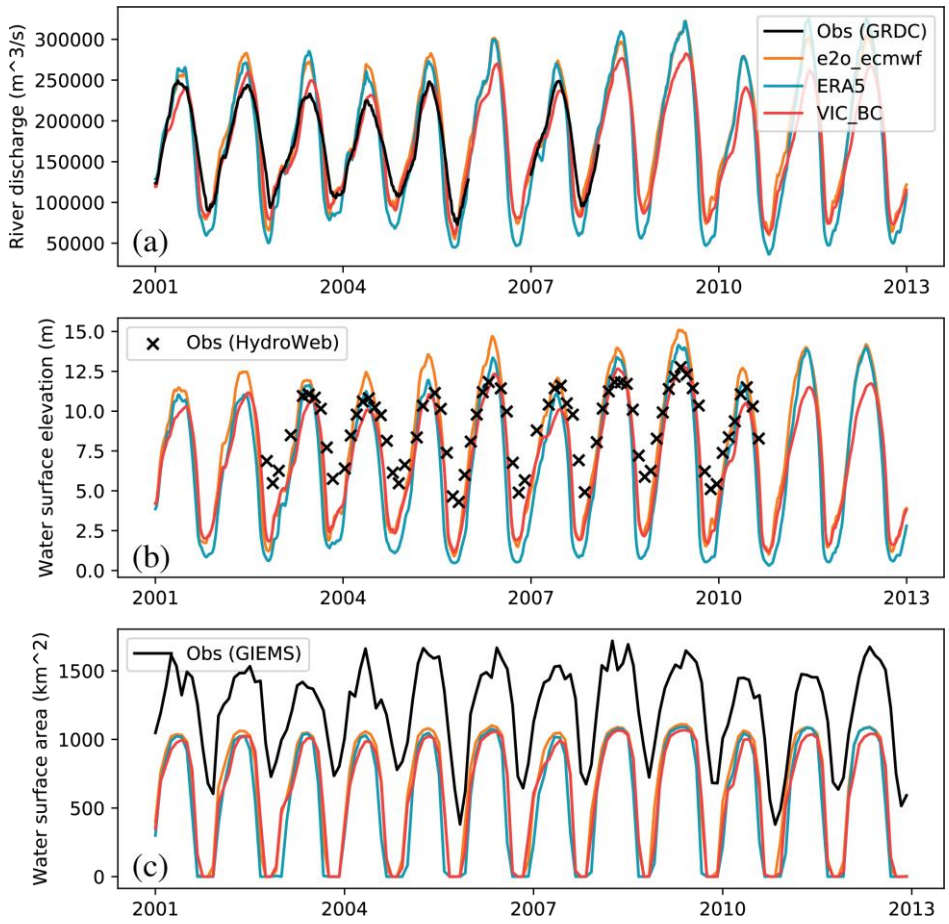

Figure 4. Time series of river discharge, WSE, WSA at Obidos, Amazon River basin, in different scenarios. The water surface area represents the total water area at the 0.5-degree grid near Obidos.

4.1.2 Evaluation Metrics

All evaluation metrics over the various variables are calculated with integrated standard algorithms and temporally saved into intermediate files. These files are prepared for further visualization and easier to be shared for intercomparisons in case of any restrictions on raw data.

A very brief summary of the evaluation metrics is listed in Table 2, showing that there are large discrepancies in the magnitude of investigated variables or metrics. For Q, VIC_BC shows better performance than the other two for all three metrics. While for WSE or WSA, e2o_ecmwf occasionally has better performance. Nevertheless, it is very difficult to conclude the relative preference of a single model from various metrics, and more advanced assessment is needed.

505 **Table 2. Evaluation metrics of the sample region - Obidos**

Variables	Models	p_bias(%)	correlation	rRMSE(%)
Q	e2o_ecmwf	3.7	0.938	14.8
	ERA5	-4.7	0.932	18.5
	VIC_BC	2.4	0.956	9.60
WSE	e2o_ecmwf	-11.1	0.931	26.4
	ERA5	-27.1	0.906	38.5
	VIC_BC	-20.2	0.957	25.4
WSA	e2o_ecmwf	-47.5	0.850	51.2
	ERA5	-55.7	0.822	59.6
	VIC_BC	-51	0.890	53.3

506

507 ## 4.2 River discharge

508 ### 4.2.1 Gauge evaluation

509 The global map of the model performance at gauges clearly shows the spatial variation of the
510 model performance (e.g., the *kge* value of Q in Fig. 5, the first row) and the comparisons against
511 a baseline model (e.g., the second and third row in Fig. 5). The range of color bar is pre-defined
512 for each metric. Although, users can modify them to increase readability of the figures.
513 Moreover, to make the changes more remarkable, the values close to zero is always set as white,
514 even though the mean value is not zero.

515 In the test case, after selection of the investigated gauges by data coverage, the available GRDC
516 gauges are mainly limited in North America and Europe (Fig. 5). Higher *kge* in green color can
517 be found in a large number of gauges, especially in lower North America, Europe, upstream of
518 Amazon and a few gauges in Asia and northern Australia. Simulations with all three runoff
519 inputs show poor results in the arid mountainous area (e.g., Rocky Mountains, Southern Andes),
520 indicating that either all runoffs are not well reproduced, or the routing process of CaMa-Flood
521 needs improvement over such regions. However, according to the *pbias* of the runoff, which is
522 less determined by the routing process (which mainly changes the timing rather than the
523 amount), all the current three experiments overestimated the runoff, probably due to
524 underestimation of evapotranspiration in those regions.

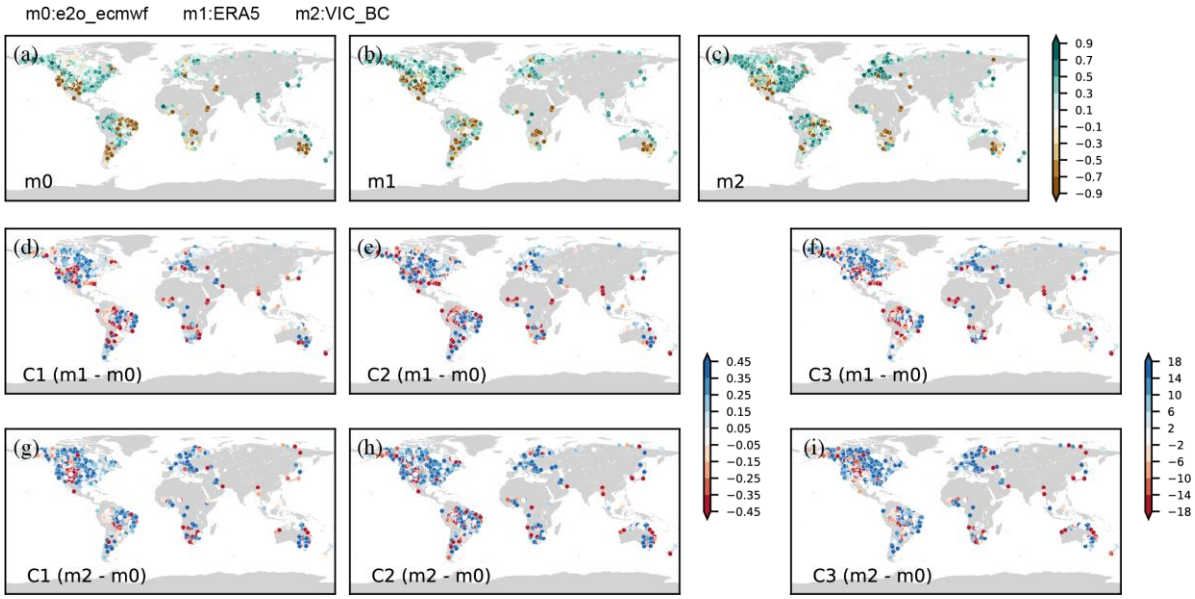

Figure 5. Sample of maps of the *kge* value for river discharge at available GRDC river gauges. The first row (a,b,c) is the evaluation of three simulations against river discharge. The second row is the comparison of the evaluation metrics between m1 (ERA5) over m0 (e2o_ecmwf), and the third row is the comparison between m2 (VIC_BC) over m0 (e2o_ecmwf). Other maps for various evaluation variables (e.g., *pbias*, correlation) can be found in the supplementary as Fig. S1, Fig. S2.

Fig. 5d-5i shows the changes in *kge* metric over the baseline model. ERA5 and VIC_BC significantly improved the model in Canada, the central and eastern US, and Europe (colors in blue). Abundant ground observations of land surface components, which have been used in the assimilation for ERA5 and bias correction in VIC_BC ensured the improvement of runoff simulation. In contrast, models driven by ERA5 or VIC_BC deteriorate in the southwestern US, where discharge itself is always at a low level. Deterioration is also found in southern Africa, where little data can be achieved for assimilating or bias correction in the ERA5 or VIC_BC. The patterns of metrics change for ERA5 (second row) are similar for the three comparison metrics (Fig. 5d,e,f). This is also the case for the VIC_BC (third row). Comparison between m2 (VIC_BC) and m1 (ERA5) is not discussed but will be presented in the next subsection.

4.2.2 Overall evaluation over Q

Fig. 6 is designed to show changes of the metrics from baseline model. The first panel in Fig. 6 is the overall evaluation of the investigated variable over all gauges (shown as median value). The three panels in the right-side hand are shown as the changes of the metrics (shown as the median change for all gauges). From the left to the right are the changes in delta index (C1), improvement index (C2) and percentile index (C3). Darker blue represents the model outperforms the baseline while the deeper red represents the opposite. All metrics are mapped in one figure so that we can easily see how much and for which metric the new model changes.

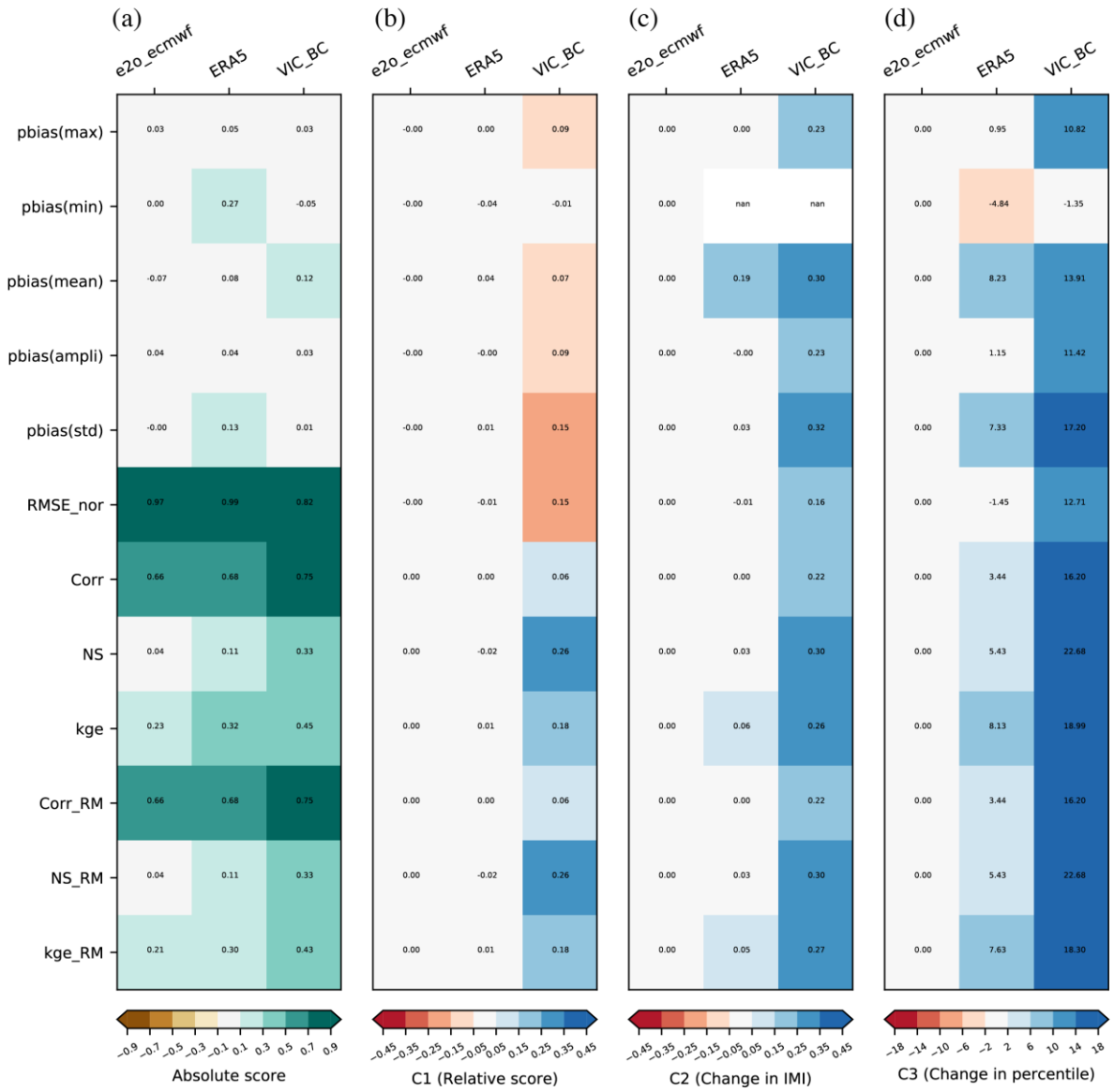

Figure 6. The overall performance of different models (a) and their comparison (b,c,d) in terms of river discharge. a. the median value of different metrics among all river discharge gauges. b,c, and d shows the changes in the metrics of the new model (m1,m2) compared to the benchmark m0, representing the relative score, change in IMI and change in percentile, respectively.

Fig. 6a shows the case for Q against observations. For e2o_ecmwf, the *pbias* has been well controlled within 10%, not only for the mean but also for the maxima, minima and relative amplitude. The *rRMSE* is 0.97 for e2o_ecmwf. The correlation is relatively good (median value 0.66), while the *NS* and *kge* values are at a low level (i.e., 0.04 and 0.23, respectively). In terms of the changes over the baseline model shown in the right panels, we see tiny changes in ERA5 over e2o_ecmwf using C1 (Fig. 6b) and C2 (Fig. 6c). While for C3 (Fig. 6d), positive changes in the percentile index (blue) mean the metrics for the new models improved to a better position

(lower values for metrics in Category 1 and higher values for metrics in Category 2). Based on Fig. 6b-d, the model driven by either ERA5 or VIC_BC has been generally improved, with a more significant improvement for the model driven by VIC_BC. Among all the metrics, the model didn't show improvement in the minimum value, indicating that either those runoff inputs are not improved in the low flow or because the CaMa-Flood streamflow is not responding correctly to changes in the runoff.

4.3 Overall evaluation over WSE

Very similar analysis can be conducted over variables of WSE, using the same visualization tools. Compared to discharge gauges, there are more virtual stations distributed in Asia and Africa due to the selection of time period. Model performance in Europe (and Amazon) and central Africa is relatively acceptable with a positive *kge* value, which overall evaluates the correlation, variability and mean values. Based on the results shown in the supplementary (Fig. S3-S5), the variability is the critical factor that deteriorates the model performance, especially in Asia. The variability is generally larger in simulations than in observations. Given that the river discharge and mean water elevation are simultaneously simulated, the river width could be larger, and the river depth could be shallower. This might be because we assume the river channel shape is rectangular while many small rivers are in other shapes depending on the geographic factors.

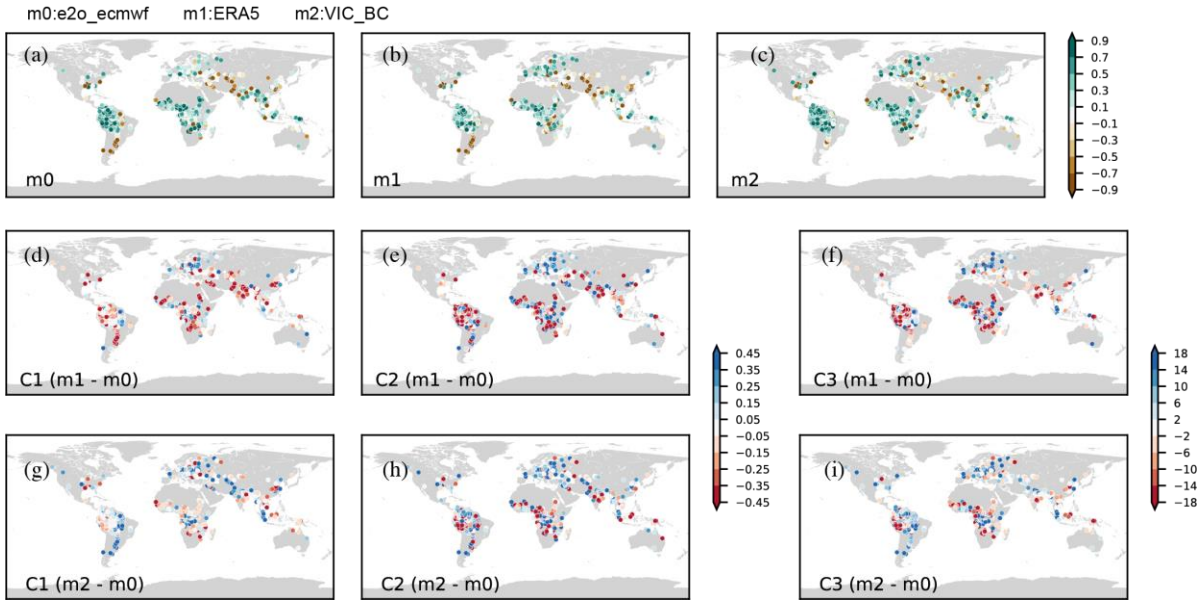

Figure 7. Sample of maps of the *kge* value for WSE at available HydroWeb virtual stations. The illustrations of the figure are same as Figure 5.

Fig. 8 evaluates the model performance with different metrics for WSE. The amplitude (maxima – minima) and variability are large with all three runoff inputs. Compared to e2o_ecmwf, ERA5 runoff increased the bias, leading to an increment in the relative score and percentile. The *NS* value is low, which is caused by the systematic bias in the simulation and observations. *NS* increased to a positive value when systematic bias was removed (*NS_RM*). Among the three runoff inputs, the model driven by VIC_BC shows better performance of WSE than e2o_ecmwf

(Fig. 8d, with blue colors), although the change is less significant than that for Q. ERA5 is worse (with red-ish colors), which is in contrast with conclusions from Q. One major reason is that the global distribution of the virtual stations and river discharge gauges is different. There are main gauges in the US where ERA5 or VIC_BC show improved model performance, while there are very few gauges in the WSE observations. Local factors for instance the accuracy in river cross-section profile, slope, DEM will also cause unpredictable impact on the WSE simulation. Nevertheless, VIC_BC shows obvious improvements in both Q and WSE, compared to the other two runoff inputs.

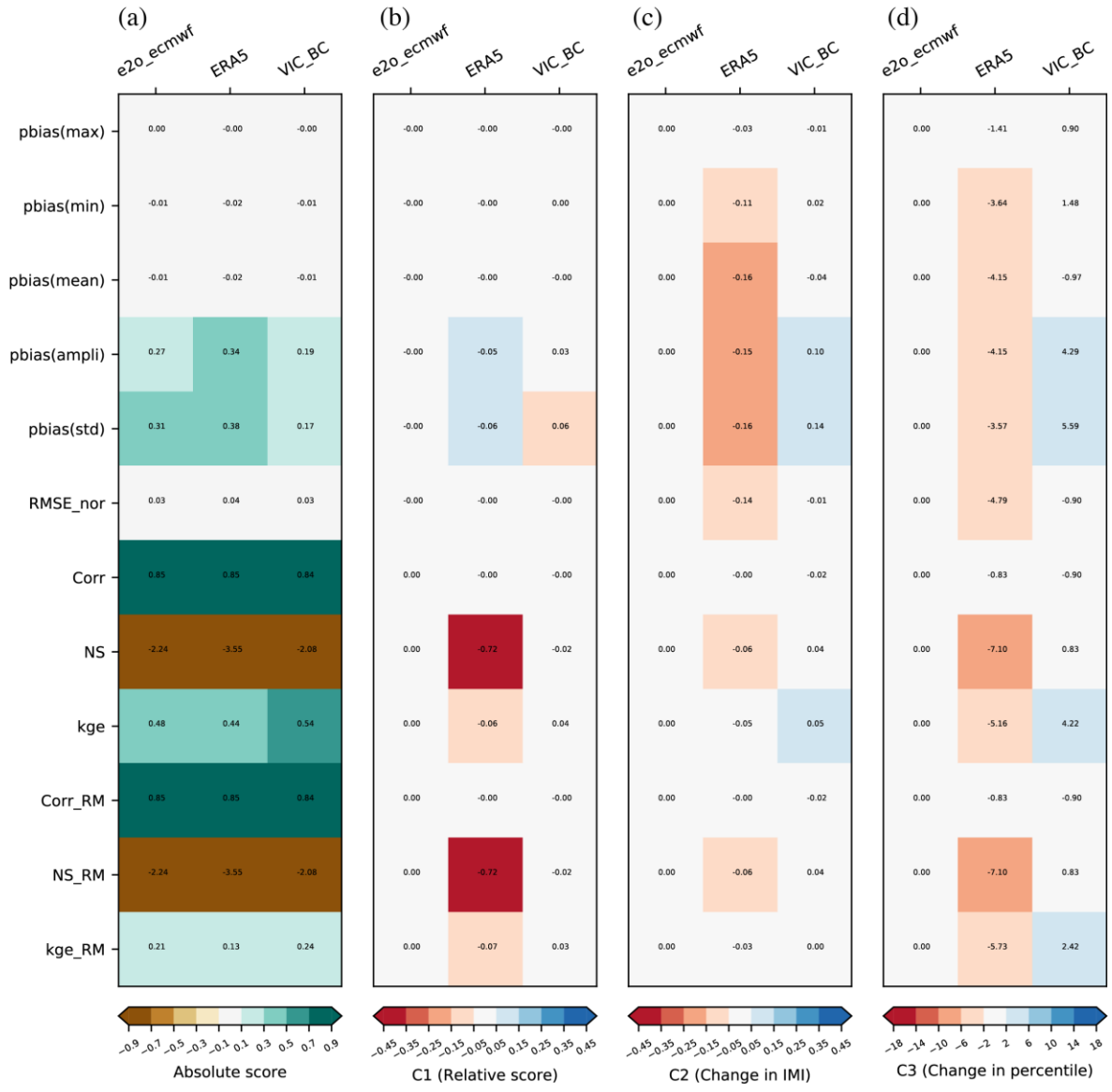

Figure 8. The overall performance of different models (a) and their comparison (b,c,d) in terms of WSE.

4.4 Overall evaluation of WSA

Comparing time series of WSA presents considerable challenges for several reasons. Firstly, the comprehensive validation of remotely sensed WSA accuracy is still lacking. The discrepancies between model simulations and sensing data, particularly in areas with smaller water surfaces, have not been adequately addressed. The GIEMS data, despite its coarse spatial resolution, tends to overestimate WSA, as it often misclassifies saturated soils as water surface. Therefore, while we have conducted model comparisons against WSA data, a more thorough analysis is needed to extract valuable insights. WSA is 2-D map data, therefore, we assessed it at each grid (using same codes for Q and WSE) and showed the results on the map in the same way. Those grids with very low mean WSA (either simulated or observed value is zero) are excluded since the results in those grids are very sensitive.

Fig. 9 illustrates the global evaluation of WSA, with the *kge* values for three different models generally falling below 0. Although the model performance is suboptimal, it still allows for a broader evaluation across various regions worldwide. Utilizing the evaluated grids (where both observed and simulated WSA exceed zero), the overall comparison is presented in Fig. 10, revealing no significant variations in state metrics across different models. However, the VIC_BC model stands out by providing more detailed process information compared to the other two models.

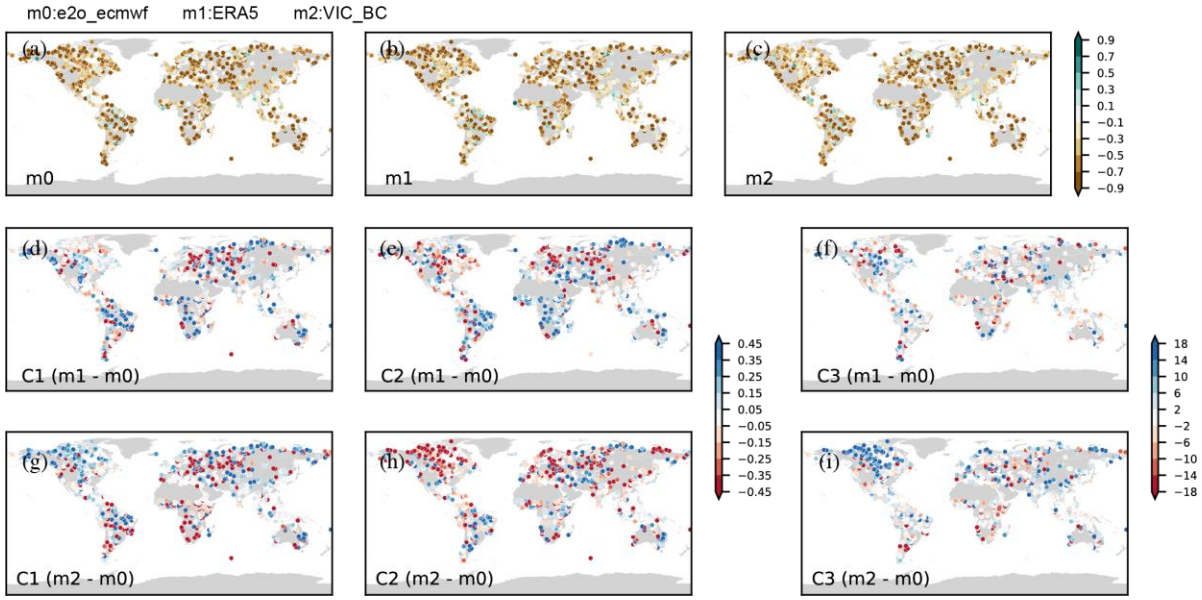

Figure 9. The global distribution of evaluation metrics (herein *kge* value) and comparisons among different models for the WSA. Note that the grids with either estimated or observed WSA as zero are excluded from the evaluation.

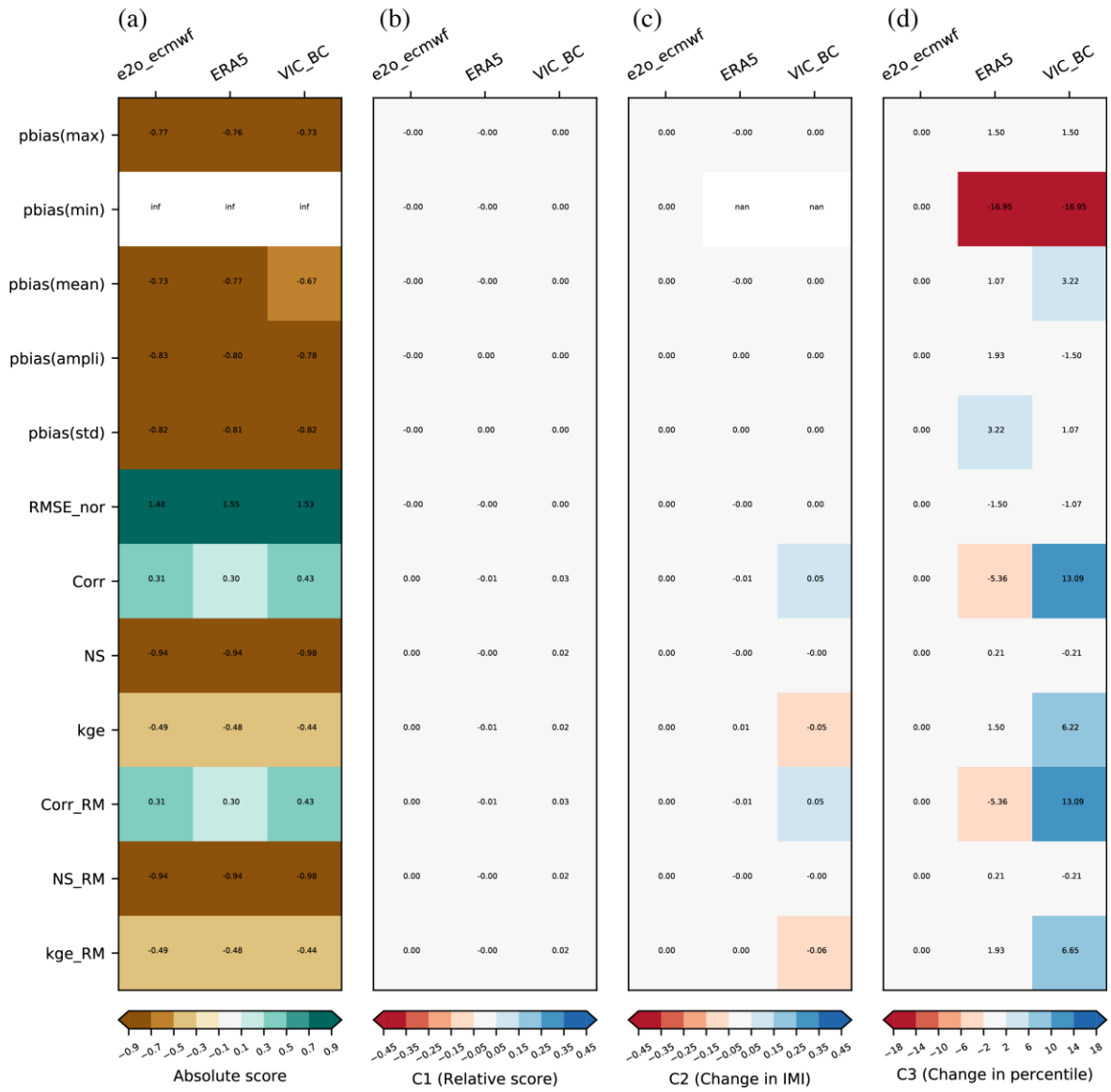

Figure 10. The overall performance of different models (a) and their comparison (b,c,d) in terms of WSA. The value is calculated as the median value for all grids which are included in the visualization map in Figure 9.

5 Discussions

5.1 Data coverage

An overview of how three data overlap in space is shown as Fig. 11. Following the rules of gauge selection, only 929 grids (0.5 by 0.5 degree) are monitored with discharge gauges. Using WSE VSs has increased almost 100% of the grids where water dynamics can be analyzed. Though, only 74 (<10%) of the grids have both Q and WSE observations. WSA observation is the most sufficient way to overcome the data scarcity problem, while only ~5% of the grids with

larger than 10km^2 have observations of Q or WSE or both (34 grids). Although we assume that the water dynamics are correlated for the three variables, investigation over the grids with more than one kind observations is needed. The three variables also have their own strengths and shortcomings which are relevant to the accuracy, spatial resolution, temporal intervals. One cannot replace the other in most occasions. Therefore, integrating three variables is sometime difficult but very important for large-scale analysis.

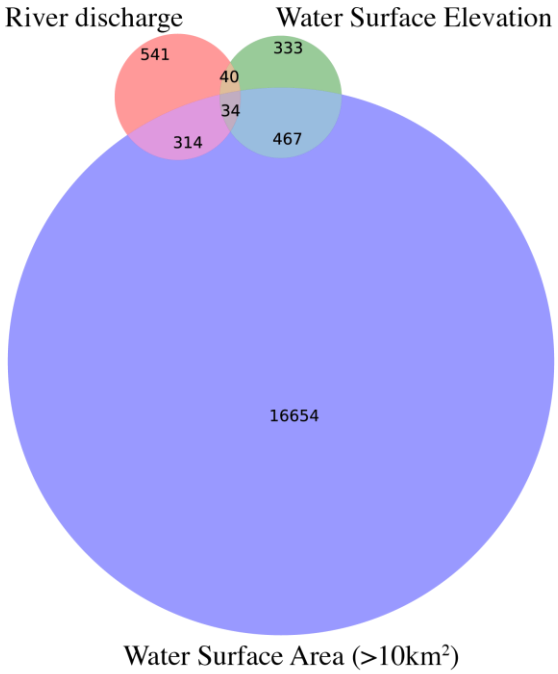

Figure 11. Statistics of the grids (0.5*0.5 degree) with various observational data. The number value shows the total number of grids with one kind of observation, no matter how many gauges are located in the grid. The overlap areas show the number of grids with two or three kinds of observations.

The constantly evolving technologies offer opportunities to incorporate new and advanced data sources into our system. For instance, laser-altimetry data from ICESat2 (Parrish et al., 2019) and higher spatiotemporal resolution water extent data from sources like MODIS (Ji et al., 2018), and Sentinel-3 (Jiang et al., 2023) can enrich our evaluations. However, despite the spatial coverage, the revisit time also matters in evaluating flow dynamics. For instance, ICESat2 has a larger surface coverage but its limited revisit frequency (i.e., 91days) determines that its improvement in short-term flow events will be limited. Satellite discharge algorithms (~2weeks, mostly based on Landsat, (Lin et al., 2023)) provides adequate frequency but not-so-high accuracy. Surface Water and Ocean Topography (SWOT, (Biancamaria et al., 2016)) mission ensures a weekly revisit time (7-10day) for most land area with wide swath altimetry, thus, has a great potential to be used in improving the benchmark system and future model developments.

5.2 Comparison metrics

In this study, we proposed three metrics (C1, C2, C3) to quantify changes in model performance. In Fig. 12, we illustrate the comparisons of *kge* value (Fig. 11a) and the differences among using three metrics (Fig. 12b-d). The color represents the upstream area of the investigated gauge, with the deeper color representing the larger upstream area. In general, the model performance driven by two runoff inputs is correlated (i.e., e2o_ecmwf and VIC_BC, Fig. 12a). When using the first relative score (C1), it can be found that when the baseline performance is good (m0, for example, >0.5), the compared model cannot show a distinguishable difference from the reference. The most obvious model improvements or deteriorations are mainly reflected when the baseline was worse (Fig. 12b, for example, <-1.0). Therefore, the change in the metric value mostly highlighted changes when baseline model is bad. In terms of the improvement index (C2), it evaluates the degree of change from the baseline value relative to the best value (i.e., *kge*=1). It is found that the largest improvement can be achieved when the baseline value is between 0 and 0.5 (Fig. 12). However, it is difficult to distinguish the magnitude of changes since points are very concentrated. Moreover, for gauges with larger baseline values, even if the simulation metric (*kge*) degrades in a small value, a significant degradation metric can be generated. Although the advantage is that the effect of the same change amount is related to the original position, it is more sensitive in gauges with larger initial values. In addition, both the previous metrics have the limitation that they are too concentrated in certain areas but too sparse in most areas. This makes it difficult to grasp the overall performance because changes in certain gauges will significantly vanish the results.

Regarding the third percentile index, the points are more evenly distributed over the entire space. This will eliminate the impact of certain extreme values when calculating the mean value over all the gauges. On the other hand, the percentile evaluation is insensitive to the magnitude or distribution of the evaluation metric itself or for different variables. Thus, it is better to be used for integrating different metrics together to show overall performance.

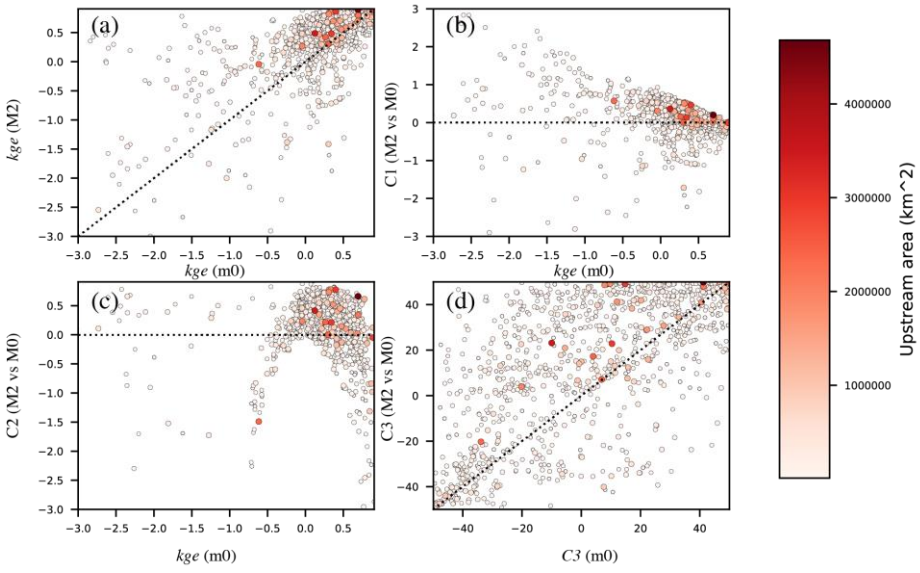

Figure 12. Comparisons between changes in comparison metric with against original value. (a) compares the original value of *kge* for the last model (m2, i.e., VIC_BC) and the baseline

model (m0). (b)-(d) shows the comparison of the difference between the last model and the baseline model with regard to the three different metrics.

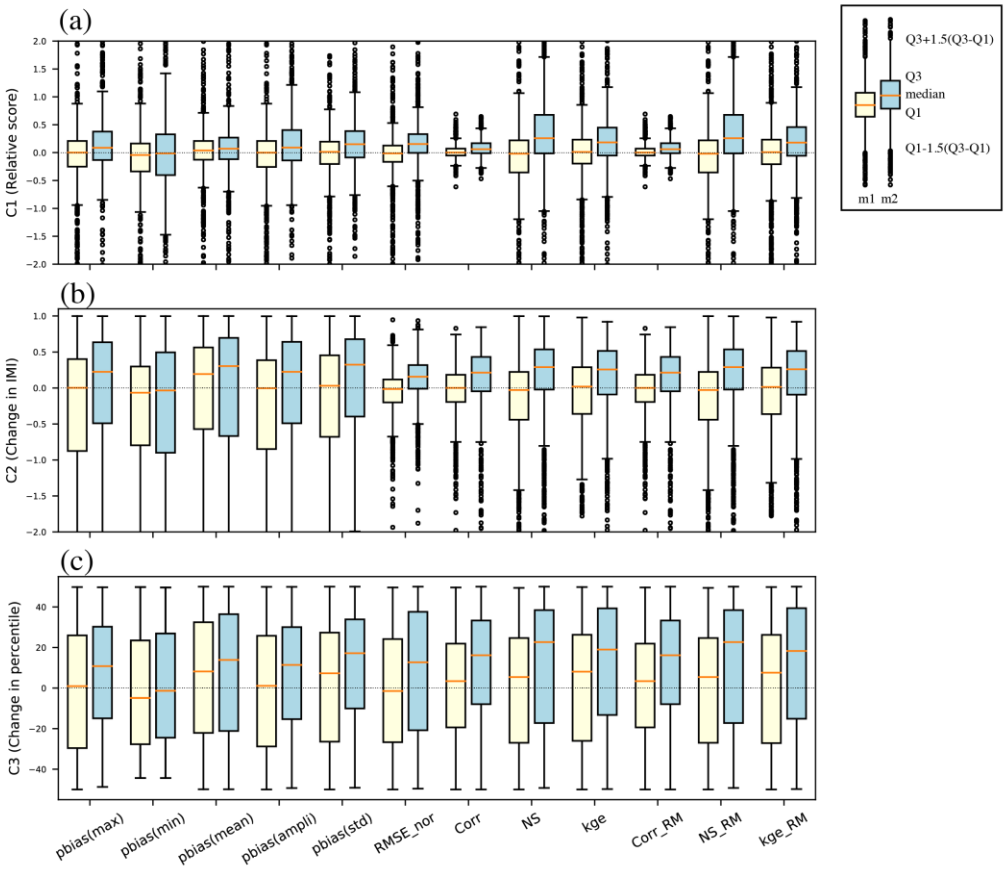

Figure 13. Boxplots of three comparison metrics for all performance indices. Outliers are identified as out of range [Q1-1.5*(Q3-Q1), Q3+1.5*(Q3-Q1)]. The yellow color represents comparisons between m1 and m0. The blue color represents comparisons between m2 and m0.

Fig. 13 illustrates the characteristics of comparison metrics, presenting median values, ranges, and outliers. The medians reflect previous colormaps for various variables, providing additional insights into the distribution of comparison metrics. In Fig. 13a, a notable proportion of comparison metrics are identified as outliers for the first and second comparison metrics. Regarding the first metric, both correlation and its normalized form exhibit a highly constrained range. Conversely, for the second comparison metric (Fig. 13b), the upper limit is 1, suggesting a model improvement to a perfect status, which is not realistically attainable. All metrics display large ranges of negative values, indicating a deteriorated model performance. However, a challenge arises in selecting an appropriate metric, as aggregating multiple metrics remains sensitive to a few extreme values, potentially leading to a misleading overall evaluation.

Concerning the third comparison metric (Fig. 13c), the constrained range of metrics falls between -50 and 50, with consistent width in the quantile range. The median values align with the overall colormaps depicted in previous figures, showcasing similar patterns for different metrics. This underscores the insensitivity of selecting a smaller number of metrics or

aggregating metrics for various evaluated variables to achieve an overall evaluation across models.

5.3 Observational data preprocessing

The benchmark system relies heavily on having sufficient observational variables to effectively validate and calibrate river models. Unfortunately, data shortages have been a limiting factor in this process, which subsequently hampers the wider application of river models. To address this issue, our study seeks to overcome data limitations by synthesizing in-situ river discharge (Q) and remote sensing data on water surface elevation (WSE) and water surface area (WSA), enabling assessments over larger and data-scarce regions in Asia and Africa.

To ensure more accurate evaluations, the gathered observations require preprocessing. In section 2.1, we applied an allocation method for river discharge and altimetry observations. However, a challenge arises when other users wish to apply our system with different underlying river networks (e.g., MERIT Hydro). To address this, we recommend that users follow specific instructions to prepare the allocation based on the river network they used in their simulations. While the error in allocating river discharge is generally small due to limited discharge changes within a short distance, users should be cautious when dealing with gauges near river confluence channels.

Allocating altimetry data presents a more complex issue, as significant changes in water surface elevation can occur along rivers within a short distance. This has a notable impact on mean values, subsequently affecting evaluation metrics that rely on mean status (e.g., NS). However, correlation-based metrics remain unaffected by this allocation challenge. The benchmark system still works if only the dynamic WSE is evaluated when the offset is not applicable.

When comparing WSA, we acknowledge limitations related to the nature of the MERIT unit-catchment data, which may not perfectly match grid boundaries. To mitigate this, we conducted simulations at a 0.1-arcdegree resolution but aggregated them to 0.5-degree resolution for comparison, resulting in acceptable evaluation results. Nevertheless, when assessing WSA at a local scale, we recommend users perform downscaling before conducting comparisons.

5.4 Future applications

As mentioned in the Introduction, this benchmark system serves as a valuable tool for evaluating the development of various models. In our study, we tested the system with different forcing inputs to CaMa-Flood, and in the supplementary material, we provided a case study by testing both kinematic wave and dynamic wave equations in hydrodynamic simulations. We also conducted comparisons to test different parameters, such as river bankfull height. It is essential to consider the varying sensitivity of evaluation metrics or variables based on different model developments. For instance, forcing inputs have a more significant impact on river discharge but a smaller impact on water level simulations. Meanwhile, water level simulations are more sensitive to river bankfull height variations.

One critical aspect to keep in mind is that the evaluation relies on observational data that may not be evenly distributed across regions. Consequently, the weight of comparisons should differ based on data availability and regional variations. For example, forcing inputs could cause large

variations in Asia and Africa because very limited observations are there. To enhance the robustness of the benchmark system, efforts should be made to collect more data and ensure its even distribution. Both governmental organizations and the scientific community can play a vital role in data collection for scientific research.

Apart from assessing model development, this benchmark system allows for the intercomparison of various global river models. However, there is a lack of standard datasets, and some models may have limitations in simulating water level and water area variables, which hinders more comprehensive comparisons. However because the evaluation over three variables is independent, the system can evaluate hydrological models that only provide discharge simulations. For instance, we compared our physically-based simulations with river discharge forecasts from the Google Flood Initiative (which are based on machine learning, (Nearing et al., 2024), Fig. S6). We are also able to compare river discharge simulations from the Global Flood Awareness System (GloFAS, Fig. S7) or The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). It is also feasible to compare routing simulations from other open-source runoff achieves, e.g., the Coupled Model Intercomparison Project (CMIP), for past and future periods.

Regarding data-sharing policies, we regret that we cannot share all observations with users. However, we offer two solutions to address this limitation. First, users can prepare the necessary observations following the provided instructions. Second, users can share their raw simulations or intermediate files (e.g., extracted raw simulations at specific gauges) to reduce data size and protect their information from leaking. We are committed to conducting evaluations of their results and comparing them to any available baseline to support their model assessments. We also have shared the evaluation results, which can be used by others for their own comparison.

5.5 Factors affecting the evaluation

In this subsection, we delve into the factors that can potentially influence the evaluations. When assessing water levels, our approach assumes that the allocation bias primarily stems from the elevation difference between the virtual station and the outlet of the unit catchment. However, this approach overlooks the intricate water dynamics within the channel itself, given that factors such as river width and slope undergo significant changes. Consequently, the flow dynamics at the visual station and outlet often diverge, rendering the removal of elevation differences does not necessarily eliminate the allocation errors. Addressing this issue demands a more extensive dataset that encompasses elevation data with finer resolution, coupled with a thorough understanding of river channel bathymetry variabilities, in order to enhance the precision of bias calculations. In the current context, downscaling the simulated water level could enhance the reliability of comparisons, although this comes at the cost of computational demands, particularly in the context of daily global-scale assessments. To achieve accurate comparisons, it remains imperative to enhance simulation capabilities in small-scale hydrodynamics. Moreover, the utilization of various elevation data and river networks will also affect the assessment, while the correct gauge allocation will to some degree alleviate the impact.

Furthermore, the assessment of water surface area at a broader scale still lacks a comprehensive and rigorous analysis. The precision of this assessment is curtailed by limitations inherent in various satellite sensors. For instance, optical sensors are impeded by dense cloud cover and

vegetation, whereas microwave sensors tend to overestimate water extent in areas with saturated soils. Moreover, the spatial resolution of these sensors dictates their efficacy in capturing various forms of water extent. Notably, the MODIS data, with a resolution of 250 meters, struggles to identify water bodies narrower than this threshold. The revisit intervals of satellites further compound the challenge, leading to the potential rapid events of short duration (e.g., flash floods). Simultaneously, different models introduce their own assumptions and exhibit varying capabilities in modeling distinct types of water surface area. Consequently, it is essential to meticulously consider the congruence between comparison types. In this study, we outline a framework for evaluating water surface area. This framework accommodates future updates in observational data, with the intent of facilitating more robust and refined comparisons of water surface area in subsequent analyses.

5.6 Expendability

The simple structure of the benchmark system allows the users to modify and extend the system easily. For instance, new evaluation metrics can be designed to evaluate various flow dynamics of interest (e.g., flow peaks, peak timing, baseflow). The regional maps can be easily customed by modifying parameters in the codes. Accompanying the information for categorization (e.g., continents, river size, climate zones), a more detailed analysis of how different factors affecting the evaluation can be assessed.

6 Conclusions

This study introduces a novel benchmark system designed to assess and compare the performance of global flood models. We proposed and established methodologies which can systematically integrate river-related observation data for global river model benchmark system. By integrating in-situ river discharge observations and remote sensing data, such as water surface elevation and water surface area, the system enables the evaluation of model performance in areas with limited ground observations. Notably, this approach allows for the evaluation of river models that simulate flow dynamics on a large scale, providing valuable insights for enhancing global river modeling.

The benchmark system employs a range of metrics to evaluate and compare various aspects of model performance, each representing different facets of model capability. A novel percentile comparison metric has been developed to offer a comprehensive assessment of model changes (e.g., improvements or deteriorations). This metric, combined with traditional methods, facilitates a comprehensive understanding of how the model evolves over time.

The versatility of the benchmark system is underscored by its capacity to assess model development and conduct intercomparisons across multiple models using diverse outputs related to flow dynamics. It is important to note that the success of this system relies on collaborative efforts within the scientific community, particularly in terms of gathering observational data and simulation outputs. Adding new metrics for evaluation is also a possibility according to users' needs. In pursuit of greater robustness and adaptability, the benchmark system acknowledges the need for innovation in data collection and analysis. By addressing the challenges and embracing emerging data sources, the system aspires to enhance its effectiveness in contributing to global

river modeling and management practices. The study also extends an invitation to users for their valuable feedback, contributions of comparative data, and advancements in code development.

Open Research

The codes for the benchmark system are shared under CC-BY-4.0 license. User can find the source code from https://doi.org/10.5281/zenodo.10903211 (Zhou et al., 2024). The instruction manual of system settings and execution is included. The raw data for the sample case can be shared by request due to its large size. The statistical results for this case study are shared in the source codes, thus, uses can reproduce most figures included in this study and for their own comparisons. The allocation algorithm is shared independently in the repository https://doi.org/10.5281/zenodo.10893741 (Yamazaki, 2024).

Acknowledgements

We acknowledge the supports from KAKENHI 20K22428 by Japan Society for Promotion of Science (JSPS), JP21500379 from New Energy and Industrial Technology Development Organization (NEDO), SIP3 project (JPJ012289) by CITI, and Google. We thank Asher Metzger and Grey Nearing and for their advice and financial support to this research and data from Google Flood Initiative. We thank Abdul Moiz for his help in reviewing the codes.

References:

Aires, F., Miolane, L., Prigent, C., Pham, B., Fluet-Chouinard, E., Lehner, B., & Papa, F. (2017). A global dynamic long-term inundation extent dataset at high spatial resolution derived through downscaling of satellite observations. *Journal of Hydrometeorology*, *18*(5), 1305–1325. https://doi.org/10.1175/JHM-D-16-0155.1

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

Bernhofen, M. V., Whyman, C., Trigg, M. A., Sleigh, P. A., Smith, A. M., Sampson, C. C., Yamazaki, D., Ward, P. J., Rudari, R., Pappenberger, F., Dottori, F., Salamon, P., & Winsemius, H. C. (2018a). A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique. *Environmental Research Letters*, *13*(10), 104007. https://doi.org/10.1088/1748-9326/aae014

Bernhofen, M. V., Whyman, C., Trigg, M. A., Sleigh, P. A., Smith, A. M., Sampson, C. C., Yamazaki, D., Ward, P. J., Rudari, R., Pappenberger, F., Dottori, F., Salamon, P., & Winsemius, H. C. (2018b). A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique. *Environmental Research Letters*, *13*(10). https://doi.org/10.1088/1748-9326/aae014

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van Den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., & Vuichard, N. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Biancamaria, S., Lettenmaier, D. P., & Pavelsky, T. M. (2016). The SWOT Mission and Its Capabilities for Land Hydrology. *Surveys in Geophysics*, *37*(2), 307–337. https://doi.org/10.1007/s10712-015-9346-y

Burek, P., & Smilovic, M. (2023). The use of GRDC gauging stations for calibrating large-scale hydrological models. *Earth System Science Data*, *15*(12), 5617–5629. https://doi.org/10.5194/essd-15-5617-2023

Chen, H., Liu, J., Mao, G., Wang, Z., Zeng, Z., Chen, A., Wang, K., & Chen, D. (2021). Intercomparison of ten ISI-MIP models in simulating discharges along the Lancang-Mekong River basin. *Science of the Total Environment*, *765*. https://doi.org/10.1016/j.scitotenv.2020.144494

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., & Randerson, J. T. (2018). The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018MS001354

Dingman, S. L. (2015). *Physical Hydrology: Third Edition*. Waveland Press.

Do, H. X., Gudmundsson, L., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth System Science Data*, *10*(2), 765–785. https://doi.org/10.5194/essd-10-765-2018

Eilander, D., Couasnon, A., Ikeuchi, H., Muis, S., Yamazaki, D., Winsemius, H. C., & Ward, P. J. (2020). The effect of surge on riverine flood hazard and impact in deltas globally. *Environmental Research Letters*, *15*(10). https://doi.org/10.1088/1748-9326/ab8ca6

Eilander, D., Ikeuchi, H., Yamazaki, D., Couasnon, A., Winsemius, H., & Ward, P. (2018).

Global fluvial flood modelling – a sensitivity analysis. *Geophysical Research Abstracts EGU General Assembly*, *20*, 2018–8397. https://meetingorganizer.copernicus.org/EGU2018/EGU2018-8397.pdf

Elmi, O., Tourian, M. J., Saemian, P., & Sneeuw, N. (2024). Remote Sensing-Based Extension of GRDC Discharge Time Series - A Monthly Product with Uncertainty Estimates. *Scientific Data*, *11*(1). https://doi.org/10.1038/s41597-024-03078-6

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt, K. D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., … Williams, K. D. (2016). ESMValTool (v1.0)-a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*(5), 1747–1802. https://doi.org/10.5194/gmd-9-1747-2016

Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. *Earth System Science Data*, *10*(2), 787–804. https://doi.org/10.5194/essd-10-787-2018

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., & Tanaka, K. (2008). An integrated model for the assessment of global water resources - Part 2: Applications and assessments. *Hydrology and Earth System Sciences*, *12*(4), 1027–1037. https://doi.org/10.5194/hess-12-1027-2008

Hanazaki, R., Yamazaki, D., & Yoshimura, K. (2022). Development of a Reservoir Flood Control Scheme for Global Flood Models. *Journal of Advances in Modeling Earth Systems*, *14*(3). https://doi.org/10.1029/2021MS002944

Hirpa, F. A., Lorini, V., Dadson, S. J., & Salamon, P. (2021). Calibration of Global Flood Models. In *Global Drought and Flood* (pp. 201–211). https://doi.org/10.1002/9781119427339.ch11

Hoch, J. M., & Trigg, M. A. (2019). Advancing global flood hazard simulations by improving comparability, benchmarking, and integration of global flood models. *Environmental Research Letters*, *14*(3). https://doi.org/10.1088/1748-9326/aaf3d3

Hou, Y., Guo, H., Yang, Y., & Liu, W. (2023). Global Evaluation of Runoff Simulation From Climate, Hydrological and Land Surface Models. *Water Resources Research*, *59*(1), 1–25. https://doi.org/10.1029/2021wr031817

Ikeuchi, H., Hirabayashi, Y., Yamazaki, D., Kiguchi, M., Koirala, S., Nagano, T., Kotera, A., & Kanae, S. (2015). Modeling complex flow dynamics of fluvial floods exacerbated by sea level rise in the Ganges-Brahmaputra-Meghna Delta. *Environmental Research Letters*, *10*(12). https://doi.org/10.1088/1748-9326/10/12/124011

Ikeuchi, H., Hirabayashi, Y., Yamazaki, D., Muis, S., Ward, P. J., Winsemius, H. C., Verlaan, M., & Kanae, S. (2017). Compound simulation of fluvial floods and storm surges in a global coupled river-coast flood model: Model development and its application to 2007 Cyclone Sidr in Bangladesh. *Journal of Advances in Modeling Earth Systems*, *9*(4), 1847–1862. https://doi.org/10.1002/2017MS000943

Ji, L., Gong, P., Wang, J., Shi, J., & Zhu, Z. (2018). Construction of the 500-m Resolution Daily Global Surface Water Change Database (2001–2016). *Water Resources Research*, *54*(12), 10,270-10,292. https://doi.org/10.1029/2018WR023060

Jiang, L., Madsen, H., & Bauer-Gottwein, P. (2019). Simultaneous calibration of multiple hydrodynamic model parameters using satellite altimetry observations of water surface

elevation in the Songhua River. *Remote Sensing of Environment*, *225*(September 2018), 229–247. https://doi.org/10.1016/j.rse.2019.03.014

Jiang, L., Zhao, Y., Nielsen, K., Andersen, O. B., & Bauer-Gottwein, P. (2023). Near real-time altimetry for river monitoring—a global assessment of Sentinel-3. *Environmental Research Letters*, *18*(7), 074017. https://doi.org/10.1088/1748-9326/acdd16

Kettner, A. J., Brakenridge, G. R., Schumann, G. J., & Shen, X. (2021). DFO — Flood Observatory. In *Earth Observation for Flood Applications*. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-819412-6/00007-9

Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., Burrows, R. M., DelVecchia, A. G., Fritz, K. M., Shanafield, M., Burgin, A. J., Zimmer, M. A., Datry, T., Dodds, W. K., Jones, C. N., Mims, M. C., Franklin, C., Hammond, J. C., Zipper, S., Ward, A. S., … Olden, J. D. (2022). Assessing placement bias of the global river gauge network. *Nature Sustainability*, *5*(7), 586–592. https://doi.org/10.1038/s41893-022-00873-0

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., Lorenz, R., Perez-Zanon, N., Righi, M., Schlund, M., Senftleben, D., Weigel, K., & Zechlau, S. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 - Diagnostics for emergent constraints and future projections from Earth system models in CMIP. *Geoscientific Model Development*, *13*(9), 4205–4228. https://doi.org/10.5194/gmd-13-4205-2020

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. https://doi.org/10.1002/hyp.9740

Liang, H., & Zhou, X. (2022). Impact of Tides and Surges on Fluvial Floods in Coastal Regions. *Remote Sensing*, *14*(22), 5779. https://doi.org/10.3390/rs14225779

Lin, P., Feng, D., Gleason, C. J., Pan, M., Brinkerhoff, C. B., Yang, X., Beck, H. E., Prata, R., & Frasson, D. M. (2023). Inversion of river discharge from remotely sensed river widths : A critical assessment at three-thousand global river gauges. *Remote Sensing of Environment*, *287*(February), 113489. https://doi.org/10.1016/j.rse.2023.113489

Mason, D. C., Horritt, M. S., Dall'Amico, J. T., Scott, T. R., & Bates, P. D. (2007). Improving river flood extent delineation from synthetic aperture radar using airborne laser altimetry. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(12), 3932–3943. https://doi.org/10.1109/TGRS.2007.901032

Modi, P., Revel, M., & Yamazaki, D. (2022). Multivariable Integrated Evaluation of Hydrodynamic Modeling: A Comparison of Performance Considering Different Baseline Topography Data. *Water Resources Research*, *58*(8). https://doi.org/10.1029/2021WR031819

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J. N. (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

Musa, Z. N., Popescu, I., & Mynett, A. (2015). A review of applications of satellite SAR, optical, altimetry and DEM data for surface water modelling, mapping and parameter estimation. *Hydrology and Earth System Sciences*, *19*(9), 3755–3769. https://doi.org/10.5194/hess-19-3755-2015

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis,

S., Tekalign, T. Y., Weitzner, D., & Matias, Y. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, *627*(8004), 559–563. https://doi.org/10.1038/s41586-024-07145-1

Parrish, C. E., Magruder, L. A., Neuenschwander, A. L., Forfinski-Sarkozi, N., Alonzo, M., & Jasinski, M. (2019). Validation of ICESat-2 ATLAS bathymetry and analysis of ATLAS's bathymetric mapping performance. *Remote Sensing*, *11*(14), 1634. https://doi.org/10.3390/rs11141634

Prigent, C., Jimenez, C., & Bousquet, P. (2020). Satellite-Derived Global Surface Water Extent and Dynamics Over the Last 25 Years (GIEMS-2). *Journal of Geophysical Research: Atmospheres*, *125*(3), 1–18. https://doi.org/10.1029/2019JD030711

Revel, M., Zhou, X., Modi, P., Yamazaki, D., Calmant, S., & Cretaux, J. (2023). AltiMaP : Altimetry Mapping Procedure for Hydrography Data. *Earth System Science Data*, *February*, 1–20. https://doi.org/https://doi.org/10.5194/essd-2022-438

Schellekens, J., Dutra, E., Martínez-De La Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J. C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., Van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., … Weedon, G. P. (2017). A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset. *Earth System Science Data*, *9*(2), 389–413. https://doi.org/10.5194/essd-9-389-2017

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T. A., & Slayback, D. A. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, *596*(7870), 80–86. https://doi.org/10.1038/s41586-021-03695-w

Tharme, R. E. (2003). A global perspective on environmental flow assessment: Emerging trends in the development and application of environmental flow methodologies for rivers. *River Research and Applications*, *19*(5–6), 397–441. https://doi.org/10.1002/rra.736

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G., & Fewtrell, T. J. (2016). The credibility challenge for global fluvial flood risk analysis. *Environmental Research Letters*, *11*(9). https://doi.org/10.1088/1748-9326/11/9/094014

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, *50*(9), 7505–7514. https://doi.org/10.1002/2014WR015638

Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., & Bouwman, A. (2013). A framework for global river flood risk assessments. *Hydrology and Earth System Sciences*, *17*(5), 1871–1892. https://doi.org/10.5194/hess-17-1871-2013

Wood, M., Hostache, R., Neal, J., Wagener, T., Giustarini, L., Chini, M., Corato, G., Matgen, P., & Bates, P. (2016). Calibration of channel depth and friction parameters in the LISFLOOD-FP hydraulic model using medium-resolution SAR data and identifiability techniques. *Hydrology and Earth System Sciences*, *20*(12), 4983–4997. https://doi.org/10.5194/hess-20-4983-2016

Wu, H., Kimball, J. S., Zhou, N., Alfieri, L., Luo, L., Du, J., & Huang, Z. (2019). Evaluation of real-time global flood modeling with satellite surface inundation observations from SMAP. *Remote Sensing of Environment*, *233*(August), 111360. https://doi.org/10.1016/j.rse.2019.111360

Yamazaki, D. (2024). *global-hydrodynamics/AllocRiverGauge: AllocRiverGauge version 1.0 (release_v1.0)*. Zenodo. https://doi.org/10.5281/zenodo.10893741

Yamazaki, D., Kanae, S., Kim, H., & Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, *47*(4), 1–21. https://doi.org/10.1029/2010WR009726

Yamazaki, D., Lee, H., Alsdorf, D. E., Dutra, E., Kim, H., Kanae, S., & Oki, T. (2012). Analysis of the water level dynamics simulated by a global river model: A case study in the Amazon River. *Water Resources Research*, *48*(9), 1–15. https://doi.org/10.1029/2012WR011869

Yamazaki, D., Sato, T., Kanae, S., Hirabayashi, Y., & Bates, P. D. (2014). Regional flood dynamics in a bifurcating mega delta simulated in a global river model. *Geophysical Research Letters*, *41*(9), 3127–3135. https://doi.org/10.1002/2014GL059744

Yang, Y., Pan, M., Lin, P., Beck, H. E., Zeng, Z., Yamazaki, D., David, C. H., Lu, H., Yang, K., Hong, Y., & Wood, E. F. (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

Zaitchik, B. F., Rodell, M., & Olivera, F. (2010). Evaluation of the Global Land Data Assimilation System using global river discharge data and a source-to-sink routing scheme. *Water Resources Research*, *46*(6). https://doi.org/10.1029/2009WR007811

Zhao, F., Veldkamp, T. I. E., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., Portmann, F. T., Leng, G., Huang, M., Liu, X., Tang, Q., Hanasaki, N., Biemans, H., Gerten, D., Satoh, Y., Pokhrel, Y., … Yamazaki, D. (2017). The critical role of the routing scheme in simulating peak river discharge in global hydrological models. *Environmental Research Letters*, *12*(7), 075003. https://doi.org/10.1088/1748-9326/aa7250

Zhou, X., Ma, W., Echizenya, W., & Yamazaki, D. (2021). The uncertainty of flood frequency analyses in hydrodynamic model simulations. *Natural Hazards and Earth System Sciences*, *21*(3), 1071–1085. https://doi.org/10.5194/nhess-21-1071-2021

Zhou, X., Prigent, C., & Yamazaki, D. (2021). Toward Improved Comparisons Between Land-Surface-Water-Area Estimates From a Global River Model and Satellite Observations. *Water Resources Research*, *57*(5), e2020WR029256. https://doi.org/10.1029/2020WR029256

Zhou, X., Yamazaki, D., Revel, M., Zhao, G., & Modi, P. (2024). *Benchmark Framework for Global River Model (version 1.0) (release_v1.0)*. Zenodo. https://doi.org/10.5281/zenodo.10903211