scientific data

OPEN

DATA DESCRIPTOR

A Machine Learning-Reconstructed Dataset of River Discharge, Temperature, and Heat Flux into the Arctic Ocean

Zihan Wang[1,2], Fengming Hui[1,2] ✉ & Xiao Cheng[1,2]

Arctic rivers deliver 11% of the global river discharge volume into the Arctic Ocean, influencing ocean circulation, sea ice, and coastal ecosystems. Our understanding of these patterns is limited by substantial data gaps. To address this, we present the Reconstructed Arctic-draining river DIscharge and Temperature (RADIT) dataset, a comprehensive record of reconstructed daily discharge, temperature, and heat flux for 25 major Arctic rivers from 1950 to 2023. Based on machine learning regression methods and ERA5-Land reanalysis data, we designed distinct reconstruction frameworks for discharge and temperature, considering the different characteristics of the observational data. We achieved high reconstruction accuracy, with median Nash–Sutcliffe efficiency (NSE) values of 0.861 for discharge and 0.906 for temperature. The RADIT dataset, with extensive spatial and temporal coverage, is a valuable resource for understanding Arctic hydrology and its response to climate change. It will improve Arctic freshwater budget quantification, climate model calibrations, and assessments of river impacts on the Arctic Ocean, enhancing our understanding of the role of the Arctic Ocean in the global climate system.

Background & Summary

The Arctic Ocean, as Earth's smallest and shallowest ocean basin, constitutes a unique component of the global ocean system[1]. It is characterized by distinctive physical features, including perennial sea ice cover, strong seasonal variations, and complex interactions among the atmosphere, cryosphere, and ocean circulation patterns[2,3]. One of the key features of the Arctic Ocean is its disproportionately high influence from riverine inputs. Despite comprising only 1% of the global ocean volume, it receives 11% of the global river discharge[4].

This substantial freshwater influx fundamentally shapes the Arctic environment via multiple mechanisms. River discharge directly influences ocean salinity and stratification, sea ice formation processes, and thermohaline circulation patterns[5,6]. Moreover, rivers transport considerable amounts of heat, nutrients (e.g., nitrogen and phosphorus), dissolved organic matter, and particulate organic carbon to the Arctic Ocean, thus profoundly affecting coastal and marine ecosystems[7–10].

Among these riverine influences, discharge, water temperature, and heat flux are three key parameters for characterizing river–ocean interactions. These hydrological variables influence various physical and ecological processes, such as river ice dynamics[11,12], coastal erosion[13], and aquatic ecosystems[14], highlighting the interconnected nature of these systems. Recent studies have revealed significant changes in these variables, with both discharge volumes and water temperatures exhibiting notable trends across many Arctic regions[15–19]. Therefore, understanding the temporal and spatial patterns of such changes is crucial for assessing the response of the Arctic system to climate change and its implications for both regional processes and global climate dynamics.

Although the importance of these parameters in shaping Arctic physical and ecological processes has been recognized, our ability to fully understand and predict changes in these critical variables faces several fundamental challenges. The most pressing issue is the lack of a unified monitoring network that provides long-term, comprehensive coverage across the pan-Arctic region. Historically, observations have been collected by different

[1]School of Geospatial Engineering and Science, Sun Yat-sen University, and Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519082, China. [2]Key Laboratory of Comprehensive Observation of Polar Environment (Sun Yat-sen University), Ministry of Education, Zhuhai, 519082, China. ✉e-mail: huifm@mail.sysu.edu.cn

national agencies via various protocols and measurement frequencies, leading to data quality and temporal resolution inconsistencies. This situation has been aggravated by the continuous decline in monitoring stations since the mid-1980s[20,21]. Moreover, the spatial coverage of observations is highly uneven, with significant data gaps for both North American and Eurasian Arctic rivers[22,23].

These data limitations have substantially limited the scope and reliability of relevant research. While recent efforts, including those using reanalysis-based datasets[16,24], have significantly improved the spatial coverage of Arctic river discharge estimates, many observation-based studies still concentrate on a few major rivers with relatively complete records, such as the Yenisey, Lena, and Mackenzie Rivers[18,25,26]. For other Arctic-draining rivers with significant data gaps, previous studies have often relied on simplified interpolation techniques[27] or have been restricted to periods with adequate observations[17]. This selective coverage and simplified treatment of missing data create potential biases in our understanding of pan-Arctic river systems, as rivers with short or discontinuous observational records are often overlooked despite their potential importance to regional hydrology[28]. Given these observational limitations, researchers increasingly employ hydrological models as complementary tools to simulate discharge patterns and thermodynamics across various spatial and temporal scales[24,29]. However, these models often exhibit substantial uncertainties[24,30] due to the complexity of Arctic hydrological processes, such as snowmelt[31] and permafrost–hydrology interactions[32,33]. These limitations collectively restrict our ability to comprehensively assess the freshwater and heat budgets of the Arctic Ocean, highlighting the urgent need for more comprehensive and reliable datasets across the entire Arctic-draining river system.

The aim of this study was to develop a comprehensive dataset covering 25 major rivers flowing into the pan-Arctic Ocean to resolve existing data limitations. To achieve this goal, we created the Reconstructed Arctic-draining river DIscharge and Temperature (RADIT) dataset, which provides continuous daily records of river discharge, temperature, and heat flux spanning from 1950 to 2023. This dataset encompasses rivers in various geographical regions, including Eurasia (e.g., Ob, Yenisei, and Lena) and North America (e.g., Mackenzie and Yukon), collectively accounting for the majority of the freshwater input to the Arctic Ocean.

The reconstruction of this dataset presented several methodological challenges. First, the highly irregular and heterogeneous temporal distribution of missing values in historical records substantially limits the applicability of traditional time series approaches. Second, substantial differences in hydrological and climatic conditions across Arctic river basins complicate the task of accurately reconstructing discharge from limited and uneven observational records. To address these challenges, we designed a reconstruction framework that integrates machine learning techniques with data-driven design choices specific to the characteristics of discharge and temperature data. For river discharge reconstruction, we implemented individual models for each river so that the reconstruction can be adapted to river-specific data patterns and variability. In the river temperature reconstruction process, we employed a unified model approach to capture spatial variations in temperature dynamics across the pan-Arctic region, even for rivers lacking historical measurements. This integrated framework allowed for consistent, high-quality reconstruction results, while adapting to the differing data conditions and hydrological behaviors across river systems. Evaluations against *in situ* observations showed that the RADIT dataset achieved high reconstruction performance, with Nash–Sutcliffe efficiency (NSE) values exceeding 0.8 for most rivers in both discharge and temperature reconstructions, supporting its applicability for various research purposes.

The RADIT dataset provides an important contribution to Arctic hydrological data availability, offering a comprehensive machine-learning-based reconstruction of continuous daily river discharge and temperature records across 25 major Arctic-draining rivers. Our approach successfully fills a gap in Arctic research infrastructure and facilitates more comprehensive investigations of land-to-ocean freshwater and heat fluxes in a changing Arctic environment. The daily temporal resolution of this dataset enables the investigation of hydrological processes at finer temporal scales, providing new opportunities to analyze rapid changes in Arctic river systems. The extensive temporal coverage of this dataset, spanning over seven decades, makes it a useful resource for understanding long-term Arctic system changes, improving climate models, and advancing our knowledge of land–ocean interactions in the Arctic region. These high-quality data can support various research applications, including hydrological model validation and calibration, assessments of the impacts of climate change, ocean circulation modeling, and ecosystem studies in the rapidly changing Arctic environment.

Methods

In situ river discharge measurements.

In this study, we utilized daily river discharge observations from multiple hydrological databases, which were integrated into a comprehensive dataset as the foundation for machine learning-based discharge reconstructions.

(1) *ArcticGRO*: The Arctic Great Rivers Observatory (ArcticGRO) project, of which its predecessor was initiated in the mid-1990s, was established to integrate hydrological and biogeochemical data from Arctic rivers (https://arcticgreatrivers.org/data/). In 2008, it was rebranded to its current name, establishing itself as a key component of the National Science Foundation (NSF)'s Arctic Observing Network[34]. This project aims to collect river hydrological observation data from hydrometeorological agencies across several countries, including Russia's Roshydromet, the United States Geological Survey (USGS), and the Water Survey of Canada (WSC). Initially, the ArcticGRO project focused on six major Arctic rivers: the Ob, Yenisey, Lena, and Kolyma Rivers in Russia, along with the Yukon and Mackenzie rivers in North America. The project expanded in 2019 to include medium-sized Russian rivers, such as the Northern Dvina, Mezen, and Pechora rivers[35]. To date, ArcticGRO provides daily discharge observations for 15 rivers flowing into the Arctic Ocean, with data extending back to the 1930s that are updated regularly.

(2) *GRDC*: The Global Runoff Data Centre (GRDC) operates as an international database under the guidance of the World Meteorological Organization (WMO) (https://portal.grdc.bafg.de/). Established in 1988,

the GRDC is managed by the Federal Institute of Hydrology of Germany, located in Koblenz. Compared with the ArcticGRO project, the GRDC maintains a considerably more extensive network of worldwide hydrological stations, with over 9500 stations and records dating back to the early 19th century. This broad spatiotemporal coverage and the availability of long-term records make it one of the world's largest quality-controlled river discharge observation datasets. Additionally, GRDC products include vector data delineating watershed boundaries for each hydrological station, which facilitates basin-scale analyses.

(3) *WSC*: The Hydrometric Data (HYDAT) database of the Water Survey of Canada (WSC) serves as a vital repository for hydrological monitoring across Canada (https://wateroffice.ec.gc.ca/). It contains extensive river discharge records from a network of gauging stations distributed throughout the country, including major river systems such as the Mackenzie River. The observed data are subjected to rigorous quality control and are regularly updated.

(4) *ArcticRIMS*: The ArcticRIMS dataset is another comprehensive discharge database covering the pan-Arctic region (https://rims.unh.edu/). In this study, we utilized ArcticRIMS data, primarily for Russian rivers and the Yukon River in Alaska, although most records for these rivers terminate between the late 1990s and early 2000s.

(5) *R-ArcticNet*: The R-ArcticNet dataset, developed by the Water Systems Analysis Group at the University of New Hampshire (UNH), is a comprehensive discharge database focused on Arctic rivers (https://www.r-arcticnet.sr.unh.edu/v4.0/). In this study, we used V4.0 of the dataset, which includes a sub-dataset providing daily discharge records for 139 rivers in Russia. Similar to ArcticRIMS, most observations in R-ArcticNet terminated in the early 2000s.

After collecting the above observed data products, we first conducted a preliminary quality assessment and then merged the different datasets on the basis of their monitoring sites. In this process, a common challenge was the inconsistent naming conventions and geographic coordinates for identical rivers or sites across the different datasets. To address this issue, we systematically verified the data consistency between proximate sites on matching dates and unified the names of rivers or sites confirmed as identical. On this basis, when inconsistencies occurred in measurements from the same site and for the same date across multiple datasets, the final observation value was determined by calculating the mean or mode. Following these steps, we established an initial daily discharge observation dataset for rivers flowing into the pan-Arctic Ocean. For the purpose of this study, the following criteria were adopted to select the rivers included:

(1) The selected river should have no more than 60% missing daily discharge data from the most downstream gauging station between 1950 and 2023;

(2) The river mouth should be situated above 60°N;

(3) The average observed discharge should be no less than 100 m^3/s (for reference, this threshold is two orders of magnitude smaller than the mean discharge of three major Russian Arctic rivers — the Ob, Yenisei, and Lena Rivers — each exhibit an average discharge greater than 10,000 m^3/s).

Following these selection criteria, a total of 25 rivers were selected for the following reconstruction study. Their spatial distributions and basic characteristics are shown in Fig. 1a and Supplementary Table 1, while Fig. 1b illustrates the temporal coverage of daily observations for each river. These rivers represent the majority of medium-sized and large rivers draining into the pan-Arctic Ocean, thus ensuring comprehensive spatial coverage of the region. The discharge, temperature, and heat flux of these 25 rivers were subsequently reconstructed. Across the selected monitoring stations, 29.2% of the daily discharge observations were missing for the 1950–2023 period. Notably, even major river systems such as the Kolyma and Yukon Rivers presented substantial data gaps, with missing data ratios of 38.6% and 40.9%, respectively. The significant data limitations highlight the importance of continuous long-term hydrological records for understanding river systems, as they are essential for detecting long-term trends and seasonal patterns in river behavior.

In situ river temperature measurements. River water temperature data are markedly less accessible to the public than discharge measurements are. In this study, we primarily utilized the ART-Russia dataset, which provides long-term temperature observations from the Russian Arctic region[17] (https://www.r-arcticnet.sr.unh.edu/RussianRiverTemperature-Website/). All the measurements in this dataset were collected before 2003. For the 25 rivers listed in Supplementary Table 1, this dataset comprises observations from 14 rivers monitored across 16 stations. The observations were conducted by Roshydromet of Russia and were subsequently compiled and published via a collaborative effort between scientists from the State Hydrological Institute (SHI) of Russia and the UNH of the United States. Given the high variability in river temperature, measurements were conducted twice daily, at 8:00 AM and 8:00 PM, with a precision of 0.1 °C[17]. The data are presented as 10-day average values rather than daily values, yielding three monthly records per station on the 5th, 15th, and 25th of each month. While these observations focus primarily on the warm season (ice-free period), winter measurements are also included.

The ART-Russia dataset encompasses three temperature versions: T_0, T_1, and T_2. The T_0 version contains raw data after basic quality control, the T_1 version represents a cleaned dataset with modifications at warm season boundaries, and the T_2 version is derived from the T_1 version by setting all winter temperatures to 0 °C. Although T_1 and T_2 were processed for specific applications, such as ice-free period analysis and energy flux calculations, these adjustments are not necessary for our study. Hence, we selected T_0 because it offers the largest sample size (13,470 records) while ensuring adequate quality control for our research objectives.

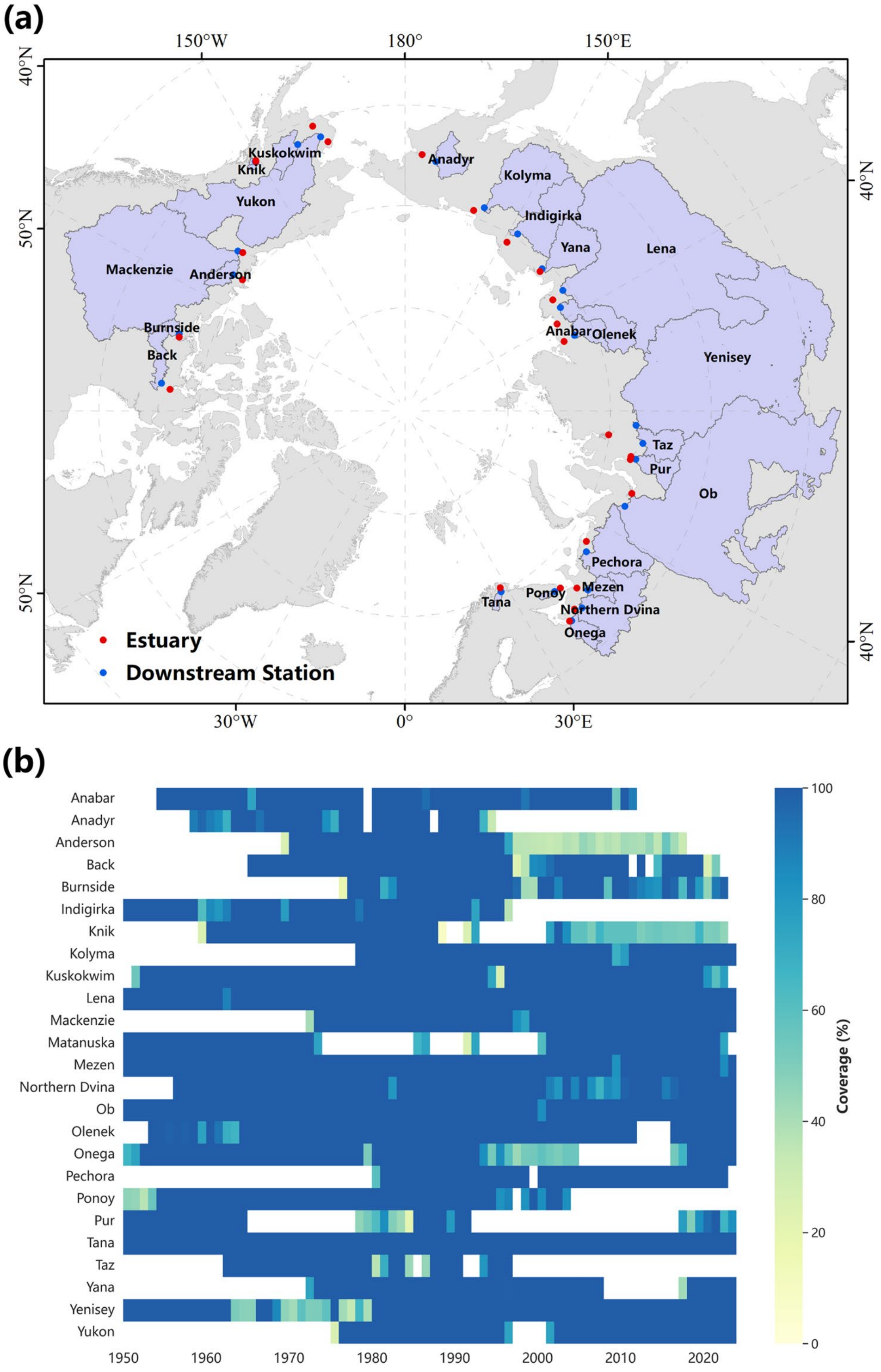

Fig. 1 (**a**) Map of the watershed, monitoring stations and estuary of the 25 selected Arctic-draining rivers in this dataset. (**b**) Temporal coverage for 25 Arctic-draining rivers from 1950 to 2023.

ERA5-Land. The high-resolution ERA5-Land reanalysis dataset is produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and is designed to provide detailed land surface information[36] (https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR). Derived from global ERA5 reanalysis with a spatial resolution of ~31 km, the ERA5-Land dataset improves upon the original dataset by providing a finer resolution of ~9 km, thus facilitating a more precise representation of land surface processes. Additionally, ERA5-Land focuses specifically on land surface variables, such as soil moisture and surface temperature, rendering it particularly valuable for terrestrial hydrology and climate research[37,38]. This dataset covers the period from 1950 to the present, providing both hourly and monthly data that enable historical reconstruction and analysis of long-term trends. The capacity to provide continuous, long-term time series of environmental parameters renders the ERA5-Land dataset particularly suitable for river discharge reconstruction studies, as it ensures temporal consistency and completeness of the input data. In this study, we utilized 14 variables from the ERA5-Land dataset to reconstruct the river discharge and temperature, leveraging its fine temporal and spatial resolutions to capture the complex interactions between climate and hydrological dynamics. We downloaded daily average data from the Google Earth Engine (GEE) to ensure alignment with river observations, including the 2-m temperature, 2-m dewpoint temperature, soil temperature of the top layer, snow depth, snowfall, snowmelt, evaporation, precipitation, surface and subsurface runoff, 10-m wind (u- and v-components), and surface solar and thermal radiation fluxes. Although ERA5-Land runoff has known discrepancies in long-term trend estimates compared with Arctic observations[24], it was used as a candidate predictor rather than a deterministic input. Our model is designed to select relevant features based on their predictive performance rather than their absolute values or trends.

Other river data.

The Simulated Topological Network (STN-30p), which is designed for hydrological and climate applications, is a global river routing dataset with a 30 arc-minute spatial resolution[39]. This dataset includes several key variables that represent various aspects of river networks, such as river routing, drainage basins, and flow direction. Among these variables, the catchment layer provides clear information on the drainage area of each pixel, thereby quantifying the area contributing to river flow at specific points. We applied this dataset to convert discharge measurements from downstream sites to river mouths, as described in the "River discharge reconstruction via machine learning regression" section.

Another auxiliary river dataset is the Global River Widths from Landsat (GRWL) dataset, which is a global river width product that provides high-resolution measurements of river and stream geometries[40] (https://zenodo.org/records/1297434). Derived from Landsat satellite imagery, the GRWL dataset contains over 58 million measurements of river widths under mean annual discharge conditions, covering rivers wider than 30 m. In our study, the GRWL data support river temperature reconstruction and subsequent analysis.

Machine learning regression models.

Frequent discontinuities in the *in situ* river discharge and temperature records render traditional time series prediction methods unsuitable for reconstruction (Fig. 1b). Therefore, we adopted regression-based machine learning approaches to estimate daily missing values and reconstruct complete time series. Four widely used ensemble learning models—Random Forest (RF), Gradient Boosting (GB), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM)—were employed. These models represent two major ensemble strategies—bagging (RF) and boosting (GB, XGBoost, and LightGBM)—and were selected for their proven effectiveness in handling nonlinear relationships, missing data, and diverse feature sets in environmental modeling tasks.

1) RF: The RF model is a bagging-based ensemble method that builds multiple decision trees using randomly resampled subsets of the training data[41]. Each tree contributes to the final prediction, and the model output is obtained by averaging the predictions of all trees. Random feature selection at each node split further reduces tree correlation, improving generalization and robustness[41].
2) GB: GB is a boosting-based technique that constructs trees sequentially, with each new tree trained to minimize the errors of its predecessors[42]. It focuses on learning on samples with larger residuals, gradually reducing the overall prediction error. A learning rate is applied to control each tree's contribution and prevent overfitting.
3) XGBoost: XGBoost enhances traditional gradient boosting with several algorithmic innovations, including regularization, parallel computation, and optimized tree construction[43]. It is known for its high computational efficiency and accuracy and is particularly well suited for large-scale structured data[44].
4) LightGBM: LightGBM is a highly efficient gradient boosting framework optimized for speed and memory usage[45]. It introduces histogram-based feature binning and a leaf-wise tree growth strategy to accelerate training. LightGBM also supports parallel learning and advanced regularization techniques, making it effective for high-dimensional or sparse data.

While these models are powerful and flexible, they also have limitations. As non-temporal regression models, they do not explicitly account for long-term temporal dependencies, which may affect their ability to represent low-frequency variability. Additionally, their performance depends on the quality and distribution of the input data, and their "black-box" nature limits interpretability compared with physically based hydrological models. These limitations should be considered when applying the reconstructed results in process-oriented analyses.

Evaluation metrics.

Several evaluation metrics were employed to assess the performance of the regression models for river discharge prediction, namely, the NSE[46], Kling–Gupta efficiency (KGE)[47], and normalized root mean square error (NRMSE), which can be calculated as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - Q_{\mathrm{pred},i})^2}{\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - \bar{Q}_{\mathrm{obs}})^2} \tag{1}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{2}$$

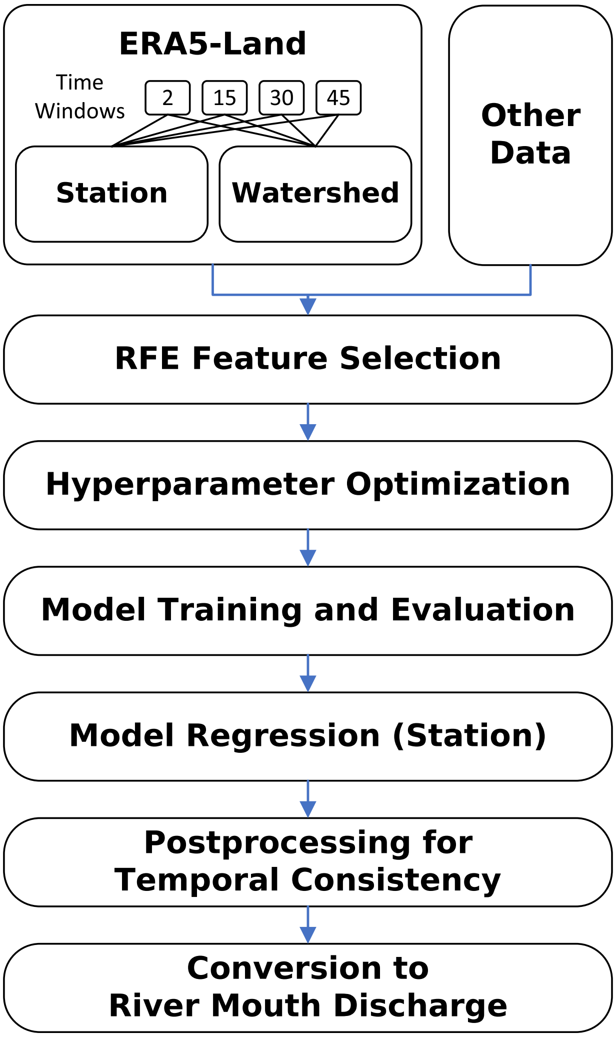
ERA5-Land
Time Windows
2
15
30
45
Station
Watershed
Other Data
RFE Feature Selection
Hyperparameter Optimization
Model Training and Evaluation
Model Regression (Station)
Postprocessing for Temporal Consistency
Conversion to River Mouth Discharge

Fig. 2 Flow chart of establishing a discharge reconstruction model for each Arctic-draining river.

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - Q_{\mathrm{pred},i})^2}}{\overline{Q}_{\mathrm{obs}}} \times 100\% \tag{3}$$

where $Q_{\mathrm{obs},i}$ is the observed discharge value at the i-th time step, $Q_{pred,i}$ is the predicted discharge value at the i-th time step, $\overline{Q}_{\mathrm{obs}}$ is the mean of the observed discharge values, r is the Pearson correlation coefficient between the observed and predicted discharge values, α is the ratio of the standard deviation of the predicted to observed values, β is the ratio of the mean of the predicted values to the mean of the observed values, and n is the total number of observations.

The NSE and KGE values range from negative infinity to 1, where 1 indicates perfect agreement between the observed and predicted values. The NRMSE approaches zero with increasing model accuracy, reflecting smaller relative prediction errors.

River discharge reconstruction via machine learning regression. To address the challenge of missing discharge data for Arctic-draining rivers, we developed a machine learning-based approach to reconstruct daily discharge values, as outlined in Fig. 2. Given that each river system exhibits unique characteristics and response patterns to environmental factors, we implemented separate models for each river. This river-specific approach allows the models to better capture the distinct relationships between environmental variables and discharge within each watershed, indirectly reflecting the influence of local geographical, climatic, and hydrological conditions through individualized model training. Our methodology comprises the following steps:

(1) Preparation of model input features
Environmental data from the ERA5-Land dataset were collected for the 1950–2023 period to construct input features on a daily basis. Following common practices in hydrological modeling, we incorporated

multiple temporal scales to capture hydrological processes operating at different timescales. The mean values of 14 environmental variables (detailed in the "ERA5-Land" section) were calculated across four time windows (2, 15, 30, and 45 days). These time windows were selected to represent different hydrological response times, from rapid discharge generation (2 days) to slower subsurface flow processes (15–45 days)[48,49]. These variables were computed separately for each river station and the corresponding upstream catchment area. This approach yielded a total of 112 features (2 locations × 4 time windows × 14 variables), ensuring that the model captured both short- and long-term hydrological processes. The upstream catchment area of each river station was delineated via the GRDC product. Additionally, the Julian day and its cosine value were incorporated as supplementary features to represent seasonal discharge patterns.

(2) Feature selection and model tuning

Given the irregular temporal gaps in discharge records and the absence of suitable continuous time series, we adopted a regression-based modeling strategy, treating each daily observation as an independent sample. Following a river-specific modeling framework, we evaluated four ensemble machine learning algorithms—RF, GB, XGBoost, and LightGBM—each of which has demonstrated effectiveness in hydrological prediction tasks.

For each river, the available daily observations were chronologically divided into training (70%), validation (15%), and testing (15%) sets. This time-ordered split ensures temporal independence between the training and testing data, better simulating real-world reconstruction scenarios.

Prior to training, we performed feature selection using recursive feature elimination (RFE), which iteratively removes the least important features. The 20 most informative predictors were retained to reduce model complexity and minimize overfitting risk.

Considering the extreme seasonal variability in Arctic discharge—particularly sharp spring peak flows—a targeted data augmentation strategy was designed. Peak flow samples (defined as those above the 90th percentile in the training set) were synthetically oversampled: each was duplicated 10 times with small random perturbations (±1%) applied to feature values. This enhanced the model's sensitivity to peak discharge events, which are both hydrologically important and often underrepresented in training data.

All the models were tuned using a random search with 5-fold cross-validation implemented via RandomizedSearchCV in scikit-learn (v1.4.2) for RF and GB, XGBoost (v2.0.3), and LightGBM (v4.5.0). This approach allows for efficient exploration of the hyperparameter space and leverages internal cross-validation to reduce the risk of overfitting[50]. The validation set was used to assess whether peak-flow augmentation improved performance. If the model trained on augmented data had a higher NSE on the validation set than did the baseline model, the augmented dataset was adopted for final training. To quantify prediction uncertainty, we implemented model-specific approaches. For RF, uncertainty bounds were derived from the distribution of outputs across all trees, while for GB, XGBoost, and LightGBM, the models were trained using quantile regression to estimate 2.5% and 97.5% prediction intervals.

(3) Final training and model selection

After determining the use of data augmentation, the training and validation sets were combined to retrain the final model. For algorithms supporting early stopping (XGBoost, LightGBM, GB), 10% of the combined data were held out as an internal evaluation set to monitor performance and prevent overfitting. Training was terminated if no improvement was observed after 40 rounds. For models that do not support early stopping (e.g., RF), overfitting was controlled via structural hyperparameters such as maximum tree depth and minimum samples per leaf.

Model performance was ultimately evaluated on the independent testing set using metrics defined in the "Evaluation metrics" section. For each river, the best-performing model was selected based on a hierarchical evaluation strategy: models without signs of overfitting were ranked by test set performance; if all models showed signs of overfitting, the model with the best trade-off between generalization and accuracy was chosen. The selected model was then applied to reconstruct missing daily discharge values.

(4) Postprocessing for temporal consistency

To improve the accuracy and coherence of the reconstructed discharge series, a two-stage postprocessing procedure was applied.

First, a correction was introduced to address the tendency of the models to underestimate peak flows. Based on the 90th percentile of the training data, a segmented linear correction was developed using training data and evaluated on the validation set. A smoothing zone (± 10% around the threshold) was used to ensure a gradual transition in correction strength. This adjustment was applied only when it led to improvements in both the overall NSE and peak flow NSE in the validation set.

Second, to ensure temporal consistency across the reconstructed segments, we implemented a continuity adjustment step. For missing segments shorter than 20 days and occurring outside the peak flow period (May–August), linear interpolation was applied, as it effectively captures gradual variations during stable flow conditions. For all other segments, a boundary-matching adjustment was applied to minimize discontinuities at the junctions between the observed and predicted values. This method optimizes a penalty function based on boundary mismatch and adjusts the first and last few days of each filled segment (up to 10% of the segment length or 10 days, whichever is smaller), while preserving the internal values. This two-step strategy improves both the amplitude accuracy and temporal smoothness of the reconstructed discharge time series.

By applying this methodology to all 25 rivers, we were able to obtain continuous daily discharge time series. Figure 3 shows examples of four typical rivers with substantial data gaps in their historical records—namely, the Onega, Indigirka, Taz, and Anadyr Rivers. These examples demonstrate how our approach reconstructs missing values and produces continuous discharge records at gauging stations.

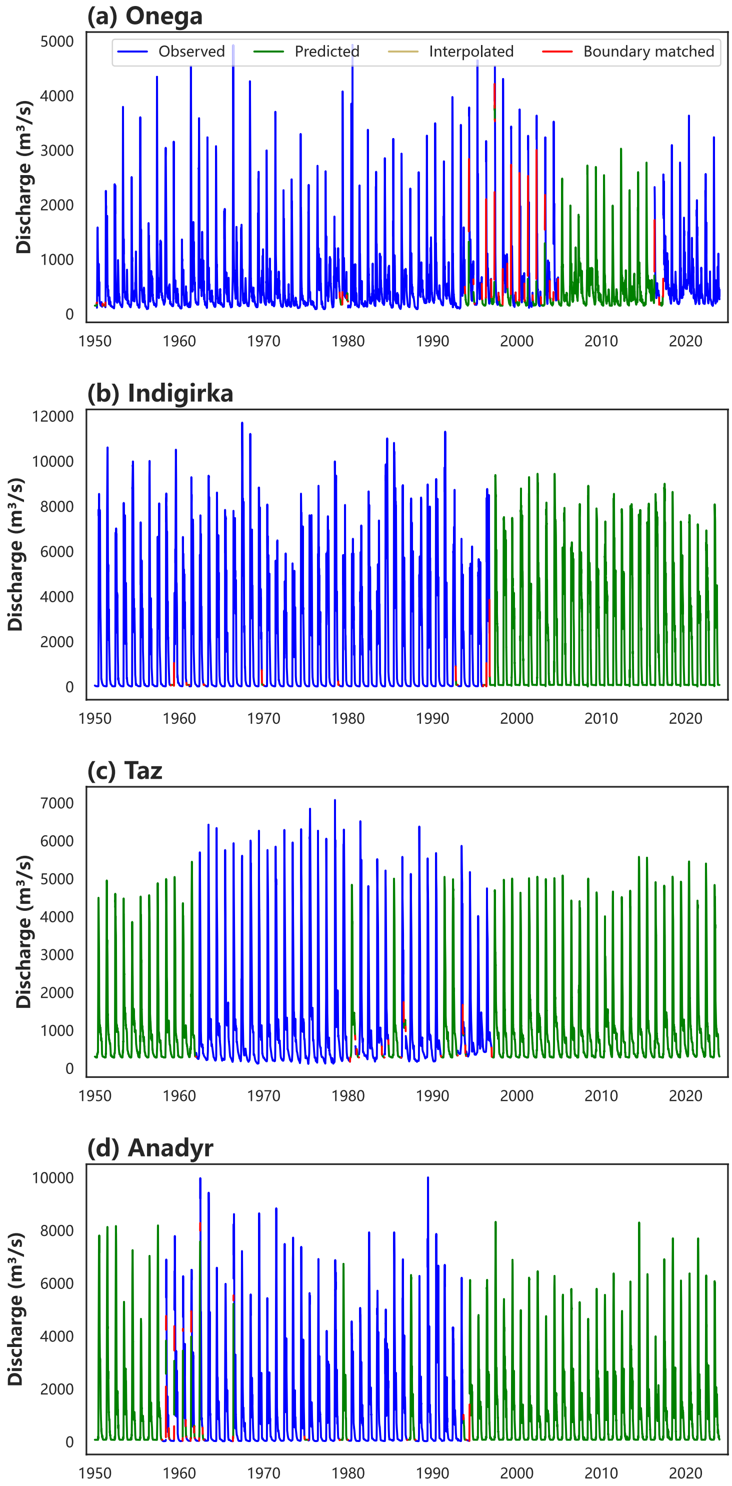

Fig. 3 Examples of reconstructed daily discharge time series for the 1950–2023 period at four gauging stations with substantial missing records: (**a**) Onega, (**b**) Indigirka, (**c**) Taz, and (**d**) Anadyr rivers. The continuous time series are composed of observed values (blue), model predictions (green), linear interpolation (yellow), and boundary-matched values (red).

(5) Conversion from station to river mouth discharge values
The steps described above yielded reconstructed continuous discharge time series at the downstream gauging stations. As this study focused on riverine freshwater input to the ocean, we then estimated the discharge at river mouths from these station-based values. Following the method of Dai and Trenberth[51], the station discharge values were scaled to obtain river mouth discharge values via catchment area ratios. For this transformation, drainage area ratios were utilized from the STN-30p product, which provides gridded

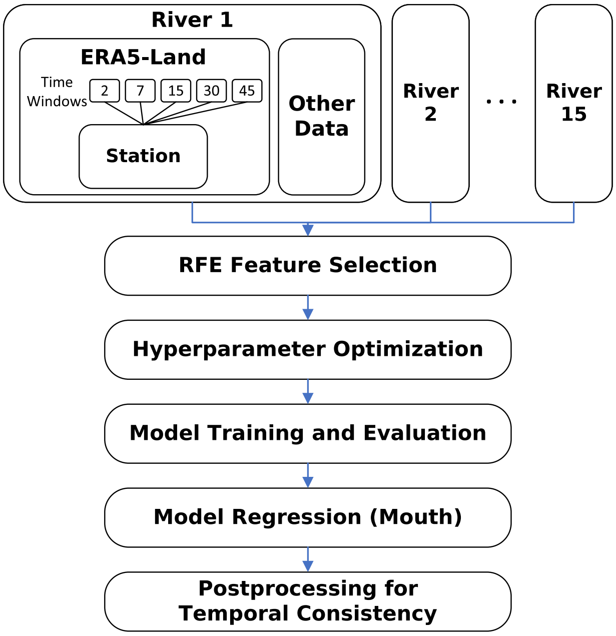

Fig. 4 Flow chart of establishing a unified river temperature reconstruction model for the pan-Arctic region.

watershed area data at a 0.5° spatial resolution. For most rivers, the gauged catchment area exceeded the STN-30p unit. Only three rivers had slightly smaller catchments, but in all the cases, the station catchment area accounted for more than 98% of the STN-30p grid area, minimizing potential mismatches. The river mouth discharge Q_{mouth} can be estimated as follows:

$$Q_{\mathrm{mouth}} = Q_{\mathrm{station}} \cdot \frac{C_{\mathrm{mouth}}}{C_{\mathrm{station}}} \tag{4}$$

where C_{mouth} and C_{station} denote the catchment areas at the river mouth and station locations, respectively, and Q_{station} denotes the discharge at the river station. We calculated the ratio between the STN-30p drainage area at the river mouth grid cell and that at the downstream gauging station grid cell. For all the rivers, this ratio ranged between 1.0 and 2.8. To ensure the appropriateness of applying area-ratio scaling, we visually inspected satellite imagery to confirm the hydrological connectivity between the gauging station and the corresponding river mouth. Based on this assessment, we applied the scaling method to all rivers in the dataset.

River temperature and heat flux reconstruction via machine learning regression. Building upon the discharge reconstruction method, we developed a modified approach for river temperature reconstruction across the pan-Arctic region (Fig. 4). While discharge observations were available for all 25 rivers in this study, water temperature measurements were available for 14 rivers, corresponding to 16 gauging stations. Two of these stations are located upstream but were included due to the availability of both discharge and temperature records, necessitating a different modeling strategy. In contrast to our discharge reconstruction approach, where river-specific models were developed for each individual river, here, we constructed a unified temperature model to estimate river temperatures at any location across the pan-Arctic region.

The feature selection process was modified from the discharge reconstruction approach to better capture local influences on river temperature. Given that river temperatures are influenced primarily by local conditions, we excluded watershed-scale variables from the ERA5-Land dataset and retained only point-scale variables at the measurement locations. To account for river-specific characteristics that influence temperature dynamics, we incorporated additional physical parameters, including the river width from the GRWL dataset and the catchment area from STN-30p. Recognizing the influence of discharge on river temperature, we also included the reconstructed discharge data from our previous analysis as input feature data. In addition to the temporal windows used in discharge reconstruction (2, 15, 30, and 45 days), we incorporated a 7-day window based on previous studies that documented the air–water temperature relationships in Arctic rivers[52]. These multiple temporal windows were applied to all dynamic variables to capture both short-term responses and longer-term dependencies in temperature patterns and their driving factors.

To validate our approach, we implemented a cross-validation strategy using data from 16 monitoring stations. For the discharge reconstruction, we evaluated four machine learning models—RF, GB, XGBoost, and LightGBM—for river temperature prediction. The model with the best overall cross-validation performance across the stations was then selected for training the final unified temperature reconstruction model. For each station, we used data from the remaining 15 stations as the training dataset while reserving the data of the target station as the testing dataset. This approach allowed us to assess the model's ability to predict temperatures at locations not included in the training data, thereby evaluating its spatial transferability. After confirming

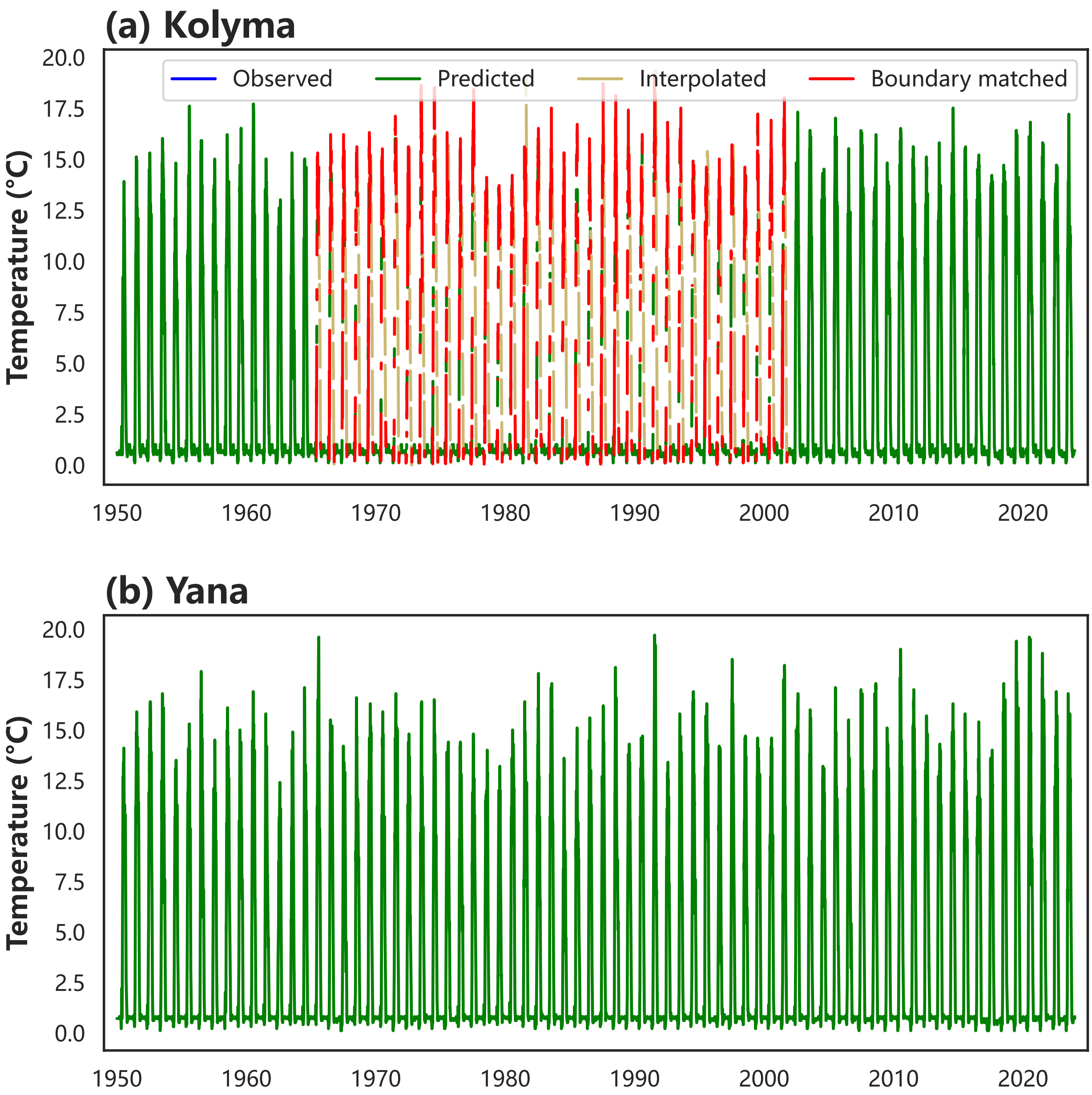

Fig. 5 Examples of reconstructed daily temperature time series for the 1950–2023 period at two river mouths: (**a**) Kolyma and (**b**) Yana Rivers. The continuous time series are composed of observed values (blue), model predictions (green), linear interpolation (yellow), and boundary-matched values (red).

satisfactory model performance via this validation process, we developed a final model trained with the complete dataset from all stations. This unified model was then applied to directly estimate temperatures at river mouths via local environmental parameters, enabling consistent temperature estimates across all Arctic-draining rivers, including those without historical temperature measurements. As with discharge reconstruction, prediction uncertainty was also quantified for river temperature estimates using model-specific approaches: ensemble-based interval estimation for RF and quantile regression for GB, XGBoost, and LightGBM.

The application of this unified model is illustrated in Fig. 5 through two contrasting examples. The reconstruction for the Kolyma River represents a case where historical observations were available for model validation, whereas the Yana River demonstrates the model's capability to generate temperature estimates for rivers without any historical measurements, highlighting the spatial transferability of our approach.

The river heat flux (relative to the water freezing point) is calculated via a universal equation that is widely applicable for quantifying heat transport in river systems[19,53]:

$$HF = 86400 \cdot Cp \cdot \rho \cdot Q \cdot W \tag{5}$$

where *HF* is the daily heat flux (10^6 MJ); *Q* is the daily mean river discharge (m^3/s); *T* is the daily mean water temperature (°C); C_p is the specific heat capacity of water, with a fixed value of 4.184 J/(g·°C); ρ is the water

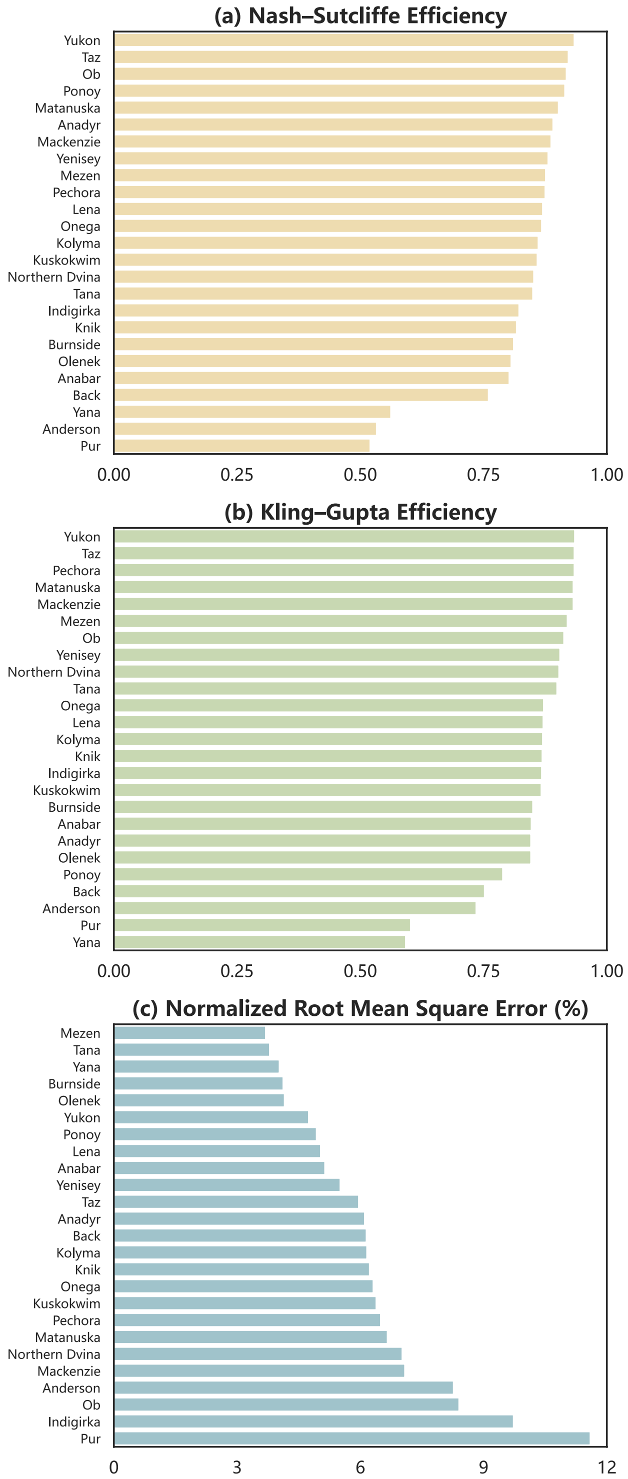

Fig. 6 Statistical metrics for discharge reconstruction for 25 Arctic-draining rivers: (**a**) NSE, (**b**) KGE, and (**c**) NRMSE.

density, with a fixed value of $10^6 g/m^3$; and 86400 is the number of seconds in one day. Both the discharge and temperature data were derived from previous reconstructions.

Data Records

The dataset is available at Zenodo (15811422)[54]. The dataset package primarily includes our reconstructed data to fill gaps in historical observational records. It comprises three main components: (1) metadata for each river, including station name, station and estuary coordinates, and catchment area; (2) tabulated results of reconstructed daily river discharge, temperature, and heat flux. These files provide the reconstructed daily records

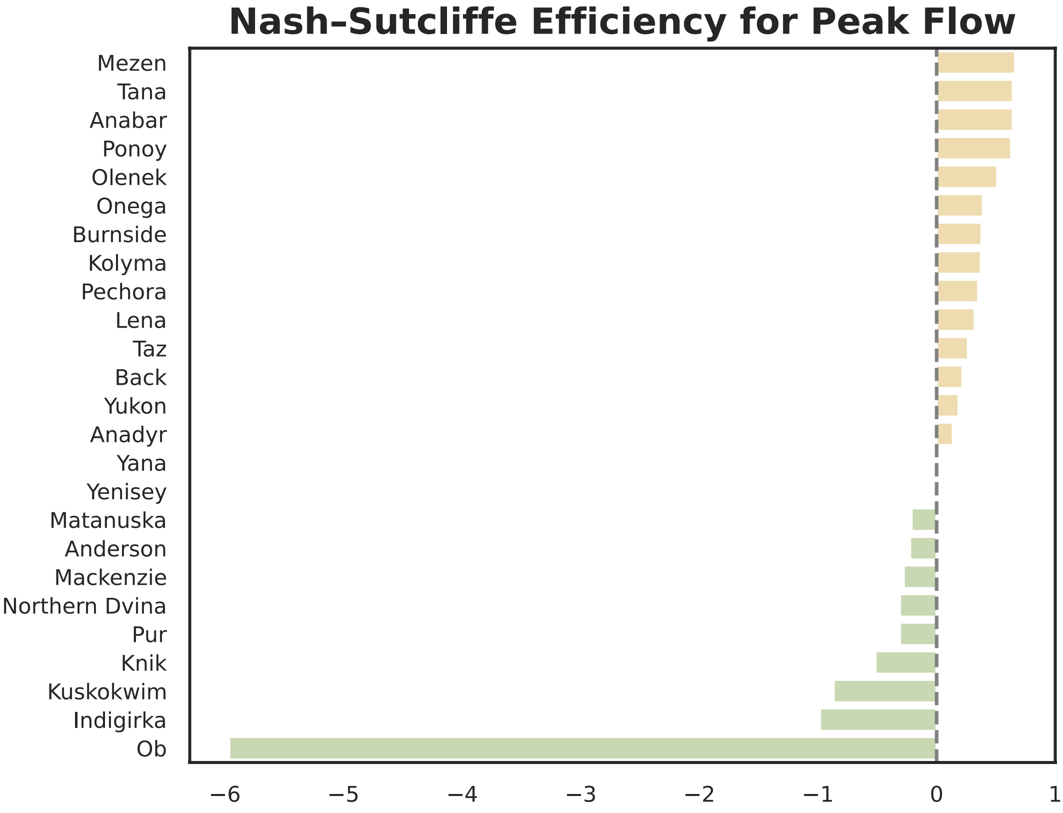

Fig. 7 NSE values for reconstructed peak discharge across 25 Arctic-draining rivers.

for the 1950–2023 period. All the data are stored in UTF-8-encoded CSV files, thus ensuring compatibility with standard analytical platforms.

To obtain a comprehensive and continuous daily dataset from 1950 to 2023, users can combine our reconstructed values with the original historical observational data. Clear instructions and links for downloading the original observational data used in this study can be found at https://github.com/zhwang24/RADIT-Reconstructed-Arctic-River-Data.

Each table file is named according to the river and station it represents (e.g., Lena__Kyusyur.csv) to ensure straightforward identification. Each file contains the following columns:

(1) Date: A timestamp in the YYYY-MM-DD format, covering the entire period from January 1, 1950, to December 31, 2023.
(2) Station discharge: The daily reconstructed discharge and its uncertainty at the monitoring station (units: m^3/s).
(3) Mouth discharge: The daily reconstructed discharge and its uncertainty at the river mouth (units: m^3/s).
(4) Mouth temperature: The daily reconstructed water temperature and its uncertainty at the river mouth (units: °C).
(5) Mouth heat flux: The daily reconstructed heat flux and its uncertainty at the river mouth (units: MJ).

Technical Validation

Validation of the reconstructed river discharge. To evaluate the reliability and accuracy of our discharge dataset, we conducted a comprehensive validation using the testing data previously derived from the *in situ* observations recorded at the 25 gauging stations. The validation was performed by comparing our estimated values with these independent measurements across rivers of different sizes.

The validation results demonstrated the robust performance of our machine learning framework in reconstructing missing discharge data across rivers.

Given that separate models were developed for each river to account for their unique hydrological characteristics, we evaluated the model performance on a river-by-river basis rather than providing aggregated statistics. The detailed performance metrics for each river are shown in Fig. 6. The results revealed a consistently high reconstruction accuracy across the rivers studied. The NSE values ranged from 0.520 to 0.934, with a median value of 0.861, indicating excellent model performance. Notably, 21 out of 25 rivers yielded NSE values above 0.8. Similarly, the KGE results demonstrated remarkable agreement between the reconstructed and observed discharge patterns, with a median value of 0.870. Twenty rivers yielded KGE values exceeding 0.8, and the lowest KGE value was 0.592. Furthermore, the NRMSE values confirmed the high accuracy of our reconstruction, with

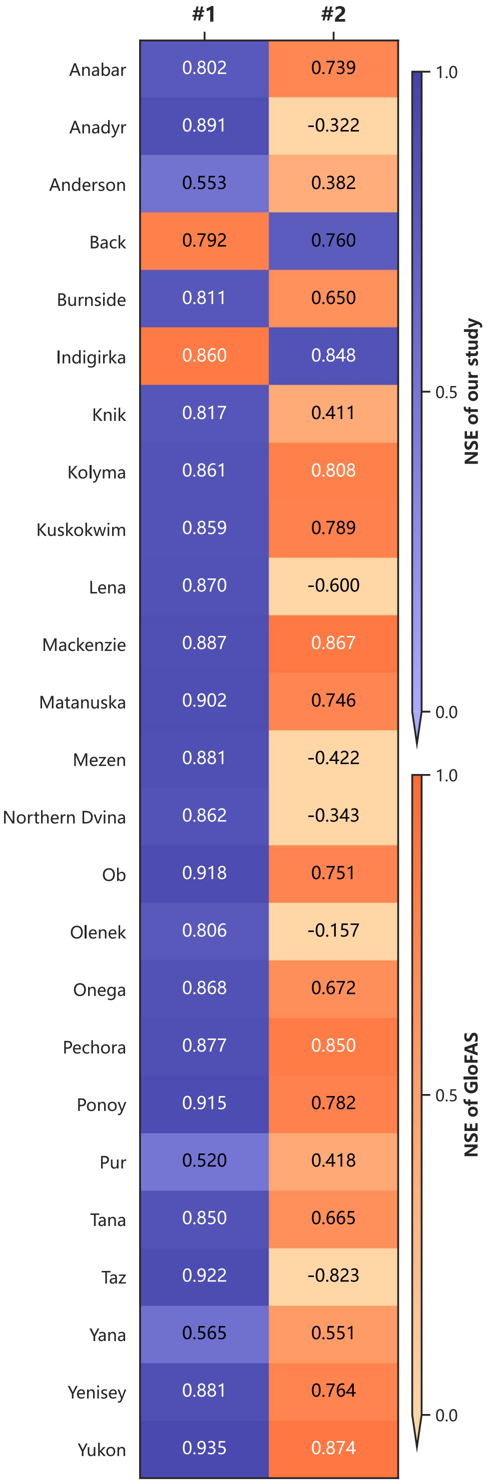

Fig. 8 Comparison of NSE performance between this study and the GloFAS product across 25 Arctic-draining rivers. Each river is represented by a pair of bars showing the better- and lower-performing products. The color indicates the data source (blue for this study and orange for the GloFAS), and depth reflects NSE values.

a median value of 6.1% and a maximum error of 11.6%, indicating that the reconstructed values closely match the observed discharge patterns across all rivers. Among the lower-performing rivers, such as Yana, Anderson, and Pur, reduced accuracy was associated with factors including extreme discharge events in the testing period or notable shifts in baseflow levels relative to those in the training period. These cases highlight the challenges of reconstructing discharge under non-stationary hydrological conditions.

In addition to evaluating overall discharge accuracy, we further assessed the model's ability to reconstruct peak discharge events, which are particularly important for hydrological extremes but are inherently more

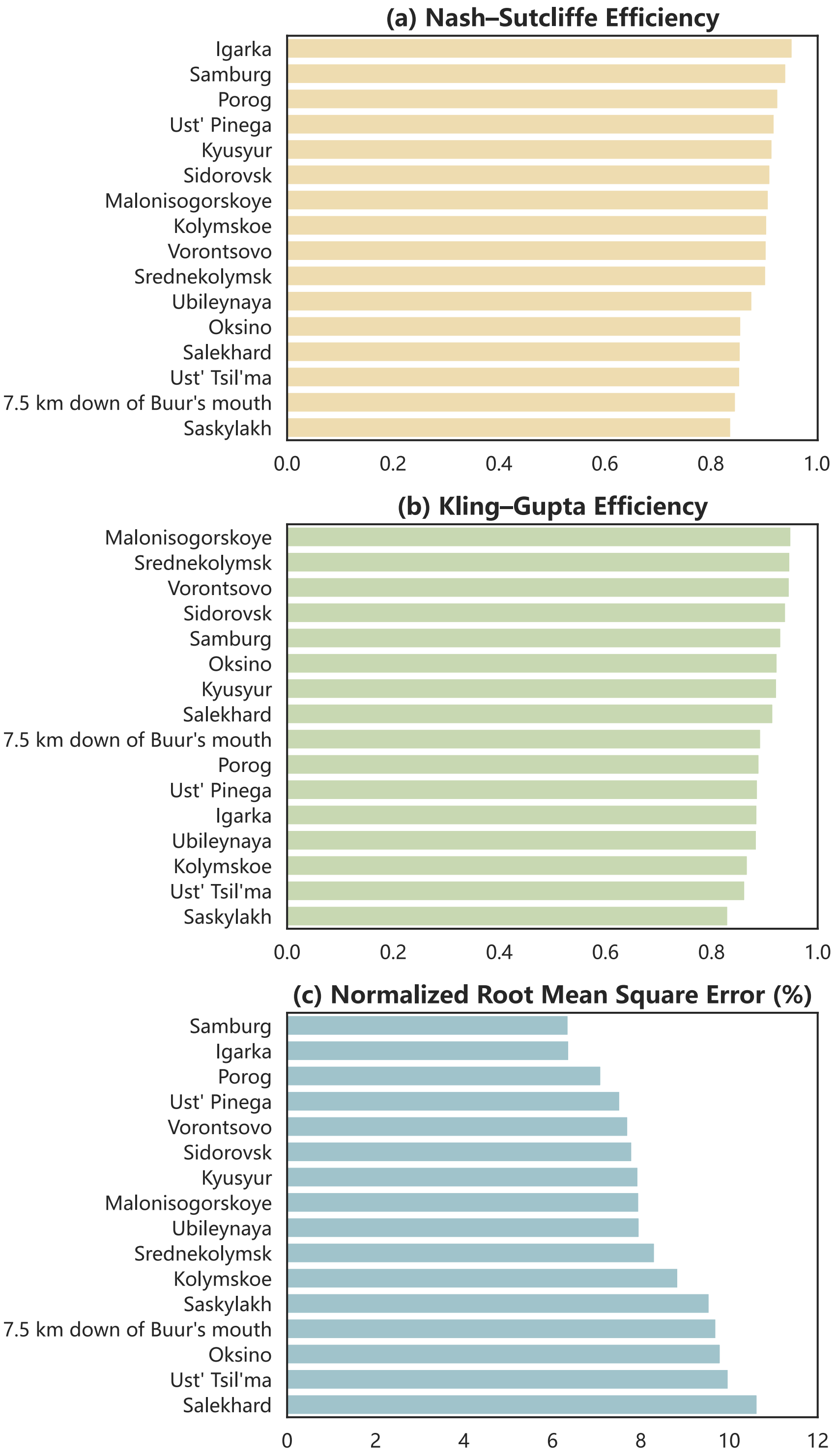

Fig. 9 Statistical metrics of the XGBoost-based temperature reconstruction model across 16 stations using leave-one-out cross-validation: (**a**) NSE, (**b**) KGE, and (**c**) NRMSE.

challenging to model. Here, we defined peak discharge for each river as values exceeding the 90th percentile within the training dataset. The NSE values calculated specifically for these peak flows are summarized in Fig. 7. The results highlight the inherent difficulty of modeling extreme flows, particularly in data-sparse Arctic basins. Among all the rivers, 16 exhibited positive NSE values for peak discharge. The overall median NSE was 0.186, with the highest value observed for the Mezen River (NSE = 0.663), indicating strong agreement for peak flow events. For rivers where the model performed less well in terms of peak flows, the NSE values generally fell between −1 and 0. Only the Ob River yielded an NSE below −1, suggesting relatively poor skill in capturing extreme discharge magnitudes in that basin. Nevertheless, the results demonstrate that the model captures peak discharge patterns with reasonable accuracy for a majority of rivers, further supporting the robustness of the reconstruction framework.

To further assess the effectiveness of our reconstruction framework, we compared it against the Global Flood Awareness System (GloFAS) product[55] (https://doi.org/10.24381/cds.a4fdd6b9), an operational global discharge dataset developed by the Copernicus Emergency Management Service (CEMS). GloFAS Version 4 provides daily river discharge estimates at a spatial resolution of 0.05° × 0.05° from 1979 to present, covering all global river systems. Figure 8 presents a comparative analysis of NSE values between our reconstructed dataset and the GloFAS across the 25 Arctic-draining rivers. For each river, the method yielding the higher NSE is shown on the left, while the lower-performing product is shown on the right. The color hues represent the different data sources (blue for this study, orange for GloFAS), and the color intensity reflects the magnitude of the NSE. The results indicate that our machine learning-based reconstruction outperforms the GloFAS on 23 of the 25 rivers. The results of the GloFAS slightly exceeded those of our method only for the Back and Indigirka Rivers. These validation results collectively demonstrate that our machine learning framework can effectively reconstruct missing daily discharge data while maintaining high accuracy and reliability across rivers that differ substantially in discharge magnitude, catchment size, and climate conditions.

Validation of the reconstructed river temperature. For river temperature reconstruction, a similar machine learning approach was adopted but with a different validation strategy due to the limited data availability. Based on the cross-validation performance across all the candidate models, XGBoost outperformed the other models and was selected as the final algorithm for temperature reconstruction. As previously described, we implemented a leave-one-out cross-validation approach, where the data of each station were iteratively used as validation data, while the data from the remaining stations served as training data.

The cross-validation results demonstrated reliable reconstruction accuracy (Fig. 9). The NSE values ranged from 0.838 to 0.954, with a median value of 0.906, and the KGE values ranged from 0.831 to 0.950, with a median value of 0.905. The NRMSE values ranged from 6.4% to 10.6%, with a median of 8.0%. These consistently high values of the performance metrics across the different validation stations indicate that our unified model has favorable adaptability and generalizability. This notable performance suggests that the model can be reliably applied to reconstruct river temperatures at ungauged locations where observational data are not available.

To further evaluate the geographical transferability of our unified model, which was trained via Russian river observations prior to 2003, we validated the model results against temperature data from the Pilot Station on the Yukon River in Alaska (the only station in the North American Arctic meeting our validation data requirements). The validation was performed using 1,609 daily observations from the post-2014 period. The validation yielded satisfactory results, with an NSE value of 0.885, a KGE value of 0.754, and an NRMSE value of 8.1%, demonstrating strong performance across both the spatial and temporal domains.

In conclusion, this study presents a comprehensive machine learning-based reconstruction of Arctic river discharge, temperature, and heat flux data, making meaningful progress toward improving data availability in regions with limited hydrological observations. The robust performance of our reconstruction approach, demonstrated by consistently high validation metrics across different temporal and spatial scales, ensures the reliability of this dataset for various research applications. The continuous daily records spanning seven decades, combined with extensive spatial coverage across the pan-Arctic region, make it a valuable resource for validating and improving climate models, especially in their representation of land–ocean interactions in the Arctic region. These advances in data availability will support more comprehensive analyses of Arctic environmental changes and their global implications. While this dataset covers the most hydrologically significant Arctic rivers, we acknowledge that contributions from smaller or ungauged basins remain unaccounted for. Future efforts may build upon this work by developing pan-Arctic statistical models that incorporate basin characteristics to extend estimates to unmonitored regions, thereby enabling more complete assessments of Arctic freshwater budgets.

Code availability

The custom GEE and Python scripts used in this study are available under the Massachusetts Institute of Technology (MIT) license. The repository at https://github.com/zhwang24/RADIT-Reconstructed-Arctic-River-Data provides script names and inline comments to support their application, covering three main aspects: (1) ERA5-Land data download: GEE scripts for retrieving daily ERA5-Land meteorological and hydrological data used in the reconstruction process. (2) Discharge, temperature, and heat flux reconstruction: Python scripts for performing the machine learning-based reconstructions of missing data, extending the time series, and generating the final RADIT dataset. (3) Integrating original observations for a complete dataset: Clear instructions and Python scripts for obtaining the original historical river discharge data from their publicly available sources and for seamlessly merging them with the reconstructed data provided in the RADIT dataset, enabling users to assemble a comprehensive and continuous daily record.

Received: 4 December 2024; Accepted: 8 July 2025;
Published online: 17 July 2025

References

1. Jakobsson, M. Hypsometry and volume of the Arctic Ocean and its constituent seas. *Geochem Geophys Geosyst* **3**, 1–18 (2002).
2. Döscher, R., Vihma, T. & Maksimovich, E. Recent advances in understanding the Arctic climate system state and change from a sea ice perspective: a review. *Atmos. Chem. Phys.* **14**, 13571–13600 (2014).
3. Docquier, D. & Koenigk, T. A review of interactions between ocean heat transport and Arctic sea ice. *Environ. Res. Lett.* **16**, 123002 (2021).

4. Shiklomanov, I. A. & Shiklomanov, A. I. Climatic Change and the Dynamics of River Runoff into the Arctic Ocean. *Water Resources* **30**, 593–601 (2003).
5. Nummelin, A., Ilicak, M., Li, C. & Smedsrud, L. H. Consequences of future increased Arctic runoff on Arctic Ocean stratification, circulation, and sea ice cover. *JGR Oceans* **121**, 617–637 (2016).
6. Park, H. *et al.* Increasing riverine heat influx triggers Arctic sea ice decline and oceanic and atmospheric warming. *Science Advances* **6**, eabc4699 (2020).
7. Wild, B. *et al.* Rivers across the Siberian Arctic unearth the patterns of carbon release from thawing permafrost. *Proc. Natl. Acad. Sci. USA.* **116**, 10280–10285 (2019).
8. Clark, J. B., Mannino, A., Spencer, R. G. M., Tank, S. E. & McClelland, J. W. Quantification of Discharge-Specific Effects on Dissolved Organic Matter Export From Major Arctic Rivers From 1982 Through 2019. *Global Biogeochemical Cycles* **37**, e2023GB007854 (2023).
9. Liu, M. *et al.* Global riverine land-to-ocean carbon export constrained by observations and multi-model assessment. *Nat. Geosci.* **17**, 896–904 (2024).
10. McClelland, J. W. *et al.* Particulate organic carbon and nitrogen export from major Arctic rivers. *Global Biogeochemical Cycles* **30**, 629–643 (2016).
11. Shiklomanov, A. I. & Lammers, R. B. River ice responses to a warming Arctic—recent evidence from Russian rivers. *Environ. Res. Lett.* **9**, 035008 (2014).
12. Yang, D., Park, H., Prowse, T., Shiklomanov, A. & McLeod, E. River Ice Processes and Changes Across the Northern Regions. in *Arctic Hydrology, Permafrost and Ecosystems* (eds. Yang, D. & Kane, D. L.) 379–406, https://doi.org/10.1007/978-3-030-50930-9_13 (Springer International Publishing, Cham, 2021).
13. Irrgang, A. M. *et al.* Drivers, dynamics and impacts of changing Arctic coasts. *Nat Rev Earth Environ.* **3**, 39–54 (2022).
14. O'Donnell, J., Douglas, T., Barker, A. & Guo, L. Changing Biogeochemical Cycles of Organic Carbon, Nitrogen, Phosphorus, and Trace Elements in Arctic Rivers. in *Arctic Hydrology, Permafrost and Ecosystems* (eds. Yang, D. & Kane, D. L.) 315–348, https://doi.org/10.1007/978-3-030-50930-9_11 (Springer International Publishing, Cham, 2021).
15. Ahmed, R., Prowse, T., Dibike, Y., Bonsal, B. & O'Neil, H. Recent Trends in Freshwater Influx to the Arctic Ocean from Four Major Arctic-Draining Rivers. *Water* **12**, 1189 (2020).
16. Feng, D. *et al.* Recent changes to Arctic river discharge. *Nat Commun* **12**, 6917 (2021).
17. Lammers, R. B., Pundsack, J. W. & Shiklomanov, A. I. Variability in river temperature, discharge, and energy flux from the Russian pan-Arctic landmass. *J. Geophys. Res.* **112** (2007).
18. Yang, D., Ye, B. & Kane, D. L. Streamflow changes over Siberian Yenisei River Basin. *Journal of Hydrology* **296**, 59–80 (2004).
19. Yang, D., Shrestha, R. R., Lung, J. L. Y., Tank, S. & Park, H. Heat flux, water temperature and discharge from 15 northern Canadian rivers draining to Arctic Ocean and Hudson Bay. *Global and Planetary Change* **204**, 103577 (2021).
20. Shiklomanov, A. I., Lammers, R. B. & Vörösmarty, C. J. Widespread decline in hydrological monitoring threatens Pan-Arctic Research. *EoS Transactions* **83**, 13–17 (2002).
21. Hannah, D. M. *et al.* Large-scale river flow archives: importance, current status and future needs. *Hydrological Processes* **25**, 1191–1200 (2011).
22. Yang, D., Marsh, P. & Ge, S. Heat flux calculations for Mackenzie and Yukon Rivers. *Polar Science* **8**, 232–241 (2014).
23. Lammers, R. B., Shiklomanov, A. I., Vörösmarty, C. J., Fekete, B. M. & Peterson, B. J. Assessment of contemporary Arctic river runoff based on observational discharge records. *J. Geophys. Res.* **106**, 3321–3334 (2001).
24. Winkelbauer, S. *et al.* Diagnostic evaluation of river discharge into the Arctic Ocean and its impact on oceanic volume transports. *Hydrol. Earth Syst. Sci.* **26**, 279–304 (2022).
25. Ye, B., Yang, D. & Kane, D. L. Changes in Lena River streamflow hydrology: Human impacts versus natural variations. *Water Resources Research* **39** (2003).
26. Nghiem, S. V., Hall, D. K., Rigor, I. G., Li, P. & Neumann, G. Effects of Mackenzie River discharge and bathymetry on sea ice in the Beaufort Sea. *Geophysical Research Letters* **41**, 873–879 (2014).
27. Durocher, M., Requena, A. I., Burn, D. H. & Pellerin, J. Analysis of trends in annual streamflow to the Arctic Ocean. *Hydrological Processes* **33**, 1143–1151 (2019).
28. Peterson, B. J. *et al.* Increasing River Discharge to the Arctic Ocean. *Science* **298**, 2171–2173 (2002).
29. Hamman, J. *et al.* The coastal streamflow flux in the Regional Arctic System Model. *JGR Oceans* **122**, 1683–1701 (2017).
30. Bring, A. *et al.* Arctic terrestrial hydrology: A synthesis of processes, regional effects, and research challenges. *JGR Biogeosciences* **121**, 621–649 (2016).
31. Park, H., Kim, Y., Suzuki, K. & Hiyama, T. Influence of snowmelt on increasing Arctic river discharge: numerical evaluation. *Prog Earth Planet Sci* **11**, 13 (2024).
32. McClelland, J. W., Holmes, R. M., Peterson, B. J. & Stieglitz, M. Increasing river discharge in the Eurasian Arctic: Consideration of dams, permafrost thaw, and fires as potential agents of change. *J. Geophys. Res.* **109** (2004).
33. Wang, P. *et al.* Potential role of permafrost thaw on increasing Siberian river discharge. *Environ. Res. Lett.* **16**, 034046 (2021).
34. McClelland, J. W., Tank, S. E., Spencer, R. G. M. & Shiklomanov, A. I. Coordination and Sustainability of River Observing Activities in the Arctic. *ARCTIC* **68**, 59 (2015).
35. Holmes, R. M. *et al.* NOAA Arctic Report Card 2021: River Discharge. *NOAA Arctic Report Card 2021* https://doi.org/10.25923/ZEVF-AR65 (2021).
36. ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) https://doi.org/10.24381/CDS.E2161BAC (2019).
37. Wang, Y.-R., Hessen, D. O., Samset, B. H. & Stordal, F. Evaluating global and regional land warming trends in the past decades with both MODIS and ERA5-Land land surface temperature data. *Remote Sensing of Environment* **280**, 113181 (2022).
38. Xu, J., Ma, Z., Yan, S. & Peng, J. Do ERA5 and ERA5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China. *Journal of Hydrology* **605**, 127353 (2022).
39. Vorosmarty, C. J. & Fekete, B. M. ISLSCP II river routing data (STN-30p). *ORNL DAAC* https://doi.org/10.3334/ORNLDAAC/1005 (2011).
40. Allen, G. H. & Pavelsky, T. M. Global River Widths from Landsat (GRWL) Database. *Zenodo* https://doi.org/10.5281/ZENODO.1269594 (2018).
41. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
42. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** (2001).
43. Chen, T. & Guestrin, C. XGBoost. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794, https://doi.org/10.1145/2939672.2939785 (ACM, San Francisco California USA, 2016).
44. Wu, J., Li, Y. & Ma, Y. Comparison of XGBoost and the Neural Network model on the class-balanced datasets. in *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)* 457–461, https://doi.org/10.1109/ICFTIC54370.2021.9647373 (IEEE, Greenville, SC, USA, 2021).
45. Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. in *Advances in neural information processing systems* (eds. Guyon, I. *et al.*) vol. 30 (Curran Associates, Inc., 2017).

46. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology* **10**, 282–290 (1970).
47. Gupta, H. V. & Kling, H. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research* **47**, 2011WR010962 (2011).
48. Cardot, R., Moradi, G., Fatichi, S., Molnar, P. & Lane, S. Basin-scale temporal evolution of the discharge and angular momentum ratios at confluences: The case of the Upper-Rhône watershed. in *EGU general assembly conference abstracts* 18902 (2018).
49. Douinot, A., Iffly, J. F., Tailliez, C., Meisch, C. & Pfister, L. Flood patterns in a catchment with mixed bedrock geology and a hilly landscape: identification of flashy runoff contributions during storm events. *Hydrol. Earth Syst. Sci.* **26**, 5185–5206 (2022).
50. Quan, S. J. Comparing hyperparameter tuning methods in machine learning based urban building energy modeling: A study in Chicago. *Energy and Buildings* **317**, 114353 (2024).
51. Dai, A. & Trenberth, K. E. Estimates of Freshwater Discharge from Continents: Latitudinal and Seasonal Variations. *Journal of Hydrometeorology*, **3**, 660-687 (2002).
52. van Vliet, M. T. H., Ludwig, F., Zwolsman, J. J. G., Weedon, G. P. & Kabat, P. Global river temperatures and sensitivity to atmospheric warming and changes in river flow. *Water Resources Research* **47** (2011).
53. Elshin, Y. River heat runoff in the European part of Russia. *Meteorol. Hydrol* **9**, 85–93 (1981).
54. Wang, Z., Hui, F. & Cheng, X. RADIT: A Machine Learning-Reconstructed Dataset of River Discharge, Temperature, and Heat Flux into the Arctic Ocean. *Zenodo* https://doi.org/10.5281/ZENODO.15811422 (2025).
55. Zsoter, E. River discharge historical data from the Global Flood Awareness System. ECMWF https://doi.org/10.24381/CDS.A4FDD6B9 (2019).
56. Muñoz-Sabater, J. *et al.* ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
57. Allen, G. H. & Pavelsky, T. M. Global extent of rivers and streams. *Science* **361**, 585–588 (2018).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 42376241). This research was supported by multiple hydrological databases. For the river discharge data, we acknowledge the Arctic Great Rivers Observatory (ArcticGRO), the Global Runoff Data Centre (GRDC), the Water Survey of Canada (WSC), ArcticRIMS, and R-ArcticNet. The river temperature data were obtained from the ART-Russia dataset[17]. We also thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the ERA5-Land reanalysis data[56] and acknowledge the Simulated Topological Network (STN-30p)[39], Global River Widths from Landsat (GRWL; Allen and Pavelsky[57]), and GloFAS[55] datasets used in this study.

Author contributions

W.Z. conducted the primary analysis and wrote the first draft of the manuscript. H.F. conceptualized and designed the research, provided funding support, and participated in the manuscript review and revision. C.X. participated in the manuscript review and revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05582-9.

Correspondence and requests for materials should be addressed to F.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025