https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Fusing Satellite Embeddings to Improve Streamflow Reconstruction Across River Networks

Haomei Lin[1], Peirong Lin[1,*], Fenghe Zhang[1], Louise Slater[2], Yuan Yang[3], Ming Pan[3], Qiming Qin[1], and Aizhong Hou[4]

[1]Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[2]School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, United Kingdom
[3]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, CA, USA
[4]Hydrological Forecast Center, Information Center of the Ministry of Water Resources of China, Beijing, 100053, China

Correspondence: Peirong Lin (peironglinlin@pku.edu.cn)

Abstract.

Reconstructing streamflow across river networks is increasingly challenging in the context of heavily modified land surface conditions. Here we present a Data Integration model with Satellite Embeddings (DISE), a reach-scale residual-learning framework that integrates Google Satellite Embeddings (SE; compact learned vector representations of satellite imagery) from the AlphaEarth Foundation Model with a recently developed discharge product (GRADES-hydroDL) by learning corrections toward gauge observations. We evaluate DISE at 41 gauging stations in the Yangtze River Basin using leave-one-station-out cross-validation, with embeddings aggregated over each reach's contributing subcatchment. Simulations incorporating SE consistently outperform the GRADES-hydroDL baseline, with mean aggregation emerging as the most balanced strategy. Improvements are most pronounced for magnitude and bias: compared to GRADES-hydroDL, median KGE increases from 0.485 to 0.594 and median NSE from 0.301 to 0.533, while correlation gains remain modest, suggesting that DISE primarily captures streamflow volume and variability rather than timing. Control experiments further show that SE enhance spatial generalization beyond both meteorological forcings and traditional hydro-environmental reach attributes (RiverATLAS): compared to the base configuration without spatial context, adding SE alone increases median KGE from 0.473 to 0.594; when SE are further added on top of RiverATLAS, median KGE increases from 0.497 to 0.567. Once SE are included, adding RiverATLAS can even slightly reduce performance. Embedding-driven gains weaken where streamflow is governed by processes not directly visible from surface imagery, particularly complex reservoir operations. Nevertheless, SE can still provide useful information when forcing-based corrections are limited. These results demonstrate that SE provide analysis-ready, information-rich representations of land surface heterogeneity that measurably strengthen streamflow reconstruction across river networks. DISE offers a scalable pathway to inject high-resolution Earth observation context into river-network modeling, improving predictions in basins where conventional forcings and hydro-environmental descriptors are often insufficient.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

1 Introduction

Streamflow integrates the effects of climate forcing and land-surface processes across river networks, reflecting their combined influence on water availability, hydrologic extremes, and ecosystem functioning. Reliable streamflow simulation is therefore fundamental for water resources assessment, flood and drought risk management, and large-scale hydro-environmental studies (Gudmundsson et al., 2026). Increasingly, these applications require models that generalize across space, including along river networks where in situ observations are sparse or unevenly distributed.

Achieving strong spatial generalization remains challenging because streamflow is controlled by spatially heterogeneous conditions that vary among reaches and subcatchments (Joseph et al., 2025). Land-surface properties such as vegetation, cropland management, irrigation, and urban expansion can reshape runoff generation and routing, while river regulation further alters flow regimes (Lin et al., 2026; Ishikawa et al., 2025; Su et al., 2019). At the global scale, human modification of rivers is widespread (Best, 2019), with only about 37% of rivers remaining free-flowing over their full length (Grill et al., 2019). These spatially varying influences can lead to strongly non-uniform simulation errors across the network, so approaches that perform well at monitored sites may degrade when transferred to new locations.

Two major modeling paradigms have been used to address these challenges, but both face limitations in representing spatial heterogeneity at scale. Process-based global hydrologic and land-surface modeling frameworks typically encode spatial heterogeneity through prescribed gridded parameter fields and subgrid tiling schemes derived from land cover, soils, and topographic datasets. For example, PCR-GLOBWB 2.0 operates at 5 arcmin resolution and represents spatial variability using mapped land-cover fractions and spatially varying soil and groundwater properties to simulate runoff generation and river routing (Sutanudjaja et al., 2018). Land-surface models adopt a similar philosophy: Noah-MP and CLM represent within-grid heterogeneity via vegetation or plant-functional-type tiles, with parameters assigned based on land-cover classes and soil texture (Niu et al., 2011; Lawrence et al., 2019). While physically interpretable, these approaches often rely on simplified and class-based representations of land-surface variability, and may not fully exploit the rapidly growing volume of high-resolution Earth-observation information (Casu et al., 2017). Recently, machine-learning approaches, particularly LSTM-based rainfall-runoff models, have demonstrated strong predictive skill and have been applied to regionalization and pseudo-ungauged settings (Kratzert et al., 2019). However, the land surface information provided to these models is commonly limited to a small set of static catchment attributes that are appended to the meteorological time series (Kratzert et al., 2018; Nearing et al., 2024; Yang et al., 2025a). Such descriptors provide only coarse summaries and cannot fully represent fine-scale, spatially heterogeneous land-surface signals that may be critical for robust spatial transfer.

Satellite embeddings (SE) offer a new pathway to explicitly introduce rich land-surface information into hydrologic prediction through embedding representations. The Google Satellite Embedding V1 dataset, produced by the AlphaEarth Foundation model, provides global, 10 m annual embeddings in which each pixel is represented by a 64-dimensional vector that summarizes multi-sensor surface conditions (Brown et al., 2025). By compressing diverse Earth-observation streams into a compact representation, SE provide analysis-ready features that can encode detailed land-surface patterns and human disturbance signatures that are difficult to incorporate using traditional hydro-environmental datasets.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Here, we investigate whether and how SE can improve the spatial generalization of reach-scale streamflow simulation. We develop a Data Integration model with Satellite Embeddings (DISE), which fuses satellite embeddings with a strong baseline reach-scale discharge product. Using the Yangtze River Basin as a testbed, we evaluate DISE under a leave-one-station-out design to directly assess station-transfer skill, and we conduct controlled experiments to separate the contributions of meteorological forcings and embeddings and to compare embeddings against commonly used hydro-environmental reach attributes (Linke et al., 2019). Our study addresses three questions:

1. Do annual satellite embeddings provide measurable improvements in reach-scale streamflow reconstruction and station-transfer performance beyond a strong baseline discharge product?
2. How should high-dimensional pixel-level embeddings be aggregated to maximize robustness and skill?
3. How do meteorological forcings, satellite embeddings, and other factors (e.g., river regulation) jointly control station-wise reconstruction gains?

2 Data and Method

The schematic representation of the proposed methodology is illustrated in Fig. 1, and the corresponding datasets and methods are described in detail below.

2.1 Data Integration Framework

We develop a Data Integration model with Satellite Embeddings (DISE), a residual-learning framework that fuses SE with a recently developed discharge product (i.e., GRADES-hydroDL (Yang et al., 2025a)) to reconstruct streamflow across river networks. We adopt GRADES-hydroDL as a strong first-guess simulation because it provides global all-reach daily discharge estimates and it achieves strong performance at gauged reaches worldwide (median KGE of the spatial-temporal test: 0.653). However, models that rely primarily on temporal input sequences learn runoff dynamics from meteorological forcing but represent spatial heterogeneity only through a small set of static attributes. As a result, their skill can degrade under strict spatial transfer or out-of-region evaluation, motivating a more explicit and information-rich representation of landscape heterogeneity. DISE therefore treats GRADES-hydroDL as a baseline and learns a data-driven correction toward gauge observations, preserving the strengths of the baseline product while compensating for locally varying, reach-specific errors. Since SE are annual, we pair them with daily meteorological forcings (CMFD v2) (He et al., 2020) to enable daily reconstruction. To harmonize heterogeneous inputs on the river network, daily gridded meteorological variables are mapped to each reach by area-weighted averaging over its contributing subcatchment, reflecting the relatively coarse resolution of meteorological grids compared with the river-network representation (Lin et al., 2018). Satellite embeddings (SE), which provide much finer-grained land surface information, are aggregated within the same subcatchment using statistical summary metrics to generate reach-level predictors. We train DISE in log space by learning the residual as $y = \log(Q_{\mathrm{obs}}) - \log(Q_{\mathrm{GRADES\text{-}hydroDL}})$, which reduces the dominance of extreme high flows (Yang et al., 2025b) and encourages stable, relative adjustments across seasons and stations.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Our study uses five key datasets. (1) Satellite embeddings are obtained from Google Satellite Embedding V1 (10 m) (Brown et al., 2025) via the Google Earth Engine platform. Each pixel is encoded as a 64-dimensional learned representation derived from multi-source Earth observation data for a given calendar year. Unlike conventional remote-sensing predictors based on individual bands, indices, or land-cover classes, these embeddings provide a compact description of land-surface conditions by integrating spatial, temporal, and multi-sensor information (e.g., optical observations from Sentinel-2 and Landsat 8/9, LiDAR data from GEDI, and environmental context from GLO-30, ERA5-Land, and GRACE). They therefore have the potential to capture fine-scale heterogeneity in vegetation, urban surfaces, water presence, and other surface characteristics relevant to runoff generation and streamflow response. (2) Daily meteorological forcings are taken from the observation-based, gauge-adjusted China Meteorological Forcing Dataset v2.0 (CMFDv2) (He et al., 2020), which provides gridded near-surface variables at a horizontal spatial resolution of 0.1°, including air temperature, surface pressure, specific humidity, relative humidity, wind speed, downward shortwave radiation, downward longwave radiation, and precipitation. (3) Baseline discharge $Q_{\mathrm{GRADES\text{-}hydroDL}}$ is provided by GRADES-hydroDL, an improved global reach-level daily discharge product spanning 1980 to the near present (Yang et al., 2025a). It is generated using the Grid LSTM-RAPID framework, in which an LSTM first estimates runoff at the global 0.25° grid scale from meteorological forcings and land-surface predictors, and RAPID subsequently routes the simulated runoff through the river network to produce reach-level discharge. The model uses MSWEP precipitation, ERA5 meteorological variables, and monthly leaf area index as dynamic inputs, and its LSTM component is trained on discharge observations from 4215 selected gauged basins worldwide before being applied globally. GRADES-hydroDL is routed on the MERIT-Basins vector river network, which is derived from the 3-arcsec (∼90 m) MERIT-Hydro hydrography and delineates ∼2.94 million reaches using a 25 km^2 channel initiation threshold (Lin et al., 2019). Globally, this network has a median reach length of 6.8 km, providing a relatively fine river-network representation for large-scale routing. Accordingly, the reaches and contributing subcatchments used in this study follow the same MERIT-Basins hydrography underlying GRADES-hydroDL, ensuring spatial consistency between the baseline discharge product and our reach-level predictor aggregation. (4) The locations of gauges are taken from GSHA (Yin et al., 2024b), a global archive of quality-controlled gauge records compiled from multiple hydrometric sources. Daily in situ streamflow of these gauges is provided by the Information Center of the Ministry of Water Resources of China and is used as the training and evaluation target. (5) Finally, we selected forest, cropland, and urban fractions from RiverATLAS, a global dataset of hydro-environmental attributes linked to individual river reaches (Linke et al., 2019). RiverATLAS provides standardized reach-scale and upstream-accumulated descriptors derived from global source datasets, covering hydrology, physiography, climate, land cover and land use, soils and geology, and anthropogenic influences at high spatial resolution. This choice was motivated by the fact that satellite embeddings are expected to encode rich land-surface information, while these land-cover fractions are also widely used as representative descriptors of human-influenced surface conditions in large-sample hydrological datasets (Yin et al., 2024b). We therefore used these reach attributes as a representative set of traditional hydro-environmental descriptors in a control experiment to examine whether satellite embeddings provide predictive information beyond conventional reach-scale covariates.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

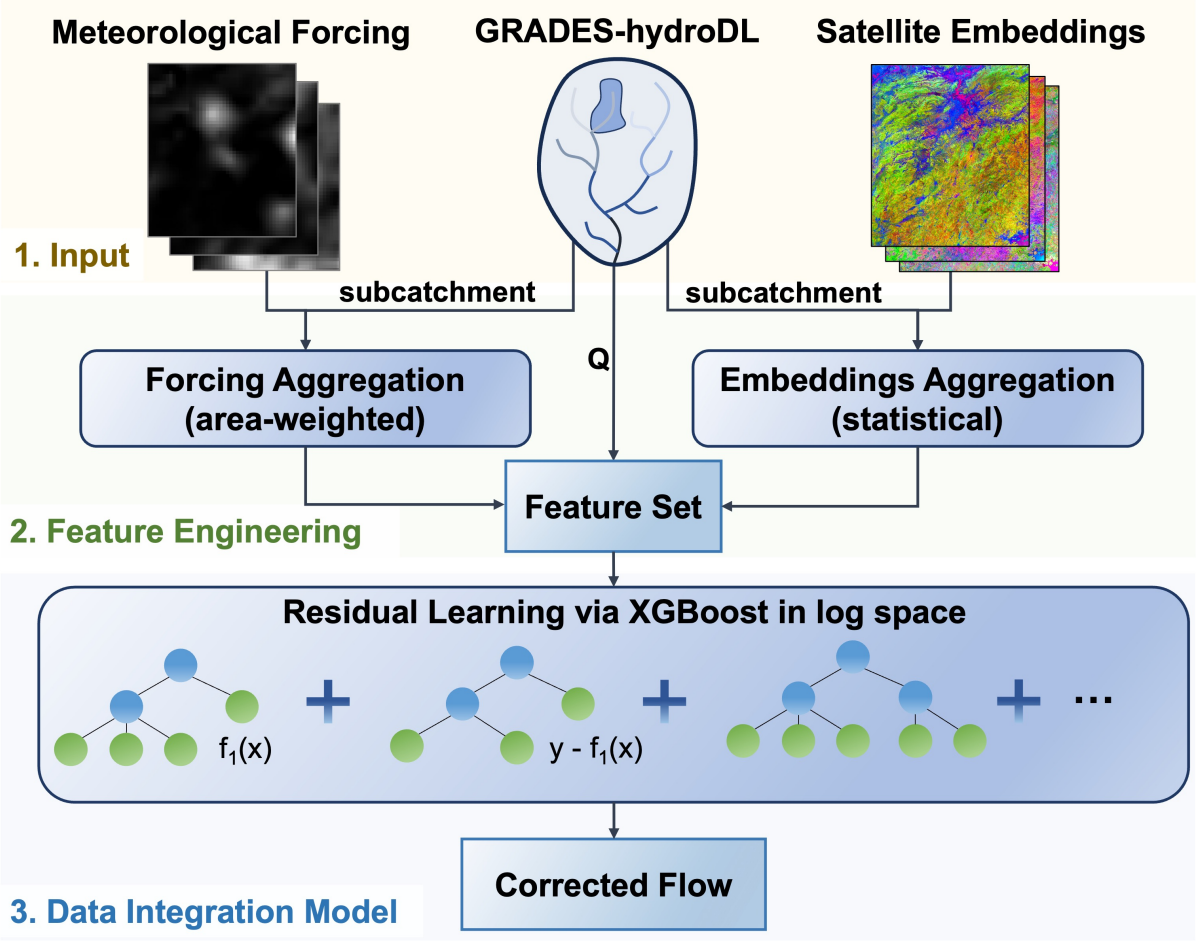

Figure 1. Framework of the Data Integration model with Satellite Embeddings (DISE). DISE integrates satellite embeddings paired with daily meteorological forcings (CMFDv2) and GRADES-hydroDL to reconstruct simulated discharge. Meteorological forcings are aggregated to each subcatchment by area-weighted averaging and satellite embeddings are summarized within each subcatchment using statistical metrics to form a unified feature set. An XGBoost model is then trained in log space to learn the residual between observed and baseline discharge. The predicted log-residual is finally transformed back to discharge space and used to correct the baseline simulation, yielding the final streamflow estimate.

2.2 Experimental Setting

120 Because SE primarily provide spatially heterogeneous information, we evaluate all model configurations using leave-one-station-out (LOSO) cross-validation, where each gauge is withheld in turn and treated as pseudo-ungauged. This protocol is widely used to assess spatial transferability and prediction skill under ungauged or poorly gauged conditions (Pool et al., 2021), which is also essential for streamflow modelling across river networks. Accordingly, our evaluation focuses on intra-

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Table 1. Experiments settings.

No.	Hps.	Experiment Name	Predictors
1	Fixed	Strategy	Base predictors (CMFDv2 daily forcings (air temperature, surface pressure, specific humidity, relative humidity, wind speed, downwelling shortwave radiation, downwelling longwave radiation, precipitation) + GRADES-hydroDL simulated discharge Q, and their 1–7 d lags.) + satellite embeddings (SE) aggregated using alternative summary statistics (mean, std, skew, and their combinations).
2	Tuned	Base	CMFDv2 daily forcings (air temperature, surface pressure, specific humidity, relative humidity, wind speed, downwelling shortwave radiation, downwelling longwave radiation, precipitation) + GRADES-hydroDL simulated discharge Q, and their 1–7 d lags.
		Base+E (DISE)	Base predictors + SE.
3	Tuned	Base+ATLAS	Base predictors + RiverATLAS reach attributes (forest, cropland, and urban fractions at subcatchment and upstream scales: `for_pc_cse`, `for_pc_use`, `crp_pc_cse`, `crp_pc_use`, `urb_pc_cse`, `urb_pc_use`).
		Base+ATLAS+E	Base+ATLAS attributes + SE.

* Hps. denotes hyperparameters and detailed hyperparameters for each experiment are listed in Table A1.

basin spatial heterogeneity and cross-location transferability, rather than temporal robustness. All experiments, including model training, hyperparameter tuning, and LOSO evaluation, are conducted using data from 2017. We include only stations with more than 300 days of observations in 2017 to ensure sufficiently complete daily discharge records.

Under LOSO, we conduct three sets of experiments (Table 1). Because SE are 10m-resolution, high-dimensional gridded layers, we first investigate how to incorporate them into river-network modeling by benchmarking different subcatchment-level aggregation strategies (mean, standard deviation, skewness, and their combinations). To isolate the effect of aggregation choices and to select a robust scheme for subsequent analyses, we fix the hyperparameter configuration across all strategies, using default settings with mild stochastic subsampling (`subsample`=0.8, `colsample_bytree`=0.8) to improve robustness under LOSO. This choice follows stochastic gradient boosting and XGBoost guidance that subsampling provides effective regularization and reduces overfitting (Friedman, 2002). And this design ensures that performance differences are primarily attributable to the aggregation scheme rather than to differences in model tuning. Second, to isolate the value of SE beyond meteorological information, we conduct a control experiment comparing the Base and Base+E configurations. These two models share identical meteorological forcings and GRADES-hydroDL discharge inputs and differ only in whether SE are included. Third, we test whether SE provides predictive information beyond commonly used spatial descriptors by comparing Base+ATLAS and Base+ATLAS+E, where RiverATLAS reach attributes are included in both configurations and SE are added only in the latter. For these two experiments, hyperparameters are tuned using station-based 5-fold cross-validation to report best-achievable performance under the LOSO evaluation setting.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

2.3 Evaluation metrics

To systematically evaluate DISE performance, we use six metrics: the correlation coefficient (CC, **Eq. (1)**), Kling–Gupta efficiency (KGE, **Eq. (2)**; Gupta et al., 2009), Nash–Sutcliffe efficiency (NSE, **Eq. (6)**; Nash and Sutcliffe, 1970), normalized root-mean-square error (nRMSE, **Eq. (8)**; Irving et al., 2018), percent bias (pBIAS, **Eq. (9)**), and relative variability (RV, **Eq. (10)**). Together, these metrics quantify temporal coherence (CC), overall efficiency relative to observed variability (NSE), and integrated agreement in correlation, variability, and mean bias (KGE). In addition, nRMSE measures the typical magnitude of residual errors normalized by the mean observed discharge, pBIAS diagnoses systematic bias in long-term flow magnitude, and RV evaluates bias in variability by comparing the relative dispersion of simulated and observed discharge. For visualization and comparison, we apply the transformations $1-\mathrm{nRMSE}$, $1-|\mathrm{pBIAS}|$, and $1-|\mathrm{RV}-1|$ so that all reported scores share a consistent interpretation, with larger values indicating better skill.

$$CC = \frac{\mathrm{cov}(Q_s, Q_o)}{\mathrm{std}(Q_s)\mathrm{std}(Q_o)} \tag{1}$$

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

$$r = CC \tag{3}$$

$$\alpha = RV \tag{4}$$

$$\beta = \frac{\mathrm{mean}(Q_s)}{\mathrm{mean}(Q_o)} \tag{5}$$

$$NSE = 1 - \frac{\sum(Q_o - Q_s)^2}{\sum(Q_o - \overline{Q_o})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum(Q_o - Q_s)^2}{N}} \tag{7}$$

$$nRMSE = \frac{RMSE}{\max(Q_o) - \min(Q_o)} \tag{8}$$

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

$$pBIAS = \left(\frac{\text{mean}(Q_s)}{\text{mean}(Q_o)} - 1\right) \times 100\% \tag{9}$$

$$RV = \frac{\text{std}(Q_s)}{\text{std}(Q_o)} \tag{10}$$

where Q_s represents the simulated streamflow and Q_o denotes the observed streamflow.

2.4 Study area

The study focuses on the Yangtze River Basin (YRB) in China (Fig. 2), the country's largest river basin, spanning diverse hydroclimatic and physiographic settings from the Tibetan Plateau to the East China Sea. Strong elevation gradients and pronounced spatial heterogeneity in precipitation and temperature produce highly variable flow regimes across the basin, ranging from snowmelt- and glacier-influenced headwaters to rainfall-dominated middle and lower reaches.

Beyond this natural heterogeneity, the basin is strongly shaped by human activity through both land-use modification and water management. Socioeconomic development is concentrated in the basin, which covers about 18.8% of China's land area but supports about 36% of its population and 40% of its GDP, reflecting intensive urbanization and land development (Zhu et al., 2020). Recent decades have also seen rapid expansion of urban land in the Yangtze River Delta, where urban built-up area increased from 4,855.30 to 44,447.15 km^2 over the past three decades (Yin et al., 2024a). Agriculture is equally prominent: irrigation is a major pressure and is reported to account for about 43.6% of total water use in the basin, largely supporting water-intensive rice production (Liu et al., 2021). Flow regimes are further modified by reservoir regulation and cascade hydropower development (Guo et al., 2021), which can alter seasonal variability and event responses. To evaluate model performance under these diverse and highly modified conditions, we compile daily discharge observations from 41 gauging stations distributed across major tributaries and the main stem, spanning a wide range of drainage areas and regulation intensities.

3 Results

3.1 Satellite-embedding aggregation and station-scale performance of DISE

We first examine how different subcatchment-level aggregation strategies for satellite embeddings affect DISE performance by comparing statistical summaries of SE, including the mean, standard deviation, skewness, and their combinations. To isolate the effect of aggregation, we train DISE using a fixed hyperparameter configuration for all strategies (see detailed hyperparameters in Table A1). Although these settings are not intended to deliver the best-achievable skill, they provide a controlled benchmark in which performance differences can be attributed primarily to the aggregation choice rather than to model tuning. Station-level performance across the 41 gauges, evaluated using six metrics (KGE, NSE, CC, 1−nRMSE, 1−|pBIAS|, and 1−|RV−1|; higher values indicate better skill), is summarized in Fig. 3. Fully tuned results are presented in the subsequent experiments.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

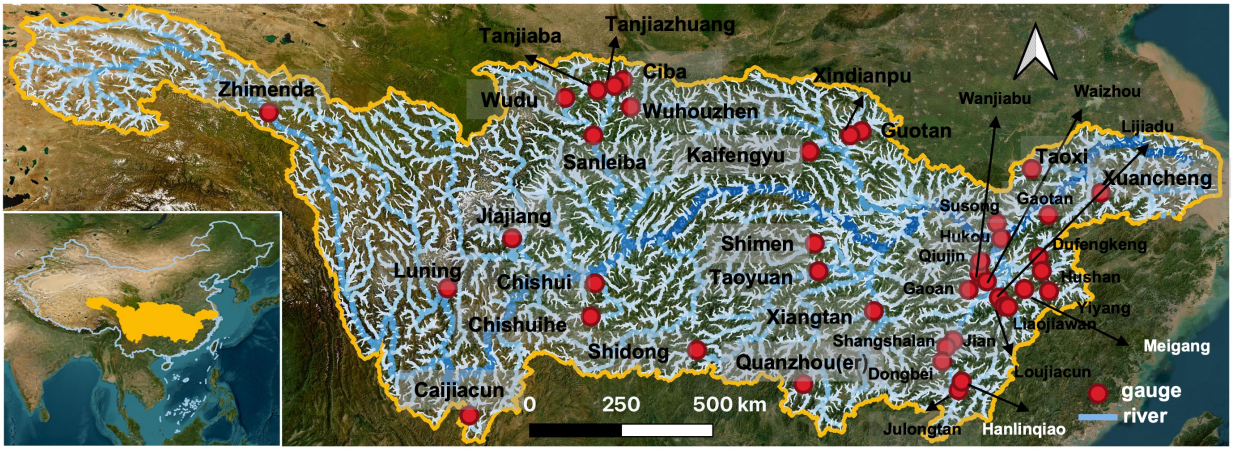

Figure 2. Research area Yangtze River Basin (YRB), China. The yellow outline delineates the Yangtze River Basin boundary, and the river network is shown in blue. Red dots indicate the 41 streamflow gauges used in this study. The names of each station are labeled nearby. The inset map shows the location of the YRB within China. Background World Imagery Map source credits: Esri, Maxer, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN and the GIS User Community | Powered by Esri.

Overall, incorporating satellite embeddings improves median performance relative to the GRADES-hydroDL baseline for most aggregation strategies, with the clearest gains appearing in NSE and the error-based metrics. For example, using mean aggregation increases the median KGE from 0.485 to 0.505 and the median NSE from 0.301 to 0.402, while also improving 1−nRMSE from 0.895 to 0.919, 1−|pBIAS| from 0.709 to 0.785, and 1−|RV−1| from 0.711 to 0.783. The strongest median gains in KGE and NSE are obtained by +Mean+Skew, which reaches 0.532 for KGE and 0.499 for NSE, whereas +Mean+Std gives the highest medians for CC (0.794) and 1−|RV−1| (0.794). However, these gains do not translate into a monotonic benefit of higher feature dimensionality. As additional summary moments are appended, performance generally becomes more variable across stations. The highest-dimensional setting (+All) illustrates this trade-off most clearly: its median KGE and NSE drop to 0.420 and 0.261, both below the GRADES-hydroDL baseline, and it shows the widest spread for several metrics.

Among all strategies, mean aggregation provides the best balance between skill, inter-station stability, and feature dimensionality. Although +Mean+Skew and +Mean+Std achieve slightly higher medians for some individual metrics, +Mean yields consistently tighter station-level distributions, with a KGE interquartile range of 0.322 and an NSE interquartile range of 0.489, compared with 0.438–0.447 and 0.493–0.537 for the other mean-based combinations, and 0.602 and 0.986 for GRADES-hydroDL. This indicates that mean aggregation delivers more reliable improvements across stations while keeping the feature representation compact. We therefore adopt mean aggregation in all subsequent analyses.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

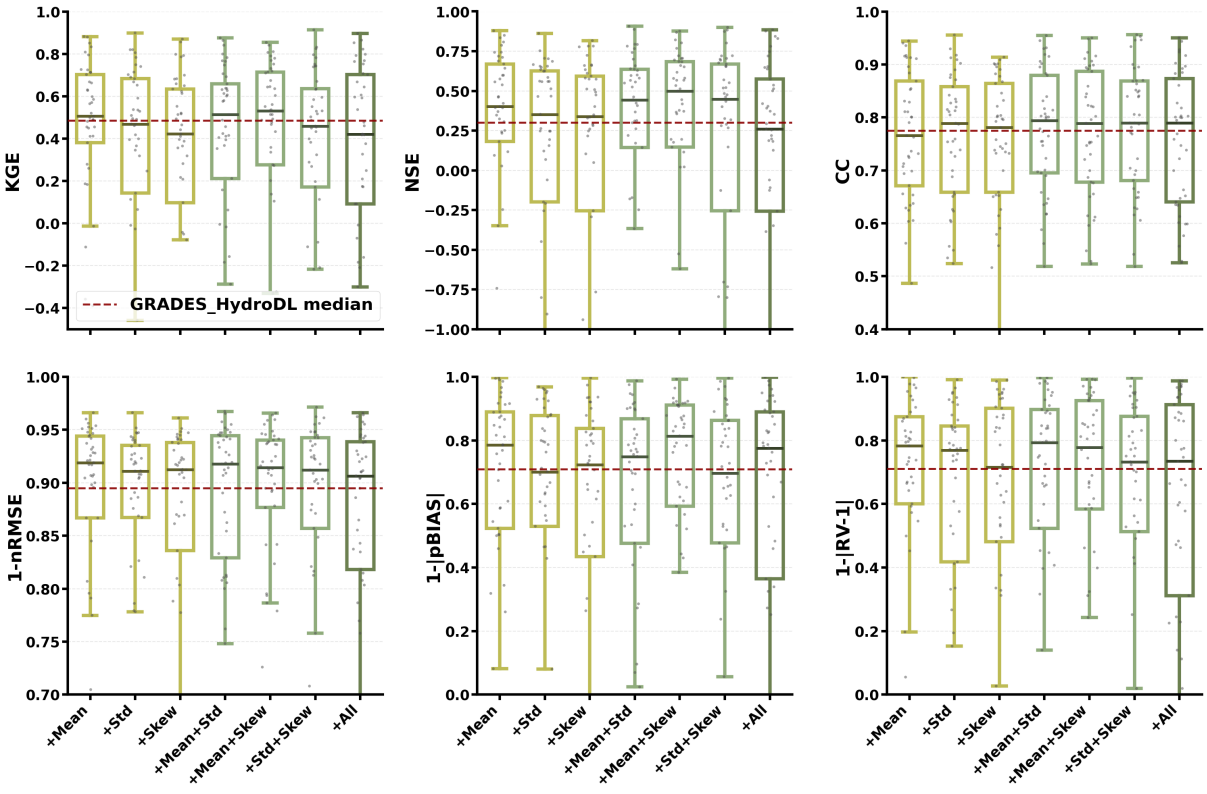

Figure 3. Performance of DISE under different satellite-embedding aggregation strategies. Each panel presents boxplots of model performance across the 41 stations for one metric (KGE, NSE, CC, 1-nRMSE, 1-|pBIAS|, and1- |RV-1|). DISE is trained with a fixed hyperparameter setting, while varying the embedding feature inputs (mean, standard deviation, skewness, and their combinations, including all statistics). Color intensity (deeper green) denotes higher feature-set dimensionality. The red dashed line indicates the median performance of the GRADES-hydroDL baseline for the corresponding metric.

After selecting mean aggregation as most reliable SE aggregation strategy, we next tune the XGBoost hyperparameters to quantify the best achievable skill of DISE and summarize the spatial distribution of station-wise performance in Fig. 4. Because DISE is trained as a residual correction to a strong first-guess simulation, its attainable improvement is expected to depend on the baseline skill. We therefore first examine the spatial pattern of GRADES-hydroDL performance before interpreting where DISE yields the largest gains. GRADES-hydroDL exhibits pronounced spatial heterogeneity (Fig. 4a). Skill is generally higher (yellower in Fig. 4a) in the Poyang Lake Basin, whereas performance degrades (bluer in Fig. 4a) in parts of the upper YRB, including tributaries such as the Jialing River. Across metrics, CC and $1-$nRMSE are relatively strong and spatially coherent, while KGE and NSE are lower and more variable.

Incorporating SE improves performance at many stations (Fig. 4b), and the difference maps highlight where these gains occur (Fig. 4c). Across the network, the most consistent improvements appear in efficiency- and error-structure metrics, with widespread positive shifts in ΔKGE and ΔNSE, accompanied by gains in $\Delta(1-|pBIAS|)$ and $\Delta(1-|RV-1|)$. In contrast,

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

ΔCC is generally small, indicating that DISE primarily improves discharge magnitude, bias, and variability rather than timing. Spatially, the strongest gains cluster in the upper basin and in the Poyang Lake Basin. Improvements in the upper basin coincide with regions where GRADES-hydroDL shows lower initial skill, suggesting greater scope for residual correction. In the Poyang Lake Basin, where land use and land cover have been strongly modified, the pronounced gains are consistent with SE capturing spatially heterogeneous land-surface signals that are not well represented in the baseline simulation. A small number of stations still show degraded performance, notably Zhimenda in the headwaters and Hukou near the Poyang Lake outlet. These sites are influenced by processes such as high-elevation cryospheric contributions and complex lake-river interactions, which are less likely to be fully captured by surface imagery alone and may therefore limit the effectiveness of DISE.

To further demonstrate how satellite embeddings translate into hydrograph-level changes, Fig. 5 presents representative daily hydrographs for eight gauges. The stations are organized into four groups according to the KGE change of DISE relative to GRADES-hydroDL, representing high, moderate, and low improvement, as well as no improvement. Across the improvement groups, DISE generally enhances performance by adjusting flow magnitudes during low-flow periods and refining selected event peaks, which is consistent with the stronger gains in KGE, NSE, and the error-based metrics than in CC.

For the high-improvement gauges (Chisui and Tanjiazhuang), DISE substantially reduces the excessive baseline discharge during low-flow periods and dampens several overestimated medium-to-high peaks, producing more realistic peak magnitudes and improved low-flow behavior. These large hydrograph corrections are consistent with the marked increases in both KGE and NSE. For the moderate-improvement gauges (Jian and Shangshalan), where the baseline already captures the timing of variability reasonably well, DISE makes more targeted refinements, including better low-flow levels and sharper, better-aligned peaks for selected events. For the low-improvement gauges (Shidong and Gaoan), the corrections are smaller and more localized, with only modest reductions in low-flow bias or event-peak errors, in line with the limited performance gains in this group.

The no-improvement gauges reveal where DISE is less effective. At Zhimenda, DISE strongly suppresses the broad summer high-flow season and substantially underestimates the observed seasonal rise, suggesting that the residual correction over-damps the hydrograph where summer discharge is partly sustained by upstream high-mountain snowmelt. These cryosphere-related contributions originate upstream and are not directly encoded in local land-surface conditions, making them difficult to capture with the predictors used here. At Hukou, DISE sharpens the main flood peak but underestimates the elevated flows before and after the event, leading to only limited correction of the overall seasonal hydrograph. This behavior is consistent with the influence of complex lake-river interactions near the Poyang Lake outlet, where storage, backwater, and exchange processes may be difficult to recover mainly through land-surface information.

3.2 Control Experiments

To further isolate the contribution of SE beyond meteorological forcings, we conducted a controlled comparison between Base and Base+E configurations (Fig. 6(a), Fig 7). Results are summarized using a radar plot of station-wise metrics and boxplots across the 41 gauges.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

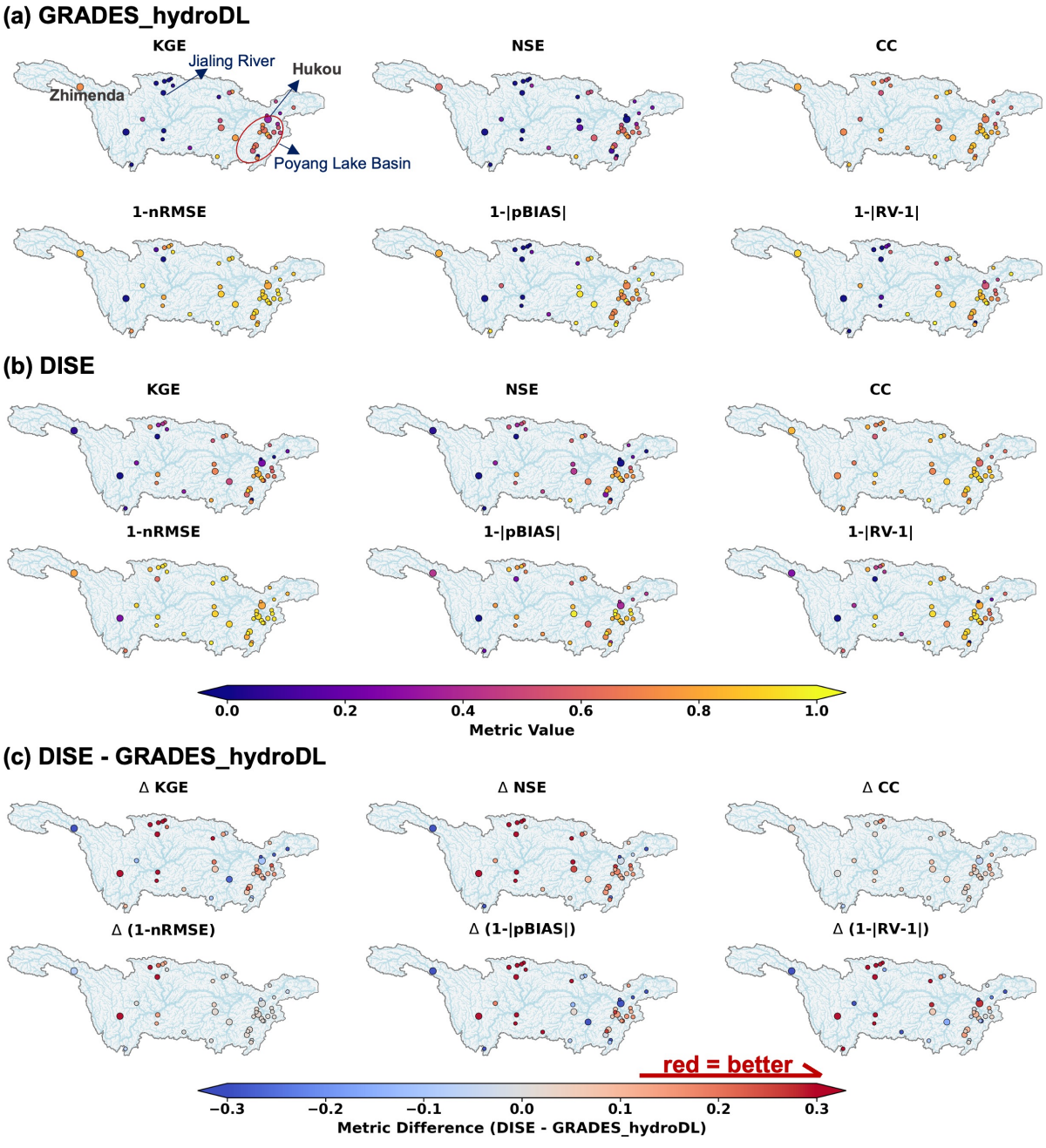

Figure 4. Spatial patterns of model performance across the Yangtze River Basin. (a) GRADES-hydroDL and (b) DISE performance at the 41 gauging stations, evaluated using KGE, NSE, CC, 1-nRMSE, 1-|pBIAS|, and 1-|RV-1|. (c) Performance differences between DISE and GRADES-hydroDL for each metric. Circles indicate station locations, with symbol size scaled by upstream drainage area; colors denote metric values (a–b) and metric differences (c), as shown by the color bars.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

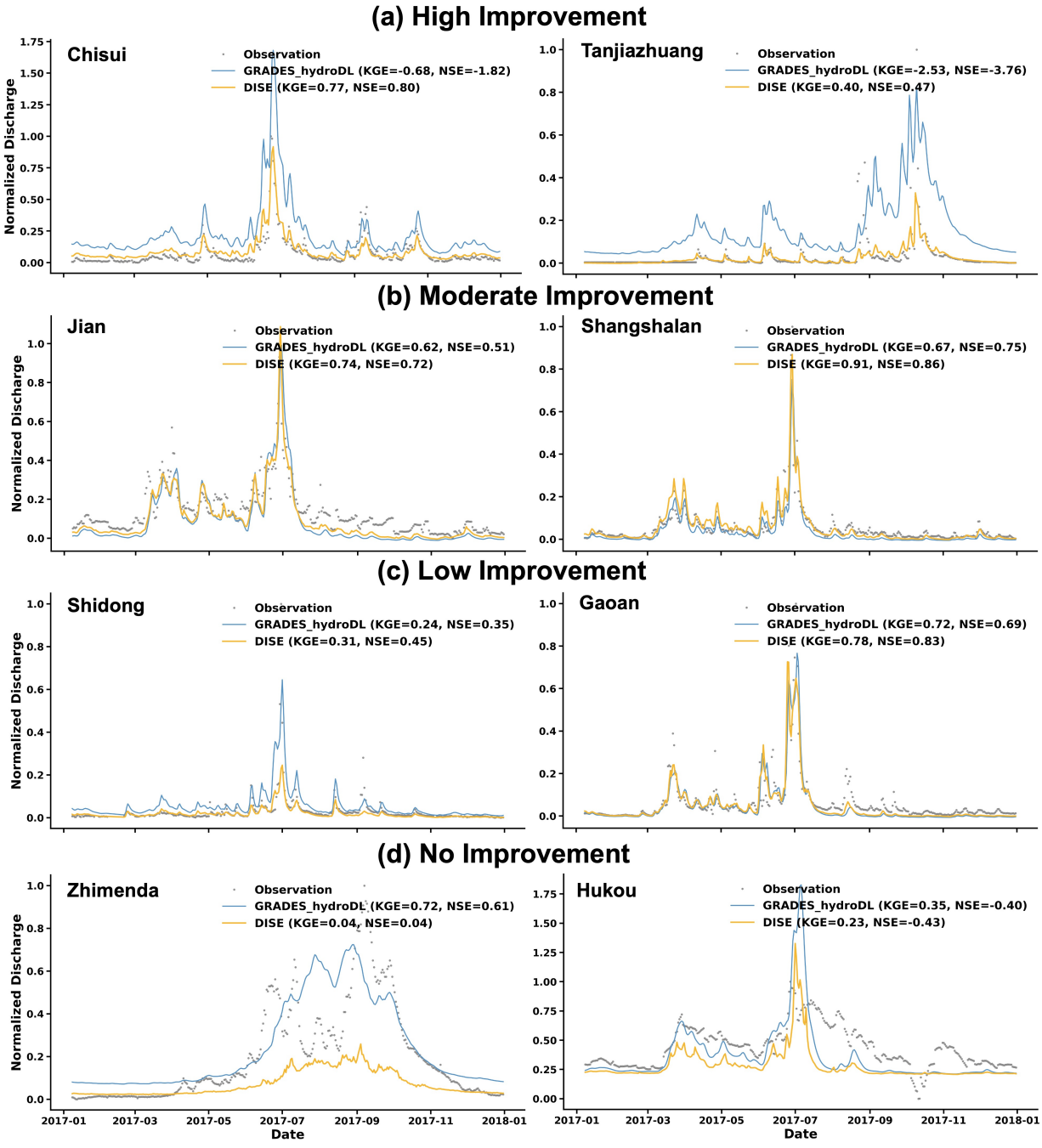

Figure 5. Daily discharge hydrographs for eight representative gauges in the Yangtze River Basin. Observed discharge is shown as grey points, together with simulations from GRADES-HydroDL (blue) and the Data Integration model with Satellite Embeddings (DISE, yellow). Six stations are grouped into three 33% bins according to the KGE gain of DISE relative to GRADES-HydroDL (high, moderate, and low improvement), with two stations randomly selected from each bin, and two additional stations are shown to represent degraded performance. For each station, KGE and NSE values for both models are reported in the legend.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

We find that Base configuration yields performance comparable to GRADES-hydroDL, with only marginal changes across metrics (blue line and grey dash line in (Fig. 6a)), indicating that meteorological forcings and lagged baseline discharge alone provide limited additional benefit in spatial generalization task. In contrast, adding SE lead to clear and consistent gains. The spatial feature space also shows distinct distribution. Compared to the inputs of base experiment, adding embeddings yields a more station-separable organization in the UMAP projections, with less inter-station mixing of daily samples (Fig. A1). Median skill increases from 0.473 to 0.594 for KGE and from 0.317 to 0.533 for NSE, accompanied by improvements in 1-|pBIAS| (0.725 to 0.792) and 1-|RV-1| (0.699 to 0.765), while CC and 1-nRMSE show smaller but positive changes (blue and yellow boxes in Fig. 7). These patterns indicate that SE contributes complementary, spatially varying information that helps correct station-dependent errors in flow magnitude, bias, and variability that are not fully captured by meteorological forcings and baseline simulations alone. Notably, Base+E also exhibits a smaller interquartile range than Base for several metrics (blue and yellow boxes in Fig. 7), suggesting more stable generalization across stations. This reduced spread implies that SE also decrease the number of poorly performing sites.

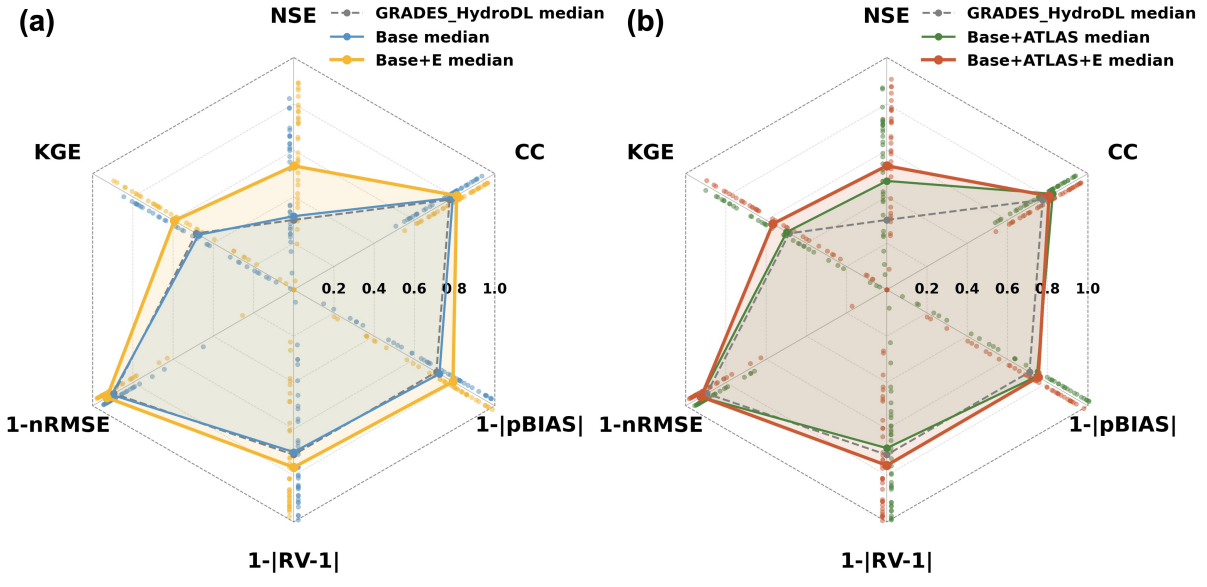

Figure 6. Performance of control experiments across 41 gauging stations. (a) Radar plot summarizing station-wise performance for six metrics (KGE, NSE, CC, 1-nRMSE, 1-|pBIAS|, and 1-|RV-1|). Blue and yellow dots show metric values at individual stations (n = 41) for Base and Base+E experiments, respectively; the connected markers indicate the median performance across stations for each model. The grey dashed line denotes the median performance of GRADES-hydroDL. (b) Green and red dots show metric values at each stations for Base + ATLAS and Base+ATLAS+E experiment, respectively; the connected markers indicate the median performance across stations for each model.

To test whether SE provides predictive information beyond commonly used hyro-environmental reach descriptors, we conduct a second control experiment comparing Base+ATLAS and Base+ATLAS+E configuration(Fig. 6b, Fig. 7).

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Incorporating traditional spatial descriptors improves generalization skill to ungauged locations. Compared with the model configurations without spatial context (blue line in Fig. 6a), adding RiverATLAS attributes yields higher median performance across several metrics (green line in Fig. 6b), indicating that even imperfect reach-level descriptors with oudated and inconsistent information can provide useful spatial differentiation for station transfer. However, adding SE on top of RiverATLAS leads to further gains (red line in Fig. 6b), most clearly for KGE (median 0.497 to 0.567), NSE (0.468 to 0.534), and 1-|RV-1| (0.682 to 0.757) (red boxes in Fig. 7). These improvements suggest that SE capture additional, spatially heterogeneous signals not fully represented by traditional hydro-environmental attributes.

Comparing the yellow boxes with the red boxes in Fig. 7 further suggests that the marginal benefit of RiverATLAS diminishes once SE are included. When SE are already present, adding RiverATLAS provides little additional skill and even slightly lower the median for the metrics except NSE. This pattern may be due to partial redundancy between the two sources of land surface information and to additional noise or conflicting signals introduced by RiverATLAS under station transfer.

These control experiments highlight the importance of land surface information for spatial generalization in streamflow modeling across river networks, and indicate that SE, as analysis-ready representations of land-surface conditions, are a promising source of spatial context for streamflow reconstruction.

3.3 Role of forcings, satellite embeddings, and regulation in controlling reconstruction gains

Streamflow errors in human-disturbed basins can arise from imperfect meteorological forcing, spatially heterogeneous land-surface conditions (e.g., irrigation and urbanization), and regulation effects, where upstream storage and release operations shape downstream flows. Because SE primarily encode land-surface context, they are expected to help most when baseline errors are tied to local landscape heterogeneity, but they may have limited ability to directly represent operation-driven signals. To examine these relative contributions, we design a two-axis attribution analysis that contrasts the gains from meteorological forcings and SE across stations (Fig. 8).

We first use the two-axis scatter plot to assess whether SE and meteorological forcings contribute complementary gains across stations, that is, whether cases with strong Base performance tend to exhibit smaller incremental improvements from adding SE. We then incorporate the regulation level as marker color and summarize the quadrant outcomes with a pie inset to facilitate interpretation across metrics, since station-level patterns in the scatter can be difficult to synthesize. In Fig. 8, the x-axis shows the gain from meteorological forcings (Base−GRADES-hydroDL), and the y-axis shows the incremental gain from SE (Base+E−Base). Point color denotes the degree of regulation (DOR), highlighting stations where operation-driven effects may constrain improvements. The four quadrants define different types: *F only* ((x>0,y<0)), *E only* ((x<0,y>0)), *F+E* ((x>0,y>0)), and *Neither* ((x<0,y<0)). Representative stations and their SE visualizations are shown in Fig. A2.

Across all six metrics, F+E dominates (14–23 of 41 stations), and E only forms the second-largest share (8–12 stations), whereas F only appears at fewer gauges (4–10 stations). This pattern indicates that forcings alone are often insufficient to fully correct baseline errors, and that SE provide complementary value across stations and metrics. The Neither category is generally small (2–5 stations for KGE, NSE, CC, and 1-nRMSE) but becomes more evident for error-structure metrics such as 1-|pBIAS|

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

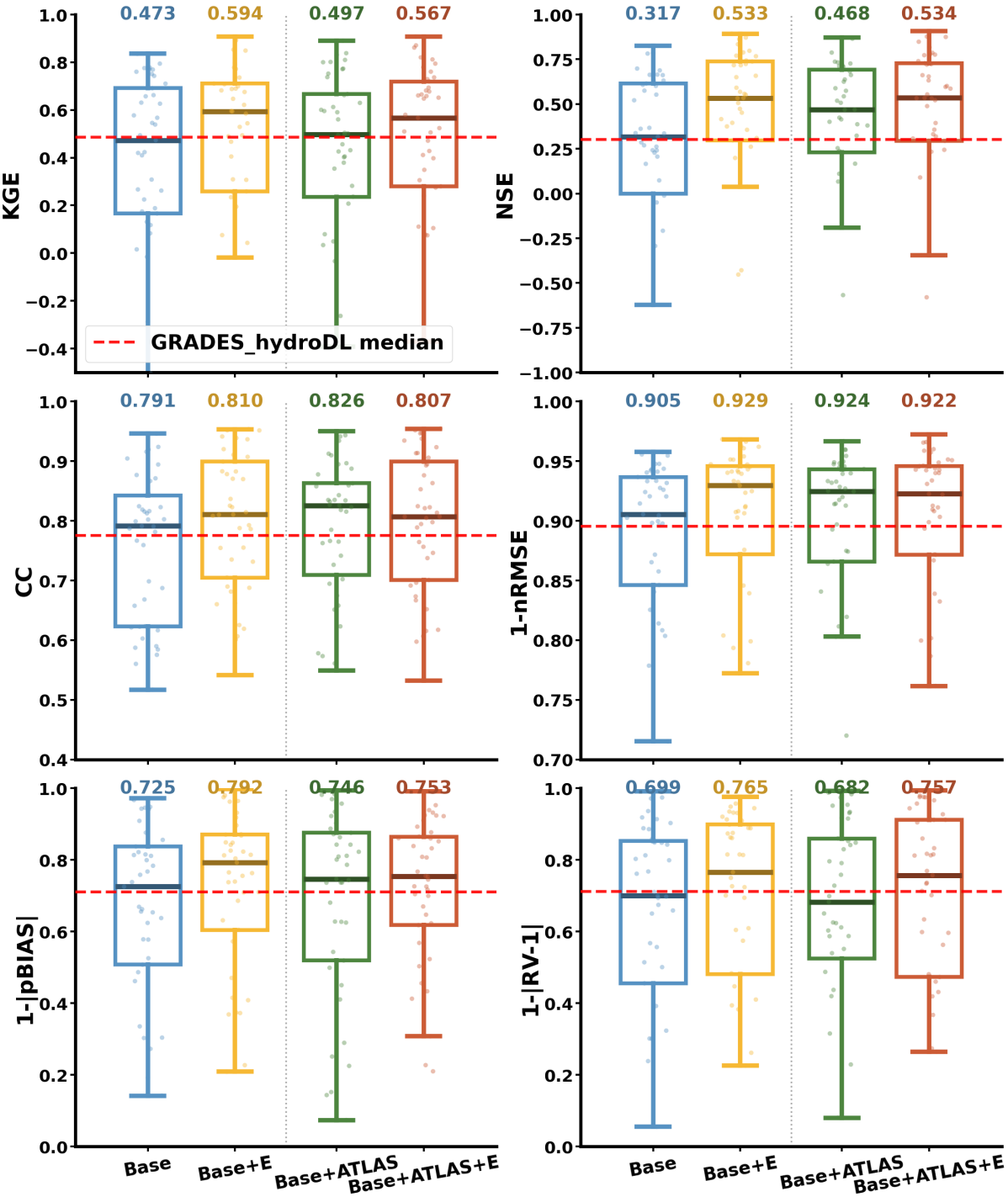

Figure 7. Boxplots of control experiments. Boxplots of six metrics across stations for Base (blue), Base+E (yellow), Base+ATLAS (green), and Base+ATLAS+E (red). Colored points denote individual stations, and the numeric labels at the top indicate the median for each configuration; the red dashed line marks the GRADES-hydroDL median.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

(7 stations), highlighting a subset of locations where biases remain difficult to correct even after incorporating satellite-derived context.

Regulation level is associated with differences in the dominant improvement pathway across stations. The *F+E* group is largely composed of weakly regulated gauges (low DOR, blue markers), where discharge remains closely coupled to meteorological variability and local land-surface conditions, allowing both forcings and embedding-derived spatial context to contribute. In contrast, *E only* stations show a higher prevalence of stronger regulation (higher DOR, warmer colors), particularly for ΔKGE, ΔNSE, $\Delta(1-nRMSE)$, and $\Delta(1-|pBIAS|)$. This pattern is consistent with the interpretation that when regulation weakens the direct forcing–runoff linkage, forcing-based corrections alone become less effective, whereas SE may still provide useful constraints by encoding landscape signatures associated with managed and human-modified catchments. For example, highly urbanized or intensively irrigated subcatchments may co-occur with upstream infrastructure that influences flow behavior. We further train linear probes to test whether SE contain clear land-surface information (Lees et al., 2022), and find that they can predict subcatchment forest, cropland, and urban fractions with clear skill, particularly for urban cover (Fig. A3). By contrast, this relationship largely disappears after shuffling the station-to-embedding correspondence, indicating that the signal is genuinely encoded in SE rather than arising from chance. The *Neither* cases are comparatively rare but tend to coincide with stronger regulation, consistent with situations where operation-driven departures from natural dynamics are not well captured by either meteorological inputs or local land-surface context alone.

4 Discussion

DISE highlights the practical value of incorporating satellite-derived spatial context for reach-scale streamflow reconstruction. By summarizing local land-surface conditions within each reach's subcatchment, satellite embeddings enhance station-transfer skill across many gauges, reinforcing that land surface information is essential for representing heterogeneous land-surface patterns in streamflow simulation across river networks. In strongly regulated reaches, however, streamflow is shaped not only by meteorological forcing and local land-surface context but also by upstream storage and release decisions that can propagate downstream. This suggests a natural extension of DISE: fusing additional river-state indicators that more directly reflect regulation signals, such as remotely sensed river width and water surface elevation, which can provide complementary constraints on managed flow behavior.

While DISE demonstrates clear skill gains, its performance is bounded by several structural constraints that operate at different levels. At the model level, DISE functions as a residual correction to GRADES-hydroDL, so its effectiveness depends on baseline errors being systematic and learnable from the available predictors. Errors that are random or driven by processes not represented in the inputs cannot be corrected within this framework. Because DISE adjusts discharge magnitude in log space rather than re-learning routing dynamics, substantial timing or structural errors in the baseline simulation also cannot be fully resolved. Furthermore, effective spatial transfer requires that baseline biases exhibit sufficient cross-station consistency to be inferred from training locations and generalized to withheld sites.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

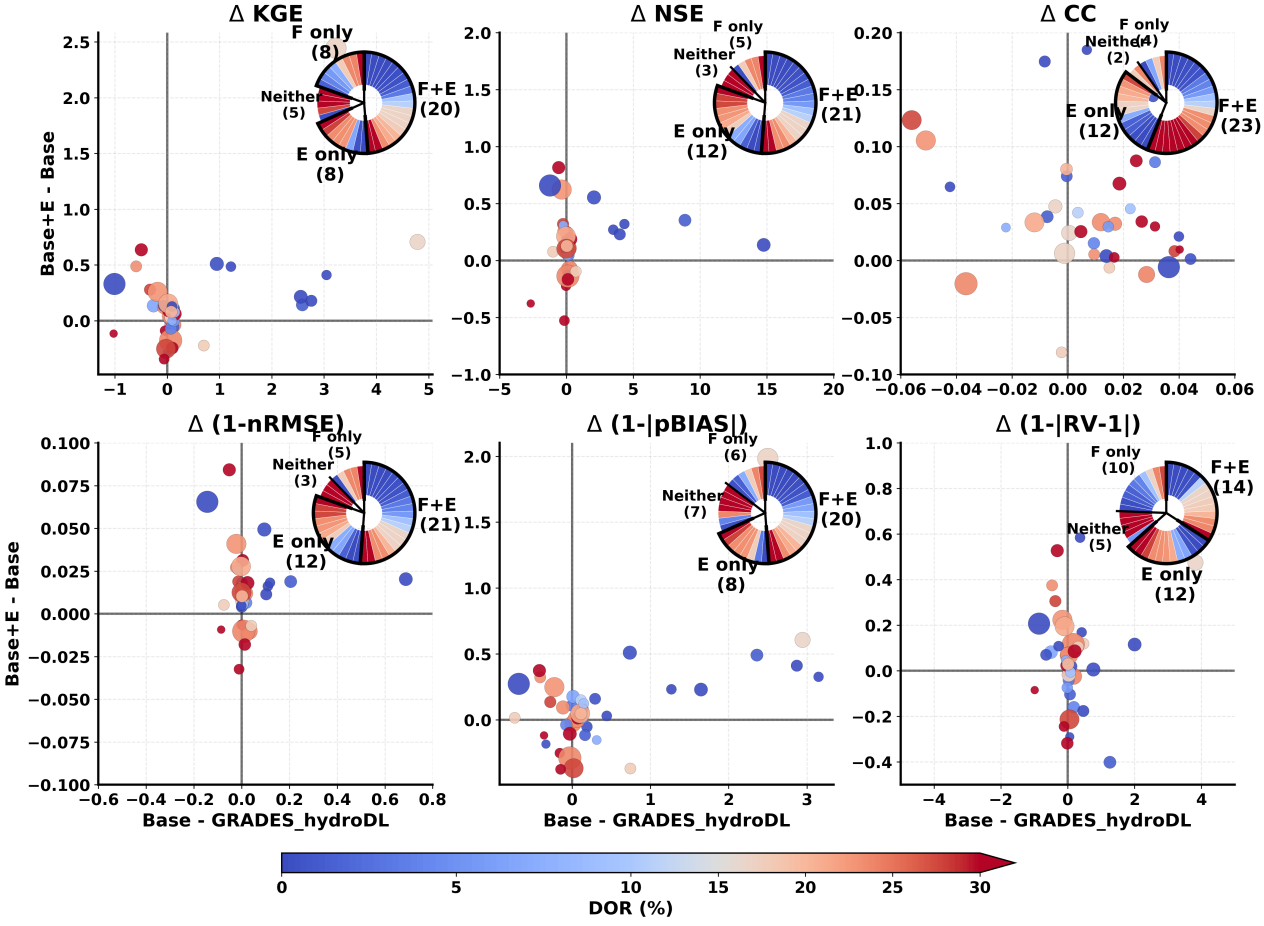

Figure 8. Station-level attribution of performance gains from meteorological forcings and satellite embeddings across six metrics (KGE, NSE, CC, 1-nRMSE, 1-|pBIAS|, and 1-|RV-1|). Each panel plots the improvement of the Base model relative to GRADES-hydroDL against the additional improvement from adding satellite embeddings (Base+E relative to Base). Points represent individual gauging stations (n = 41), with marker size scaled by upstream drainage area and colors indicating the degree of regulation (DOR). Pie insets summarize the number of stations where gains are driven by forcings only (F only), embeddings only (E only), both (F+E), or neither.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

The magnitude of DISE improvements is also inherently tied to the strength of the baseline model itself. In regions where the baseline is weaker, larger apparent gains may be expected. For example, almost no gauge observations or locally optimized forcing data from China were included in the training of GRADES-hydroDL, so baseline performance there is not expected to be strong. The larger improvements observed in China may therefore reflect this limitation, together with the benefits of incorporating land-surface context, locally optimized forcings, and streamflow observations in DISE. Further work is needed to evaluate DISE effectiveness in other regions and to better quantify how improvements relate to baseline model performance.

At the input level, a further constraint arises from the temporal resolution of the satellite embeddings. The Google Satellite Embedding V1 product provides annual representations of land-surface conditions that are treated as static within each year. While this captures spatial heterogeneity, it does not resolve intra-annual dynamics such as crop phenology, irrigation timing, or short-term disturbances. DISE therefore relies on daily meteorological forcings and the baseline simulation to represent sub-annual variability, with embeddings primarily informing spatial differentiation. Incorporating temporally richer satellite representations could further improve reconstruction skill, particularly in regions where seasonal land management strongly influences runoff generation.

Because embeddings primarily provide land surface information, the evaluation is framed around spatial generalization using a leave-one-station-out setting within a single basin and year, such that stations share partially correlated large-scale meteorological forcing. This design isolates spatial differentiation under broadly shared hydroclimatic conditions and enables a controlled assessment of the incremental value of satellite embeddings, but it does not constitute a full spatio-temporal generalization test. In addition, owing to the limited accessibility of daily streamflow observations in China — especially for recent years, the present analysis is restricted to 2017, when sufficiently complete gauge records were available. Future work should therefore examine performance under stronger domain shifts, including cross-year and cross-basin transfer, to provide a more complete picture of embedding portability across hydroclimatic and disturbance regimes.

Overall, DISE aligns with emerging context-aware strategies for hydrologic modeling across river networks. The observed skill gains and the controlled comparisons indicate that integrating spatial context is a promising pathway for improving streamflow reconstruction across river networks.

5 Conclusions

This study presents DISE, a data-integration framework that fuses satellite embeddings and a recently developed discharge product (GRADES-hydroDL) to reconstruct streamflow in human-disturbed basins. By learning corrections toward gauge observations in log space and evaluating performance under leave-one-station-out cross-validation across 41 stations, we draw three main conclusions.

1. **DISE improves streamflow reconstruction over GRADES-hydroDL.** Gains are strongest for magnitude and bias, while correlation changes are modest. Median KGE increases from 0.485 to 0.594 and median NSE from 0.301 to 0.533. Spatially, the largest improvements occur in the upper basin and in the Poyang Lake Basin.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

2. **Satellite embeddings provide complementary landscape context that improves spatial generalization.** Beyond meteorological forcings and traditional hydro-environmental reach attributes (e.g., RiverATLAS), satellite embeddings contribute additional land-surface information that improves reconstruction skill across river networks. Once SE are included, RiverATLAS adds little further skill and can even slightly reduce performance.

3. **Strong regulation limits DISE gains, but satellite embeddings still provide useful constraints where forcing-based corrections become less effective.** Improvements are generally smaller at highly regulated stations, suggesting that operation-driven signals are not fully captured within the current framework. Nevertheless, satellite embeddings can still improve reconstruction under these conditions.

Our results support the broader perspective that incorporating spatial context through embedding-based representations can improve spatial generalization and enhance the streamflow simulation skills across river networks, particularly in highly land-surface-modified basins where process-based models and data-driven approaches (e.g., LSTM) may have systematic limitations. Future work should further assess transferability across broader hydroclimatic regimes, explore richer temporal representations of satellite information, and investigate synergies with emerging entity-aware and context-aware learning strategies.

Appendix A

To determine the hyperparameters used in each experiment (Table A1), we applied Optuna combined with grouped cross-validation. Groups were first defined by station identity (`stnm|stcd`) to ensure that samples from the same station were never mixed between the training and validation subsets within a fold. For each trial, Optuna sampled a candidate hyperparameter set and evaluated it using `GroupKFold` cross-validation with five folds. Within each fold, predictors were standardized, sample weights were computed according to the selected weighting strategy, and an XGBoost model was trained on the training subset. Performance was then evaluated on the validation subset using RMSE, and the mean RMSE across folds was used as the objective value for that trial. Optuna employed the `TPESampler` to minimize this grouped cross-validated loss and return the optimal parameter set.

The optimized hyperparameters included `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `reg_alpha`, `reg_lambda`, `min_child_weight`, `gamma`, and `n_estimators`. These parameters jointly control model capacity and regularization. Specifically, `max_depth`, `learning_rate`, and `n_estimators` control tree complexity, update size, and the overall number of boosting rounds. The sampling parameters `subsample` and `colsample_bytree` improve generalization by reducing variance, while `reg_alpha`, `reg_lambda`, `min_child_weight`, and `gamma` provide additional regularization to limit overfitting.

In the linear-probe analysis, we train a separate Elastic Net model for each land-surface indicator using nested five-fold cross-validation. The outer loop is used to obtain out-of-fold predictions for unbiased performance evaluation, whereas the inner loop is used to tune the regularization parameters. Final predictions are obtained by concatenating the held-out predictions across all outer folds. As a shuffled baseline, we use the same features and cross-validation design but randomly permute the

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

training targets within each outer fold before fitting, which destroys the correspondence between embeddings and the target while keeping the predictors unchanged. This baseline provides a reference for assessing whether predictive skill arises from information genuinely encoded in the embeddings rather than chance associations. Elastic Net is particularly suitable here because the embedding dimensions may be intercorrelated, and the combined L1-L2 regularization both stabilizes estimation and suppresses uninformative predictors.

The purpose of this linear-probe analysis is not to maximize predictive accuracy, but to test whether these land-cover fractions can be directly recovered from satellite embeddings through a simple linear readout. The consistent improvement over the shuffled baseline across all fractions indicates that satellite embeddings retain genuine, non-random correspondence with land-surface characteristics rather than spurious station matching. At the same time, the weaker performance for some fractions suggests not the absence of relevant information, but that these signals are less directly organized in a linearly separable form and may require more flexible nonlinear mappings for fuller recovery.

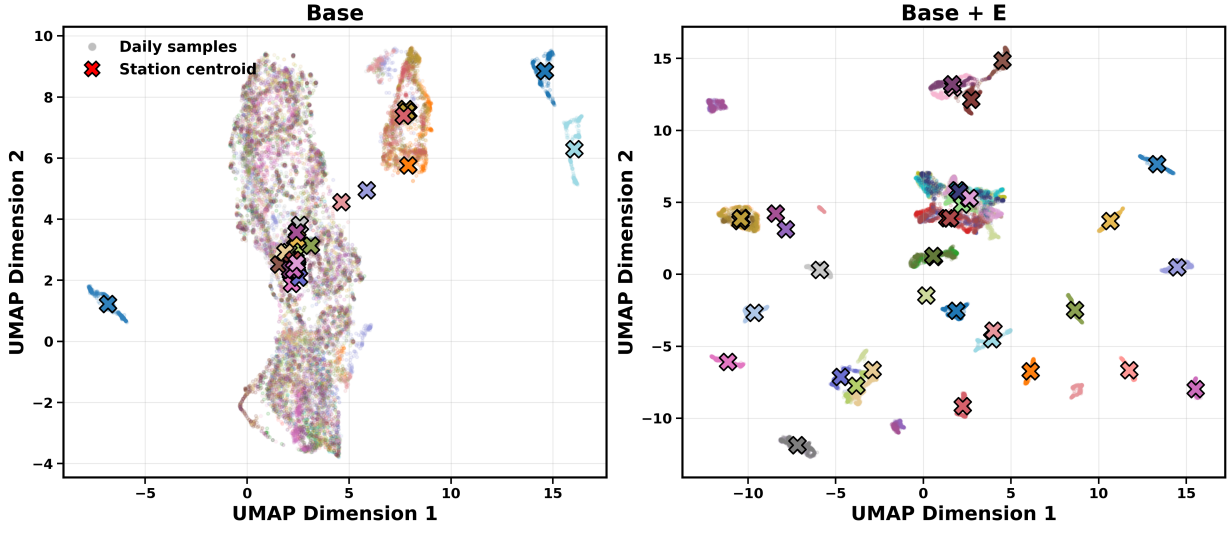

Figure A1. UMAP visualization of the input feature space of Base and Base+E configuration. Each colored dot represents a daily sample in the UMAP-projected feature space, and colored crosses denote station centroids computed as the mean of embedded daily samples for each station.

Table A1. XGBoost hyperparameters used in each experiment.

Experiment	max_depth	learning_rate	subsample	colsample_bytree	reg_alpha	reg_lambda	min_child_weight	gamma	n_estimators
Strategy (Fixed)	6	0.30000	0.80000	0.80000	0.00000	1.00000	1	0.00000	100
Base (Tuned)	4	0.01252	0.82106	0.65405	0.24408	0.97803	4	0.90932	1369
Base+E (Tuned)	6	0.01875	0.69820	0.77356	1.42439	0.56752	3	0.43813	1831
Base+ATLAS (Tuned)	4	0.06335	0.81293	0.65577	0.26054	0.72524	9	0.64216	2825
Base+ATLAS+E (Tuned)	6	0.05762	0.88797	0.57398	1.98176	0.14089	5	0.73173	1689

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

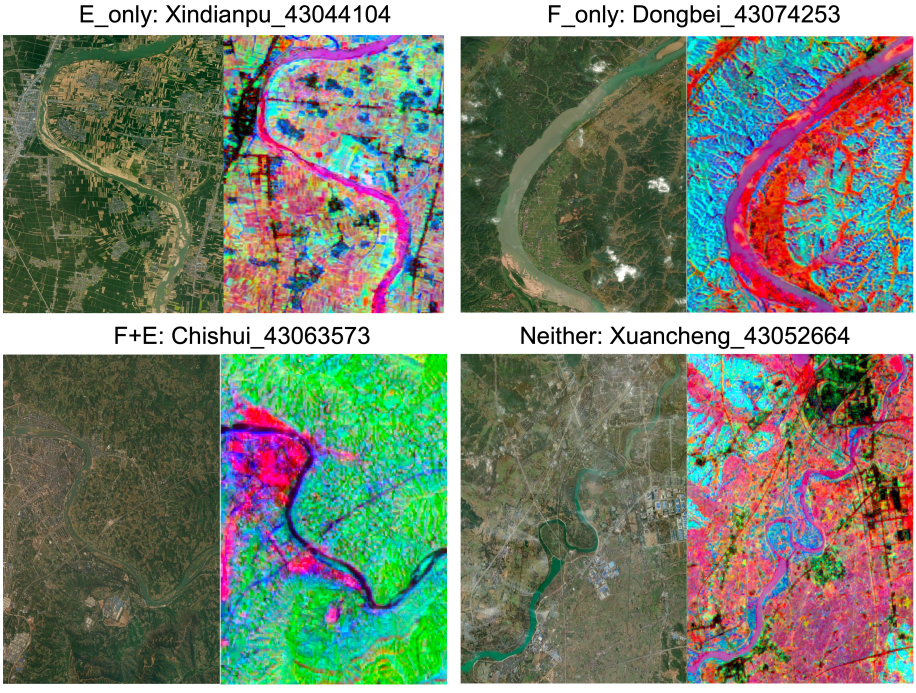

Figure A2. Visual examples of satellite embeddings for four categories. One station is randomly selected from each category (E only, F only, F+E, and Neither). For each station, the left panel shows a true-color satellite image of the surrounding reach area, and the right panel visualizes the satellite embeddings using a three-band composite formed by the three most important embedding dimensions for streamflow reconstruction at that station, as ranked by the station-specific model feature importance. Background World Imagery Map source credits: Esri, Maxer, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN and the GIS User Community | Powered by Esri.

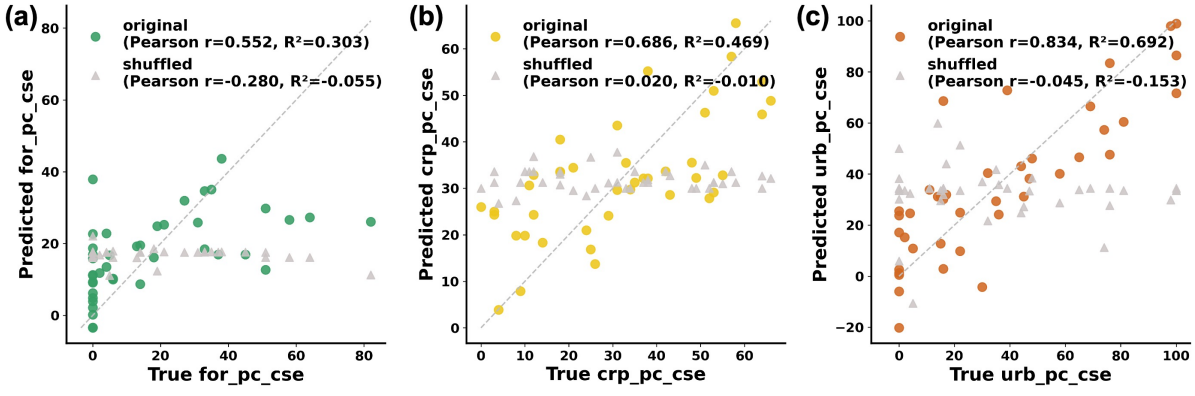

Figure A3. Linear-probe evidence that satellite embeddings encode key land-surface information. Panels (a)–(c) compare the true subcatchment fractions of forest, cropland, and urban land cover from RiverATLAS (x-axis) against the corresponding fractions predicted from satellite embeddings by linear probes (y-axis). Colored circles show results using the original station-to-embedding correspondence, whereas gray triangles show a shuffled control in which the correspondence between stations and embeddings is randomly permuted before training. Pearson correlation coefficients and R^2 values are reported in each panel. The dashed line denotes the 1:1 line.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Code and data availability. All data used in this study are publicly available from the following repositories: Google Satellite Embeddings (https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL); China Meteorological Forcing Dataset v2 (CMFD v2; https://www.tpdc.ac.cn/zh-hans/data/e60dfd96-5fd8-493f-beae-e8e5d24dece4); GRADES-hydroDL (https://www.reachhydro.org/home/records/grades-hydrodl); Global Streamflow Characteristics, Hydrometeorology, and Catchment Attributes (GSHA; https://zenodo.org/records/10433905); RiverATLAS (http://www.hydrosheds.org/page/hydroatlas); and GeoDAR reservoir location and capacity data (https://zenodo.org/records/6163413). The code used in this study is available at https://github.com/LePapilllon/residual_learning.

Author contributions. Conceptualization: PL, HL. Investigation: HL, PL. Data curation: HL, FZ. Funding acquisition: PL. Methodology: HL, PL. Visualization: HL. Writing (initial draft): HL, PL. Writing (review and editing): HL, PL, LS, YY, MP, FZ, QQ, AH.

Competing interests. The authors declare no conflict of interests.

Acknowledgements. This research was supported by the National Natural Science Foundation of China (42371481), the Beijing Nova Program (20230484302), and the Beijing Nova Interdisciplinary Program (20240484647), Beijing Key Laboratory of Spatio-temporal Perception and Urban Resilience (Peking University). The authors used AI-assisted language tools to support manuscript polishing, including grammar refinement and sentence restructuring; all scientific content was reviewed and verified by the authors.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

References

Best, J.: Anthropogenic stresses on the world's big rivers, Nat. Geosci., 12, 7–21, https://doi.org/10.1038/s41561-018-0262-x, 2019.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., and Kohli, P.: AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, https://arxiv.org/abs/2507.22291, 2025.

Casu, F., Manunta, M., Agram, P. S., and Crippen, R. E.: Big remotely sensed data: tools, applications and experiences, Remote Sens. Environ., 202, 1–2, https://doi.org/10.1016/j.rse.2017.09.013, 2017.

Friedman, J. H.: Stochastic gradient boosting, Comput. Stat. Data Anal., 38, 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2, 2002.

Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., Babu, S., Borrelli, P., Cheng, L., Crochetiere, H., Ehalt Macedo, H., Filgueiras, R., Goichot, M., Higgins, J., Hogan, Z., Lip, B., McClain, M. E., Meng, J., Mulligan, M., Nilsson, C., Olden, J. D., Opperman, J. J., Petry, P., Reidy Liermann, C., Sáenz, L., Salinas-Rodríguez, S., Schelle, P., Schmitt, R. J. P., Snider, J., Tan, F., Tockner, K., Valdujo, P. H., van Soesbergen, A., and Zarfl, C.: Mapping the world's free-flowing rivers, Nature, 569, 215–221, https://doi.org/10.1038/s41586-019-1111-9, 2019.

Gudmundsson, L., Brunner, M. I., Döll, P., Fluet-Chouinard, E., Frolova, N., Gosling, S. N., Hirabayashi, Y., Kireeva, M. B., Liu, X., Müller Schmied, H., Magritskiy, D., Slater, L. J., Stein, L., Tramblay, Y., Wang, K., Wasko, C., Yamazaki, D., and Zhou, X.: Past and future change in global river flows, Nat. Rev. Earth Environ., 7, 7–23, https://doi.org/10.1038/s43017-025-00745-z, 2026.

Guo, S., Xiong, L., Zha, X., Zeng, L., and Cheng, L.: Impacts of the Three Gorges Dam on the streamflow fluctuations in the downstream region, J. Hydrol., 598, 126 480, https://doi.org/10.1016/j.jhydrol.2021.126480, 2021.

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

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., and Li, X.: The first high-resolution meteorological forcing dataset for land process studies over China, Sci. Data, 7, 25, https://doi.org/10.1038/s41597-020-0369-y, 2020.

Irving, K., Kuemmerlen, M., Kiesel, J., Kakouei, K., Domisch, S., and Jähnig, S. C.: A high-resolution streamflow and hydrological metrics dataset for ecological modeling using a regression model, Sci. Data, 5, 180 224, https://doi.org/10.1038/sdata.2018.224, 2018.

Ishikawa, Y., Yamazaki, D., and Yang, Y.: Evaluation of a Width-Based Satellite Discharge Algorithmfor Detecting Longitudinal Flow Changes in a Human-Regulated Continental River Basin, Geophys. Res. Lett., 52, e2024GL114 191, https://doi.org/10.1029/2024GL114191, 2025.

Joseph, J., Kumar, S., Merwade, V. M., and Johnson, D. R.: Direct human interventions drive spatial variability in long-term peak streamflow trends across the United States, Commun. Earth Environ., 6, 772, https://doi.org/10.1038/s43247-025-02738-8, 2025.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F. M., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J. T. M., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model Version 5: Description of new features, benchmarking, and impact of forcing uncertainty, J. Adv. Model. Earth Syst., 11, 4245–4287, https://doi.org/10.1029/2018MS001583, 2019.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Sahu, R. K., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

Lin, H., Lin, P., and Zheng, K.: Human perturbations reshape hydrological responses in riverine systems: Insights from a reach-level quantification framework to the Pearl River Basin, Sustainable Horizons, 17, 100 174, https://doi.org/10.1016/j.horiz.2025.100174, 2026.

Lin, P., Yang, Z.-L., Gochis, D. J., Yu, W., Maidment, D. R., Somos-Valenzuela, M. A., and David, C. H.: Implementation of a vector-based river network routing scheme in the community WRF-Hydro modeling framework for flood discharge simulation, Environ. Model. Softw., 107, 1–11, https://doi.org/10.1016/j.envsoft.2018.05.018, 2018.

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky, T. M., Allen, G. H., Gleason, C. J., and Wood, E. F.: Global reconstruction of naturalized river flows at 2.94 million reaches, Water Resour. Res., 55, 6499–6516, https://doi.org/10.1029/2019WR025287, 2019.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Sci. Data, 6, 283, https://doi.org/10.1038/s41597-019-0300-6, 2019.

Liu, G., Wang, W., Shao, Q., Wei, J., Zheng, J., Liu, B., and Chen, Z.: Simulating the climatic effects of irrigation over China by using the WRF-Noah model system with mosaic approach, J. Geophys. Res. Atmos., 126, e2020JD034 428, https://doi.org/10.1029/2020JD034428, 2021.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part I—A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

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., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, J. Geophys. Res. Atmos., 116, D12 109, https://doi.org/10.1029/2010JD015139, 2011.

Pool, S., Vis, M., and Seibert, J.: Regionalization for ungauged catchments: Lessons learned from a comparative large-sample study, Water Resour. Res., 57, e2021WR030 437, https://doi.org/10.1029/2021WR030437, 2021.

Su, Z., Ho, M., Hao, Z., Lall, U., Sun, X., Chen, X., and Yan, L.: The impact of the Three Gorges Dam on summer streamflow in the Yangtze River Basin, Hydrol. Process., 34, 705–717, https://doi.org/10.1002/hyp.13619, 2019.

https://doi.org/10.5194/egusphere-2026-1834
Preprint. Discussion started: 30 April 2026

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geosci. Model Dev., 11, 2429–2453, https://doi.org/10.5194/gmd-11-2429-2018, 2018.

Yang, Y., Feng, D., Beck, H. E., Hu, W., Abbas, A., Sengupta, A., Delle Monache, L., Hartman, R., Lin, P., Shen, C., and Pan, M.: Global daily discharge estimation based on grid long short-term memory (LSTM) model and river routing, Water Resour. Res., 61, e2024WR039 764, https://doi.org/10.1029/2024WR039764, 2025a.

Yang, Y., Pan, M., Feng, D., Xiao, M., Dixon, T., Hartman, R., Shen, C., Song, Y., Sengupta, A., Delle Monache, L., and Ralph, F. M.: Improving streamflow simulation through machine learning-powered data integration and its potential for forecasting in the western U.S., Hydrol. Earth Syst. Sci., 29, 5453–5476, https://doi.org/10.5194/hess-29-5453-2025, 2025b.

Yin, C., Chen, R., Xiao, X., Van de Voorde, T., Qin, Y., Guo, X., Meng, F., Pan, L., Yao, Y., and Li, Y.: Thirty years of 3-D urbanization in the Yangtze River Delta, China, Sci. Total Environ., 949, 174 909, https://doi.org/10.1016/j.scitotenv.2024.174909, 2024a.

Yin, Z., Lin, P., Riggs, R., Allen, G. H., Lei, X., Zheng, Z., and Cai, S.: A synthesis of Global Streamflow Characteristics, Hydrometeorology, and Catchment Attributes (GSHA) for large sample river-centric studies, Earth Syst. Sci. Data, 16, 1559–1587, https://doi.org/10.5194/essd-16-1559-2024, 2024b.

Zhu, M., Zhang, Z., Zhu, B., Kong, R., Zhang, F., Tian, J., and Jiang, T.: Population and economic projections for the Yangtze River Basin based on the shared socioeconomic pathways, Sustainability, 12, 4202, https://doi.org/10.3390/su12104202, 2020.