



A hybrid framework for sub-seasonal to seasonal streamflow prediction: integrating numerical and statistical models



Lingfeng Li^{1,2}, Huan Wu^{1,2,3}✉, Lulu Jiang^{1,2}, Yiwen Mei⁴, John S. Kimball⁵, Lorenzo Alfieri⁶, Zhijun Huang^{1,2}, Ying Hu^{1,2}, Sirong Chen⁷, Shaorou Dong⁸, Yaming Hu⁸ & Wei Wu⁹

Sub-seasonal to seasonal (S2S) precipitation forecasting has long been regarded as a “forecasting desert” due to limited skill beyond seven lead days, undermining downstream hydrological forecasts. However, the higher predictability of streamflow compared to precipitation, and its disproportionate improvement relative to precipitation forecast, have often been overlooked. This study integrates a distributed hydrological model with a probabilistic statistical model to enhance S2S flood forecast by assimilating statistical hydroclimate relationships. The ensemble approach is validated at 24 hydrological stations across Pearl River Basin with complex hydrology. Its modest forecasts show mean Nash–Sutcliffe Efficiency (NSE) scores ranging from 0.36 to 0.16 for weeks 2 to 6, and a 15% improvement in Continuous Ranked Probability Score Skill (CRPSS) compared to hydrological model alone. This study underscores the value of integrating physical and statistical models to improve S2S streamflow prediction, offering a practical pathway to enhance forecast skill in flood-prone basins.

Floods account for nearly half of all the natural disaster-related economic losses and fatalities, with global economic damages reaching \$651 billion between 2000 and 2019^{1,2}. As climate change intensifies extreme rainfall events, the frequency and severity of floods are projected to rise, highlighting an urgent need for reliable flood forecasting^{3–5}. Timely and accurate predictions are essential for effective disaster management and mitigation, but forecasting at the sub-seasonal to seasonal (S2S) scale remains a formidable task due to the complex interactions between atmospheric conditions and boundary factors^{6,7}.

Effective flood prediction relies on accurate precipitation forecasts across various timescales. While short-term weather forecasts have high accuracy, the skill of S2S forecasts sharply declines beyond one week, earning it the label of “forecasting desert”^{8–10}. This is primarily due to the chaotic nature of atmospheric dynamics, where initial conditions lose their influence over time, making long-term (beyond one week) predictions inherently difficult^{11,12}. Although previous studies indicate that the

predictability of streamflow is better than that of precipitation^{13,14}, few studies have been able to report quantitative performance scores, such as the Nash–Sutcliffe Efficiency (NSE) or Kling–Gupta Efficiency (KGE), for S2S flood forecasting. When available, these scores often show limited performance, revealing a significant gap in predictive skill for these time scales^{15,16}.

However, streamflow predictions can improve disproportionately compared to precipitation forecasts^{17–19}. This is not only because of the hydrological memory that streamflow retains from long-term storage of moisture in the system (e.g., groundwater and soil moisture)^{20,21}, but also due to the amplification of precipitation errors in streamflow²². Streamflow is highly sensitive to precipitation, and even minor improvements in precipitation forecasts can lead to substantial gains in streamflow prediction accuracy^{23,24}. Current S2S flood forecasting methods primarily focus on precipitation^{25,26}, often overlooking two critical sources of predictability: hydrological memory and precipitation–streamflow error propagation. Moreover, the statistical model, by capturing the underlying relationships

¹Southern Marine Science and Engineering Laboratory (Zhuhai), and School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai, China. ²Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, Sun Yat-sen University, Zhuhai, China. ³Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA. ⁴Carbon–Water Observation and Research Station in Karst Regions of Northern Guangdong, School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. ⁵Numerical Terradynamic Simulation Group, W.A. Franke College of Forestry & Conservation, University of Montana, Missoula, MT, USA. ⁶CIMA Research Foundation, Savona, Italy. ⁷Climate Center of Guangxi, Nanning, China. ⁸Climate Center of Guangdong, Guangzhou, China. ⁹National Disaster Reduction Center of China, Ministry of Emergency Management, Beijing, China. ✉e-mail: wuhuan3@mail.sysu.edu.cn

within the data, can effectively represent nonlinear relationships that are difficult to simulate directly in physical models^{27,28}. This gives statistical models a distinct advantage in handling complex and indirectly observable phenomena, thereby enhancing predictive skill. Studies have also shown that combining hydrological models with data-driven approaches can improve predictive performance compared to using data-driven models alone^{29–31}. While most previous studies concentrated on short-range or monthly-scale forecasts, several recent efforts have also examined how forecast skill evolves with lead time at the daily scale^{32–34}. Nevertheless, relatively few studies have extended such analyses to the S2S timescale, particularly using coupled physical–statistical modeling frameworks.

In this study, we extend this line of research by implementing a hybrid framework that integrates the Dominant River Tracing-Routing Integrated with Variable Infiltration Capacity Environment (DRIVE) model with a Bayesian Joint Probability (BJP) model, specifically designed for improving S2S streamflow predictions in the Pearl River Basin (PRB). The PRB was selected for its significance as one of the world's most densely populated and economically vital regions, home to 230 million people and a GDP of \$1.9 trillion. Its rain-dominant basin (453,690 km²), hilly topography, and rapid rainfall-runoff response make it one of the most challenging regions for S2S scale forecasting and among the most high-risk flood prone areas in the world^{35,36}. This complexity underscores the value of our approach, which integrates hydrological model and statistical model to overcome difficulties in S2S streamflow forecast. While the current evaluation is limited to a single large basin, the modeling strategy builds upon the Global Flood Monitoring System (GFMS)^{37–40} based on the DRIVE model and the well-established BJP model, suggesting possible broader applicability, although its transferability requires further assessment through comprehensive multi-basin validation.

Results

Overall assessment of forecasts

The streamflow in the PRB was simulated at the sub-seasonal to seasonal scale using two integration approaches (E1 and E2) that couple the physical model (DRIVE) with the statistical model (BJP) (see “Methods”). The E1 scheme primarily couples ensemble physical models and statistical models to highlight the benefits of model integration. The E2 scheme, building upon the E1 approach, further incorporates boundary conditions and S2S precipitation data to enhance the overall predictive skill. The results from both schemes were compared with those standalone model simulations to assess the simulation.

The ensemble simulations show a considerable improvement in performance over the individual physical and statistical models as lead times increase (Fig. 1a). Specifically, for the 2–6 week forecast period, the average NSE values were 0.23 for the E2 scheme, 0.22 for E1, 0.08 for DRIVE, and 0.12 for BJP. Furthermore, the boxplots indicate a marked narrowing of interquartile ranges with extended lead times, particularly for the E2 scheme, which exhibited shorter whiskers and fewer outliers. As the forecast period extends from the first week to the sixth week, the proportion of forecasts with NSE greater than 0 decreases for each scheme (Fig. 1b). For the E2 scheme, this proportion decreases from 99.7% to 98.5%; for the E1 scheme, from 98.8% to 96.9%; for the DRIVE scheme, from 92.4% to 65.9%; and for the BJP scheme, from 99.6% to 74.3% (Supplementary Table 1). This trend indicates that the E2 scheme performs better than individual models at the S2S timescale, providing good agreement with observed data and relatively small errors.

At the one-week forecast horizon, the DRIVE model achieves the highest proportion of forecasts with NSE greater than 0.5 (46.7%), outperforming the two ensemble models (34.5% and 35.7% for E1 and E2, respectively) and the BJP model (22.8%) (Fig. 1b). This indicates that physical models exhibit superior predictive skill and accuracy under stringent criteria for shorter forecast horizons. However, the performance of the physical model declines rapidly as the forecast period extends, particularly when transitioning from sub-weekly to weekly scales and from sub-monthly to monthly scales. For example, as the forecast horizon increases from the

first week to the second week, the proportion of forecasts with NSE < 0 increases from 7.6% to 15.0%, representing a 97.4% relative increase. Similarly, when the forecast horizon extends from the fourth week to the fifth week, this proportion increases from 20.2% to 28.3%, a 40.1% rise.

However, when the forecast horizon extends to 2–6 weeks, the boxplots for the top 50% of forecasts from physical models closely align with those from the ensemble models, while remaining notably higher than those from the statistical models (Fig. 1a). This suggests that although physical models exhibit considerable uncertainty at longer lead times (2–6 week), which limits their practical usage as standalone methods, they can still provide accurate predictions under scenarios with diverse climate forcing and model parameter settings. This provides the basis for ensemble models effectively mitigating the physical modeling deterioration by the incorporation of statistical components. At a forecast horizon of 6 weeks, the proportion of forecasts with NSE < 0 is only 3.1% for E1 and to 1.5% for E2, demonstrating that model coupling enhances the predictive capabilities of the physical model while addressing its rapid performance decline over longer lead times by incorporating with the statistical model.

To further assess the applicability of the proposed methods in flood event forecasting, we identified flood events at each station within the study area (see “Methods”). The simulation results from different forecasting schemes were then evaluated against the observed streamflow records, and both the Probability of Detection (POD) and the False Alarm Ratio (FAR) were calculated (Supplementary Fig. 1).

During the short-term forecast periods (1–7 and 8–14 days), DRIVE, E1, and E2 outperform BJP with notably higher POD values. Particularly for the 1–7 days interval, all three methods exhibit median POD values above 0.6, whereas BJP lags behind at ~0.45. This suggests that the physically based and hybrid models are more capable of detecting extreme events at shorter lead times. Regarding FAR, BJP exhibits consistently higher false alarm rates, especially in the 22–28 day and later intervals, where the median FAR approaches or even exceeds 0.8. This indicates a tendency toward excessive false positives in medium-to-long-range forecasts. In contrast, E1 and E2 show relatively stable FAR values, generally below 0.5, suggesting better reliability. It is worth noting that E2 achieves relatively high POD and low FAR across most lead time intervals, highlighting its superior balance between detection and false alarm. This makes it particularly suitable for sub-seasonal to seasonal (S2S) forecasting applications.

Upstream and downstream applicability evaluation

Within the PRB, the varying topography, soil types, and vegetation across different catchment areas lead to distinct hydrological responses. Therefore, assessing model performance and applicability across the river basin and forecast lead times is crucial. Figure 2 presents the performance of the models for forecast periods ranging from 1 to 44 days across 24 stations, arranged in order of generally increasing catchment area (from up to bottom along the Y-axis).

BJP models show increasing performance as catchment area increases, while the DRIVE model shows an increasing trend up to a one-week lead time (Supplementary Fig. 2). Specifically, DRIVE outperforms BJP for short lead times, particularly within the first week (Fig. 2a, b). However, as the forecast horizon extends, DRIVE experiences pronounced performance degradation in certain catchments, highlighting challenges in numerical modeling due to inefficiencies in climate forcing inputs. In contrast, although BJP generally shows lower skill than DRIVE across most locations and lead times, its performance degradation with increasing forecast horizon and variability across catchments is less pronounced. Importantly, in regions where DRIVE exhibits poor performance at the S2S timescale (highlighted by blue strips), the BJP model better captures the persistent signals from climate and hydrological variables, providing more stable and reliable forecasts at longer lead times.

Figure 2 demonstrates that the ensemble schemes E1 and E2 effectively combine the strengths of both models across the entire PRB (Fig. 2c, d). Both ensemble schemes, E1 and E2, demonstrate notable improvements over either the BJP or DRIVE schemes across the basin and forecast lead times,

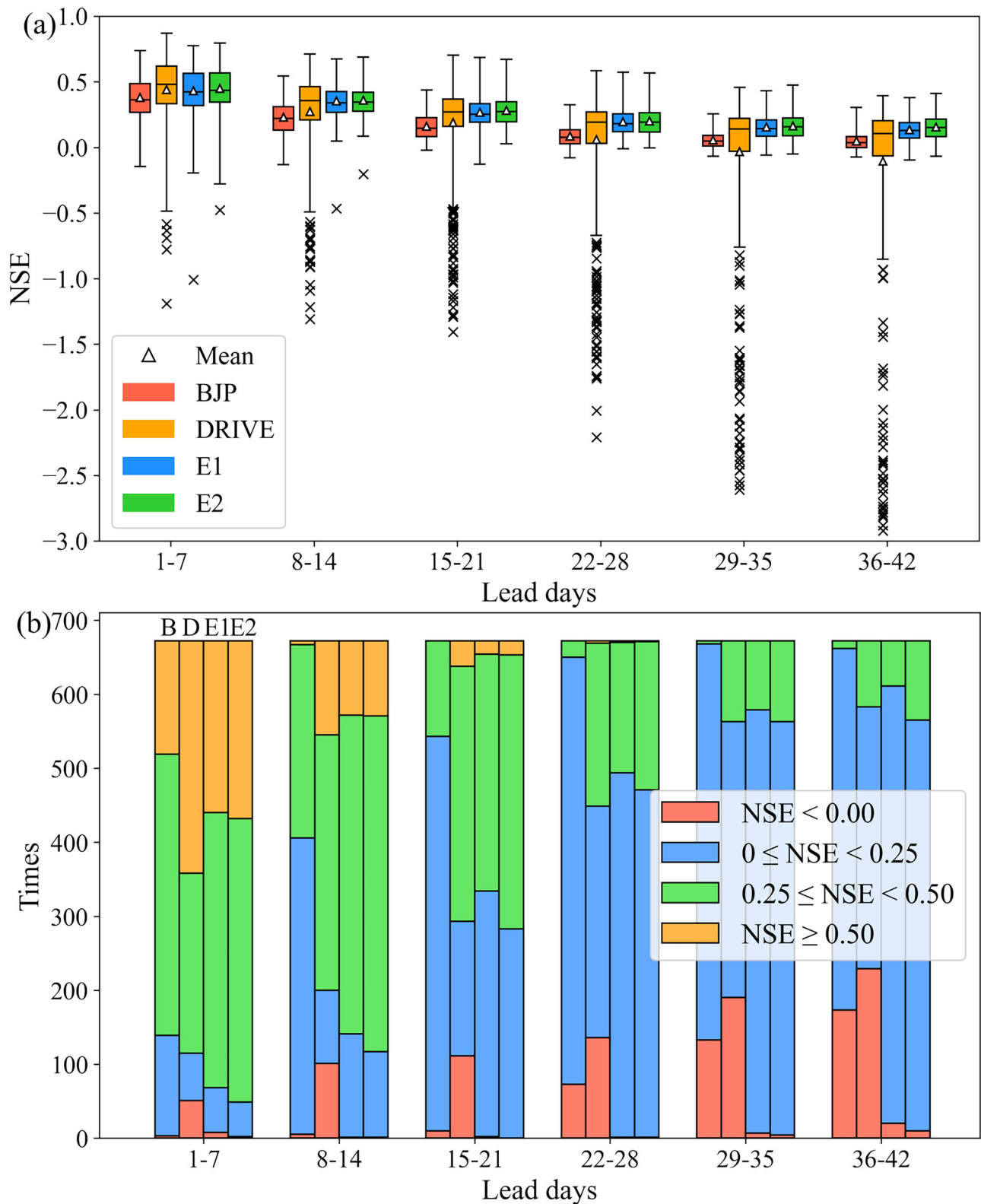


Fig. 1 | Overall distribution characteristics of NSE scores across forecasting schemes, NWP models, and lead times in the Pearl River Basin. a Statistical distribution of Nash–Sutcliffe Efficiency (NSE) scores obtained from 4 Numerical Weather Prediction (NWP) models across 24 gauging stations in the Pearl River Basin (PRB) for 6 forecast lead times. Box plots represent four forecasting schemes, with red, orange, blue, and green boxes corresponding to BJP (B), DRIVE(D), E1, and E2, respectively. For each box, the triangle indicates the mean NSE for the scheme at the corresponding lead time, while x marks denote outliers. **b** depicts the

same NSE dataset as in (a), classified into four performance categories: NSE < 0 (red), 0–0.25 (blue), 0.25–0.5 (green), and ≥0.5 (yellow), and displayed as histograms. For each of the six lead times, four bars are shown representing the four schemes (B, D, E1, and E2) to illustrate the frequency of each skill level across schemes. The result incorporates four Numerical Weather Prediction (NWP) models, 24 sites, and six leading forecast periods (lead days), resulting in a total of 672 times.

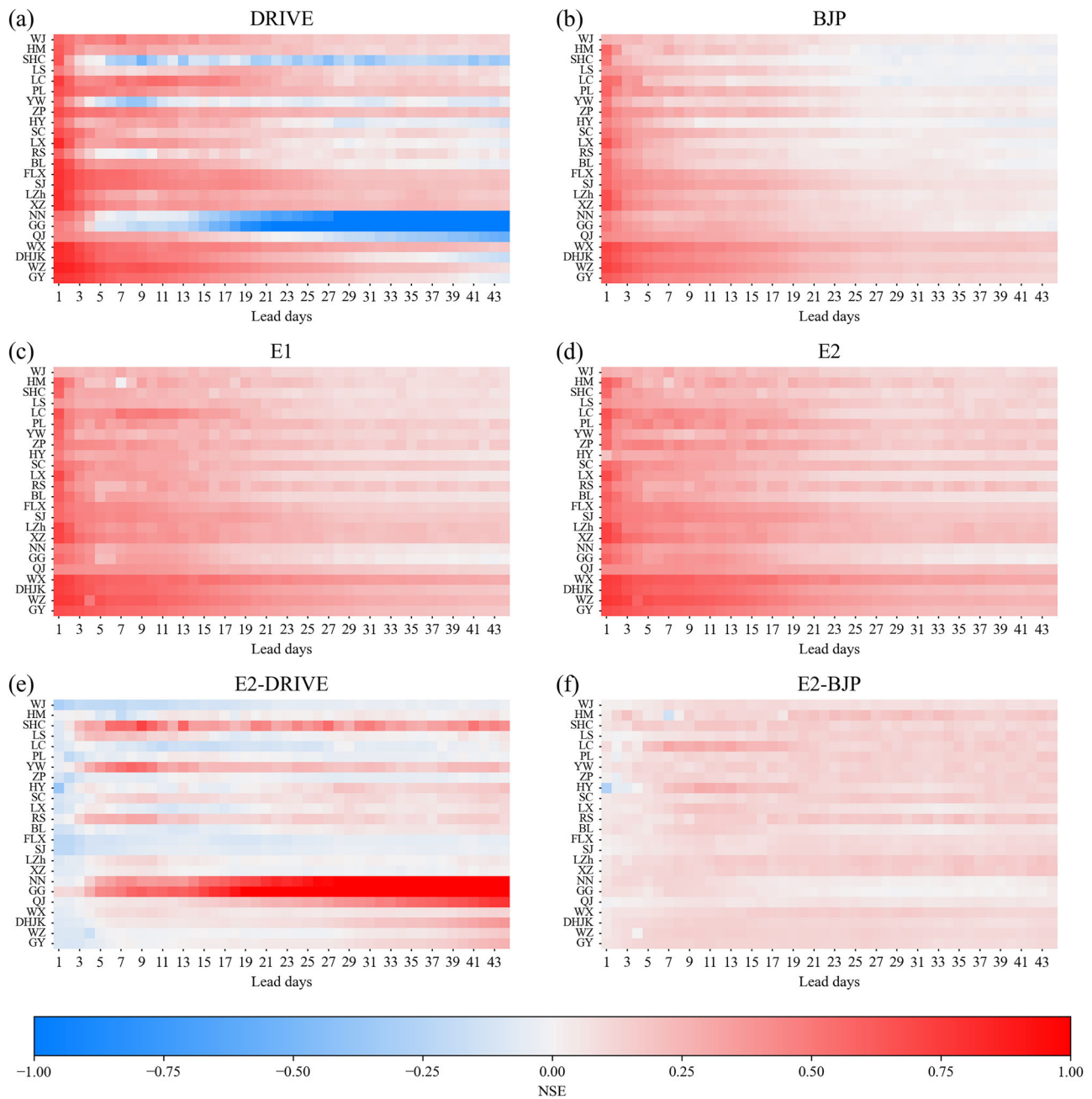


Fig. 2 | Comparison of NSE performance across schemes and improvement of the E2 scheme. a–d Present the NSE values of 24 stations across lead times of 1–44 days for four schemes (BJP, DRIVE, E1, and E2), respectively. e, f Show the differences in NSE between E2 and DRIVE, and between E2 and BJP, highlighting the

improvements introduced by the E2 scheme. The x-axis represents lead times from 1 to 44 days, and the y-axis lists stations in ascending order of catchment area from top to bottom. Station names are shown as initials, with full names provided in Supplementary Table S3.

with E2 slightly outperforming E1. The E2 ensemble mitigates the severe degradation of performance in upstream catchments by BJP and in downstream catchments by DRIVE over extended lead times (Fig. 2e, f). As a result, E2 maintains the highest skill and slowest decay, achieving a mean NSE score of 0.45, 0.36, 0.28, 0.20, 0.16 and 0.16 for weeks 1–6, respectively, demonstrating valuable forecasting skill up to six weeks. In contrast, individual methods show little to no skill for the weeks 5–6 forecasting, with DRIVE and BJP yielding negative or very low NSE scores (Supplementary Table 2), highlighting the efficacy of the proposed ensemble schemes.

Further analysis of stations with considerable improvements—such as Nanning and Guigang—shows that ensemble schemes outperform DRIVE, particularly in regions with larger catchments (Supplementary Fig. 3). As lead time increases, the decay in performance for the E2 scheme is slower

than that for DRIVE, resulting in a longer predictability horizon. Notably, the ensemble forecasts reduce the occurrence of spurious flood peaks and adjust underfitted flows, thus enhancing forecast accuracy.

Evaluation of relative performance in probabilistic forecasting

While NSE is commonly used to evaluate deterministic forecast accuracy, it does not account for the uncertainty in predictions⁴¹. In contrast, the Continuous Ranked Probability Score (CRPS) is used to quantify the difference between the predicted probability distribution and the actual observed value. Unlike a single-point forecast, CRPS accounts for both the shape of the predicted distribution and its proximity to the true observed value, providing a more comprehensive measure of forecast performance. However, the value of CRPS itself has no absolute meaning, as it is

influenced by the scale of the data and the predicted distribution. To address this issue, Continuous Ranked Probability Score Skill (CRPSS) is commonly used to assess the relative performance of probabilistic forecasts compared to a reference forecast. In this study, we use a climatology-based forecast, the DRIVE model, and the BJP model as reference forecasts to evaluate the performance of the ensemble forecasting.

A climatology-based forecast, derived solely from the historical distribution of the forecast variable (streamflow), is used as a baseline to evaluate the added value of ensemble forecasts generated from models. Results show that the two ensemble schemes (E1 and E2) consistently outperform the climatology-based forecast, demonstrating the reliability of the ensemble approach. The improvement is more pronounced at shorter lead times, indicating that the inclusion of forecasting predictors significantly enhances model performance in the short term. When the standalone physical model (DRIVE) is used as the benchmark, results show that both ensemble schemes consistently outperform it across all forecast lead times, with median CRPSS values exceeding 15%. The improvement becomes more substantial as the lead time increases. Similarly, when compared to the standalone statistical model (BJP, Fig. 3b), both ensemble schemes again show superior performance at all lead times, with median CRPSS values exceeding 8%.

In addition to CRPSS-based comparisons, the PIT histogram results indicate that the E2 scheme provides a notably better fit to the distribution of observations compared to the BJP scheme (Supplementary Fig. 4). A two-sample Kolmogorov–Smirnov test confirms that the distributions of the PIT values from the two schemes are significantly different (p -value < 0.01). Specifically, the BJP scheme exhibits a pronounced inverted U-shaped pattern, suggesting that the forecasts are over dispersed with wider distributions. This might reflect an overly conservative estimation by the model. Additionally, a sharp peak near one implies that the model fails to capture extreme flood events, while the E2 scheme substantially mitigates both issues, with the improvements more pronounced at longer lead times.

To further evaluate forecast uncertainty, we conducted an additional analysis comparing ensemble spread (i.e., the standard deviation of ensemble members) with the actual forecast error (i.e., the difference between the ensemble mean and the observation) at a representative station (Supplementary Fig. 5). The average correlation coefficient between ensemble spread and actual forecast error is 0.53 ($p < 0.001$), indicating a significant moderate positive relationship. This suggests that the ensemble forecast uncertainty can to some extent reflect the variation of the actual error. However, the average slope is only 0.39, implying that the ensemble spread generally overestimates the actual error and that there is some degree of overdispersion or miscalibration. In summary, the current ensemble forecasting system can reasonably capture changes in uncertainty, but its quantification of uncertainty can still be improved. Future work could focus on calibration methods to enhance the reliability of ensemble spread and the accuracy of forecasts.

These results highlight the advantages of the ensemble approach over individual models in probabilistic forecasting. Combining deterministic and uncertainty metrics provides a comprehensive assessment, confirming the model's accuracy and adaptability under uncertainty, thus aiding in decision-making. This integrated approach helps users evaluate the model's effectiveness and reliability in practical applications.

Discussion

In this study, we propose an ensemble approach that combines a distributed hydrological model (DRIVE) with a statistical model (BJP) to simulate S2S streamflow in the Pearl River Basin. The DRIVE hydrological model simulates the watershed hydrological characteristics, while the BJP statistical model captures relationships among initial flow, sub-seasonal precipitation, forecasted streamflow, and actual streamflow for predictions. We compare the predictive performance of this combined approach with individual models and assess its applicability across various stations within the PRB. To better illustrate the differences in S2S streamflow forecasting skills across models and how these differences may relate to predictability driven by the

atmosphere, land surface, and ocean, we compared the variation in model skills across lead times against the conceptual predictability framework proposed by Dirmeyer⁴² which is based on the well explored literature^{43–46}, as shown in Fig. 4. It is important to note that it represents a theoretical upper bound under idealized conditions for qualitative reference, rather than for direct quantitative comparison.

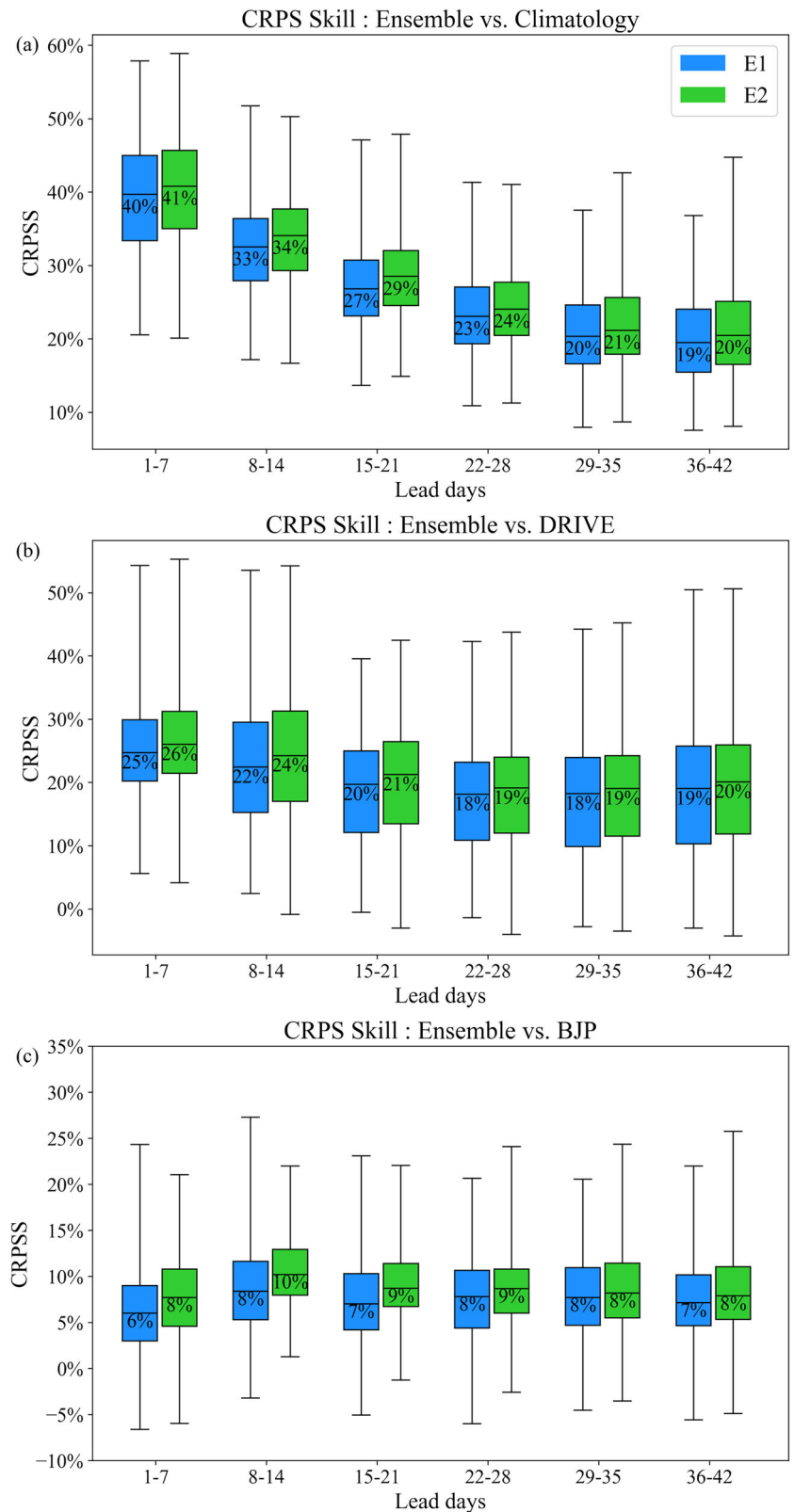
Among the models evaluated, DRIVE shows superior performance in short-term forecasts, while the E2 ensemble approach outperforms others for mid-to-long term forecasts by achieving both higher accuracy and lower variability in S2S streamflow predictions. In short-term forecasting, where precipitation forecasts are generally accurate, DRIVE effectively simulates runoff processes, providing corresponding predictions. However, for longer lead times, errors in precipitation forecasts tend to accumulate, which amplifies streamflow errors in physically based models that rely on partial differential equations. This leads to a rapid decline in simulation performance^{49,47}. However, the performance of DRIVE deteriorates after the forecast horizon exceeds one week (depending on the upstream drainage area) and may even improve when the land surface processes, influenced by hydrological memory, come into play (orange line in Fig. 4a)⁴⁸. The role of the land surface becomes increasingly important for improving forecast accuracy beyond one week⁴². Not only do atmospheric initial conditions matter, but the moisture memory of the underlying surface also influences streamflow forecasts^{49,50}. Therefore, although the forecasting ability of physical models may degrade with longer lead time, this degradation is not as substantial as that of atmospheric predictability (as seen in the orange line in Fig. 4a compared to yellow line in Fig. 4b).

Additionally, flood occurrence is heavily influenced by climatic factors. Studies suggest that these climate drivers could have a greater impact on hydro-meteorological processes than precipitation alone, complicating flood forecasting at the S2S scale^{21,50}. The introduction of the BJP model addresses the limitations of relying on physical models. It uses historical data to fit joint probability distributions, which helps capture complex relationships between key variables such as precipitation, initial flow, and streamflow. Therefore, The BJP model begins to outperform the DRIVE model at approximately three weeks (red line in Fig. 4a).

Different basins exhibit considerable variation in underlying surface characteristics, such as topographic relief, soil types, and vegetation cover, all of which influence runoff generation and concentration processes. Given the large variability in catchment areas across different stations, the spatial heterogeneity results in differing response characteristics of physical models to precipitation in runoff simulations. This is particularly evident at stations with steep slopes or highly undulating terrain, where the strong runoff response makes the physical model highly sensitive to changes in precipitation^{51,52}. As the forecast period increases and precipitation forecast accuracy decays, the S2S streamflow prediction for these stations becomes more challenging. However, the Bayesian joint probability forecasting method can mitigate this sensitivity by combining prior knowledge with new observational data through prior distributions and likelihood functions, generating a probability distribution to express uncertainty. In this study, we employed an ensemble forecasting approach to leverage the high simulation performance of the DRIVE model in the short-term forecast period, while also utilizing the BJP model to capture the complex relationships among key variables, thereby extending the predictability of streamflow simulations over a longer forecast horizon. Additionally, we incorporated initial boundary conditions and climate data to further enhance the simulation accuracy and performance.

However, the performance of these forecasting methods is not always consistent, as they are influenced by a variety of factors, including model uncertainty and the quality of input data. As a result, evaluating the effectiveness and accuracy of probabilistic forecasting becomes crucial, particularly in regions with complex basin characteristics. Probabilistic forecasting offers an effective means of addressing uncertainty and enhancing forecast reliability, particularly in complex catchments. The results indicate that ensemble forecasts outperform predictions from the DRIVE model (CRPSS median > 15%)

Fig. 3 | CRPS distributions across lead times for different ensemble schemes. Continuous Ranked Probability Score Skill (CRPS) is shown relative to a the climatology forecast, b the DRIVE model, and c the BJP model at 24 evaluation gauges. Blue boxes represent ensemble scheme E1, and green boxes represent ensemble scheme E2.



and the BJP model (CRPS median > 8%). As a result, both ensemble approaches outperform the individual BJP method and the DRIVE model in terms of forecast accuracy.

To sum up, recent advancements in NWP, DRIVE, and BJP modeling demonstrate that the integration of physical and statistical models notably improve the forecasting ability of streamflow at the S2S scale. Previous

studies indicate that streamflow forecast lead times were generally limited to 5–10 days (NSE > 0)¹⁵, and KGE values of ≤ −0.50 were observed for large flood events at a 2-week lead time¹⁶. In comparison, the results presented here extend forecast skill at longer lead times. Although ensemble methods generally improve forecast robustness, their use of mean or median aggregation may smooth out extreme signals, potentially leading to inferior

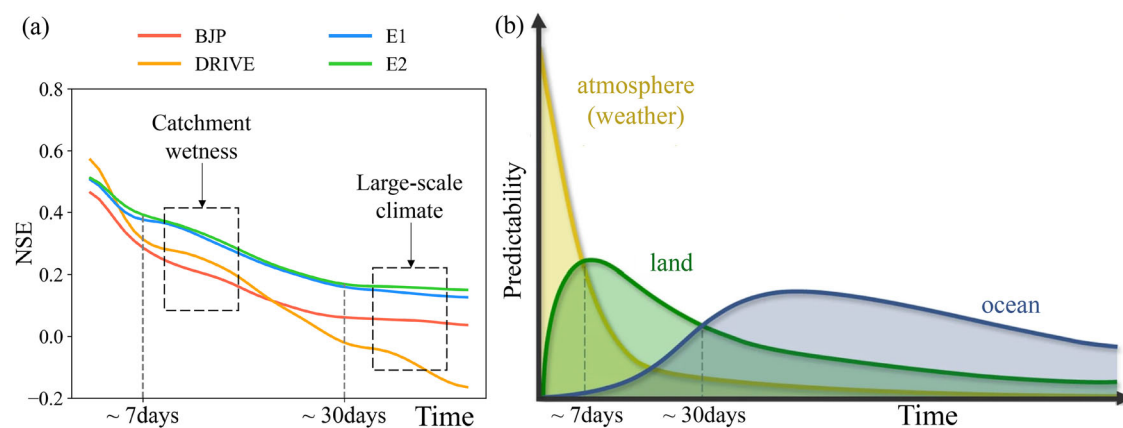


Fig. 4 | Temporal evolution of streamflow predictability. **a** Calculated streamflow predictability performance with the methods, i.e., the statistical (BJP red line) and physical process (DRIVE orange line) models, and their ensembles (E1 blue line, E2 green line), declines as forecast lead time gets longer. **b** Conceptual schematic of S2S predictability adapted from Paul Dirmeyer⁴². The roles of the atmosphere (yellow), land surface (green), and ocean (blue) are highlighted: short lead forecasts depend

mainly on the atmosphere, 2–4 week forecasts require land surface information, and forecasts beyond 30 days rely on ocean conditions such as sea surface temperature variations linked to El Niño. It represents the theoretical upper bound of forecast skill under ideal conditions. This framework is included solely to provide qualitative context for interpreting the temporal evolution of actual model skill.

performance compared to well-initialized physical models in short-term forecasts. As a result, the benefits of integration are relatively limited for short lead times (green line in Fig. 4a). The superior performance of physical models in this range is largely attributed to their high sensitivity to accurate initial and boundary conditions, which can typically be well represented over short forecasting horizons.

In contrast, at longer lead times, the integration of physical and statistical models proves more beneficial. The E2 scheme outperforms the E1 scheme likely due to the implicit interdependencies among variables embedded in initial conditions and S2S precipitation, which are better captured through a hybrid approach. This synergy enhances the predictability of S2S streamflow and underscores the value of model integration, not as a simple substitution, but as a complementary strategy that leverages the strengths of each component: process understanding from physical models and data-driven correction from statistical models.

While the combination of models may not be optimal and exhibits limitations in underestimate of flood peak magnitudes and increasing FAR at longer lead times, this study provides a valuable proof of concept for integrating physical and statistical approaches on the S2S timescale. In current operational flood forecasting, although the overall performance of the hybrid framework surpasses that of the standalone physical model, the physical model (DRIVE) exhibits superior capability in representing high-flow magnitudes and extreme events. This strength of physically based models should not be ignored as they provide process-based information that statistical approaches tend to overlook. Therefore, the hybrid framework should not be regarded as a complete replacement for physical models, but rather as a complementary component within an integrated forecasting system to enhance the lead time and reliability of flood prediction. Although restricted to the Pearl River Basin with 3.5 years of data, the findings suggest that the insights gained here can inform future applications with longer records and across diverse hydroclimatic regions. In particular, future work should include extending the record length and test the framework in multiple basins representing diverse hydroclimatic settings, to provide a more robust and generalizable evaluation. Furthermore, recent developments in fully differentiable and physics-guided data science frameworks, such as physics-informed neural networks and differentiable hydrological models, offer promising ways to incorporate physical knowledge into machine learning models. These hybrid approaches are expected to enhance the accuracy, interpretability, and reliability of S2S forecasts, especially alongside ongoing improvements in NWP precipitation forecasting and AI techniques.

Methods

Streamflow data

The Pearl River is the third longest river and the second largest in terms of basin area in China, covering 453,690 square kilometers (Fig. 5a). The Pearl River Basin (PRB), located between 18°N and 27°N and 100°E and 118°E, is characterized by a subtropical climate and diverse hydrological features. The mean annual temperature ranges from 14 °C to 22 °C, and annual precipitation varies between 1200 mm and 2200 mm⁵³. The basin's average annual runoff is 260 km³, ranking second in flow magnitude in China, sixth in Asia, and eighteenth globally⁵⁴. The influence of the East Asian monsoon leads to uneven distribution of discharge throughout the year. Specifically, the accumulated discharge from April to September accounts for ~80% of the total annual discharge, with May–August contributing to more than 50% of the total. To evaluate the simulation's accuracy, streamflow data from 24 gauging stations within the PRB were utilized (Fig. 5a and Supplementary Table 3). The streamflow data were obtained from the open source at National Hydrological and Rainfall Information Network, Ministry of Water Resources of China (http://xxfb.mwr.cn/sq_djdh.html). These data offer valuable insights into the region's diverse hydrological responses. The locations and detailed information of these gauging stations provide critical support for understanding the hydrological characteristics of the basin.

Meteorological data

In this study, meteorological data including satellite-derived precipitation, forecast precipitation, and other atmospheric variables were used. Specifically, satellite-derived precipitation data were utilized for the control experiment simulations, while forecast precipitation data were employed for streamflow predictions in the DRIVE model's forecasting mode. Additional meteorological variables were applied for the model initialization.

Satellite-derived precipitation data were sourced from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS v2.0 <https://www.chc.ucsb.edu/data/chirps>). This dataset integrates real-time station data with satellite infrared imagery to provide accurate precipitation estimates⁵⁵. This data records global precipitation from 1981 to the present, with a spatial resolution of 0.05° and daily gridded precipitation time series, covering latitudes from 50°N to 50°S.

Forecast precipitation data are obtained from the ECMWF public dataset S2S prediction, which comprises 13 models providing forecasts for variables at various pressure levels and surface variables⁵⁶. Forecast range is up to 60 days at a spatial resolution of 1.5 degrees. We established a selection criterion demanding models to provide initialized forecasts spanning from

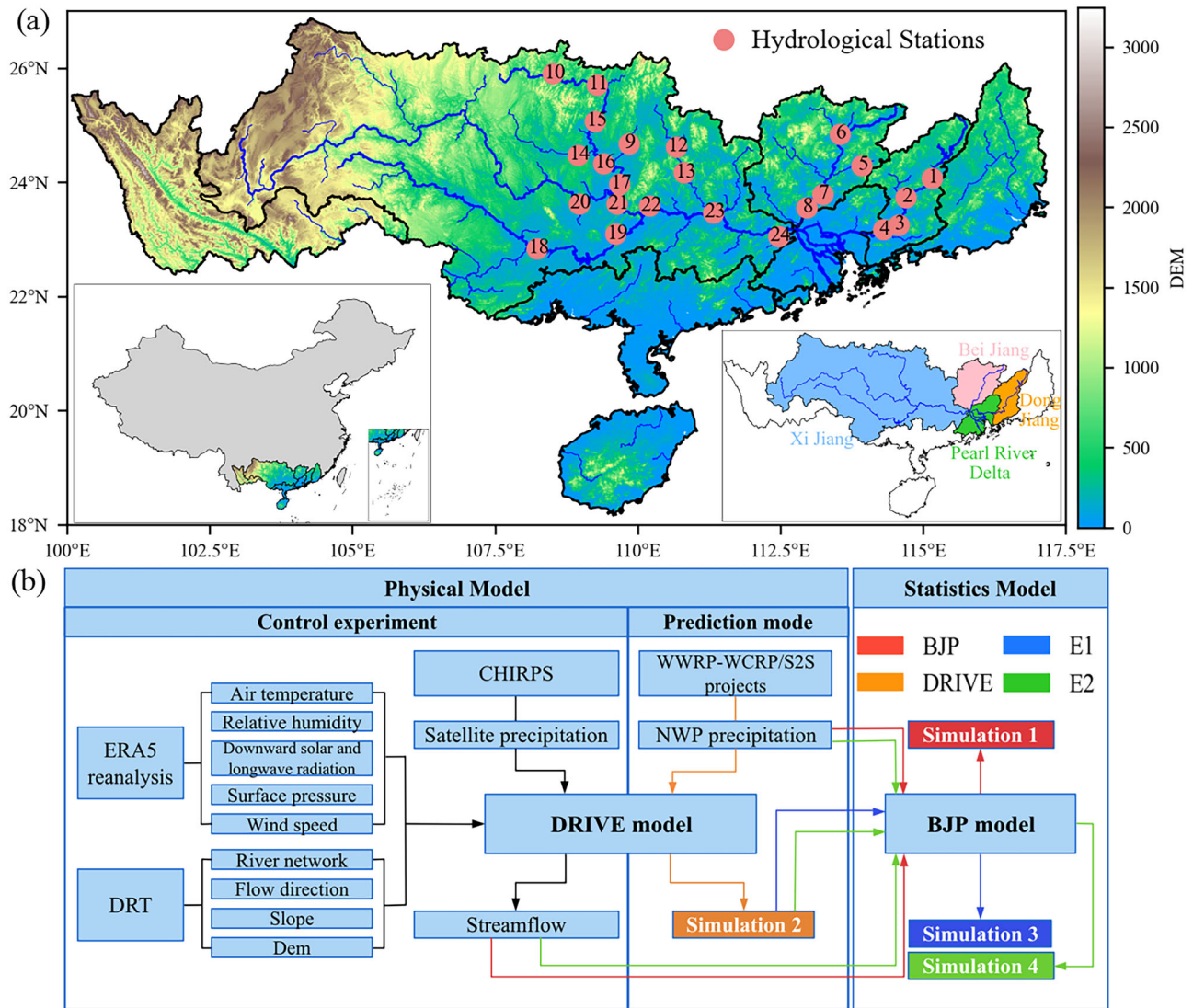


Fig. 5 | Study area and methodological framework. **a** Digital elevation map (DEM) and river network of the Pearl River Basin, with gauging locations marked by red dots and labeled by their respective catchment IDs. The inset (left) shows the location of the Pearl River Basin within China, in addition to the PRB sub-basin layout, including the DongJiang, BeiJiang, and XiJiang sub-basins as well as the Pearl

River Delta (right). **b** Overall flowchart of this study, illustrating the integration of the DRIVE model (orange line), the BJP model (red line), and their combined modeling approaches (E1: blue line; E2: green line). All outputs represent streamflow.

2016 to 2022, ensuring continuous forecast availability, in near-real-time by 2023, with at least a 44-day lead time. Four models, namely ECMWF, KMA, UKMO, and NCEP, met this criterion, as detailed in Supplementary Table 4. The forecast quality of these models has been independently evaluated⁴⁸, which showed their relative strengths and limitations. Therefore, we do not repeat the detailed precipitation skill evaluation here and instead focus on assessing the hydrological predictability based on these validated inputs.

Other atmospheric variables including air temperature, downward solar and longwave radiation, relative humidity, surface pressure, and wind speed are from ERA-5 reanalysis, offering hourly temporal resolution and 0.25-degree spatial resolution⁵⁷.

Hydrological modeling

The DRIVE model integrates the Variable Infiltration Capacity Macroscale Hydrologic Model (VIC) with the Dominant River Tracing based Routing Model (DRTR)^{58–61}. VIC calculates surface and subsurface runoff by explicitly representing precipitation partitioning into rainfall and snowfall, snow accumulation and melt, infiltration into a multi-layer soil column, evapotranspiration (including canopy interception, soil evaporation, and

vegetation transpiration), and drainage that generates both surface runoff and baseflow. This model employs a variable infiltration curve that accounts for multi-layer soil infiltration and various land cover types to compute the runoff volume for each grid cell ($0.125^\circ \times 0.125^\circ$)^{62,63}. The meteorological forcing was downscaled using nearest neighbor interpolation. The DRTR routing module simulates the movement of water in terrestrial and river grid cells, conducting flood forecasting computations by solving the kinematic wave equation for natural watershed systems, based on Strahler order river network coding⁵⁸. This provides results for each time step, including river flow, surface water storage, inundation depth, and extent^{61,64}.

The performance of DRIVE has been extensively validated and widely applied in various research fields, providing crucial scientific support for flood prevention and disaster reduction efforts^{61,65–70}. In a companion study⁴⁸, DRIVE forced with S2S precipitation was benchmarked against the Ensemble Streamflow Prediction (ESP) method, demonstrating reasonable performance and providing detailed calibration and validation for the Pearl River Basin. Building on this foundation, the present study integrates DRIVE (VIC + DRTR) with the probabilistic BJP model and positions the resulting forecast skill

within the context of prior S2S research^{15,16}, showing that the integrated framework achieves skill comparable to or exceeding previously reported values in rain-dominated settings.

Statistical modeling

The BJP model, originally designed for seasonal streamflow prediction^{71–75}, has been extended to effectively forecast sub-seasonal to seasonal streamflow patterns^{30,76}. By capturing the statistical relationships between key hydrological variables, the BJP model has demonstrated improved predictive performance across various lead times. It describes the relationship between predictors and predictands using marginal and conditional probability distributions, which allows flexible handling of non-linear and asymmetric dependencies. While the full derivation of the underlying distributions and the Gibbs sampling procedure is detailed in previous studies, including the algorithm and pseudocode by Wang et al.⁷⁷, we provide only a brief summary of the implementation relevant to this study.

To standardize all variables and improve model performance, we apply a log-sinh transformation as follows:

$$z = \frac{1}{\lambda} \log(\sinh[\varepsilon + \lambda y]) \quad (1)$$

where y is the pre-transformation variable, z is the post-transformation variable, and ε and λ are transformation parameters. The Bayesian Maximum Posteriori method⁷⁸ is used to estimate optimal parameter values, assuming the transformed variables follow a multivariate normal distribution:

$$z \sim N(\mu, \Sigma) \quad (2)$$

where μ and Σ are the mean and covariance matrices, respectively. Through Bayesian inference, we explore the posterior distribution:

$$p(\theta|D) \propto p(\theta)p(D|\theta) \quad (3)$$

where θ represents the parameters, including μ and Σ . $p(\theta)$ is the prior distribution of the parameters, $p(D|\theta)$ is the likelihood function, and $D = \{z(t), t = 1, 2, \dots, n\}$ is the dataset. We use the non-informative multivariate Jeffreys prior⁷⁹, a common objective prior in Bayesian statistics. It is designed to minimize subjective input, allowing the data to play a dominant role in the inference process.

$$p(\theta) = p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2} \quad (4)$$

where d represents the combined number of predictands and predictors.

Using the BJP model, we calibrate new prediction results $y(t')$ or $z(t')$ as follows:

$$f(z_2(t')) = \int p(z_2(t')|z_1(t'), \theta)p(\theta|D)d\theta \quad (5)$$

We obtain samples of $z_2(t')$ through Gibbs sampling. Gibbs sampling simplifies the computation process of high-dimensional complex models by sequentially sampling from the conditional distributions, thereby effectively estimating the probability of flooding and related uncertainties. In this study, we assessed convergence based on the stability of posterior parameters and the length of the sampling process. For the BJP sampler, we ran 6000 iterations and discarded the first 2000 as burn-in. The trace plots of key parameters (means and covariance elements) show that most sampled values lie within ± 2 standard deviations of the posterior mean, exhibiting only minor fluctuations around the mean and no long-term drift (Supplementary Fig. 6). The visual diagnostics indicate that the chains are well-mixed, stationary, and have successfully converged. Finally, the backward transformation method is applied to convert the transformed variables $z_2(t')$ into forecasted streamflow values $y_2(t')$.

Integrating numerical and statistical modeling

This study explores the integration of the DRIVE and BJP models to enhance S2S streamflow forecasting. The DRIVE model is executed using catchment attribute datasets and meteorological forcing data, including ERA5 reanalysis and CHIRPS precipitation data. These results serve as a benchmark experiment, providing a well-characterized depiction of the basin's initial conditions for accurate forecasting (black line in Fig. 5b).

The coupling strategy between DRIVE and BJP was designed to systematically utilize physically based hydrological states as predictors in the statistical framework. Specifically, DRIVE outputs daily streamflow simulations for each ensemble member under S2S precipitation forecasts. For the integration, we extracted ensemble-mean streamflow values from DRIVE forecasts and used them as predictor variables for BJP, thereby incorporating physically simulated streamflow information into the statistical post-processing.

To further enhance predictive performance, two configurations of coupling were tested:

- (1) E1: BJP model uses DRIVE-forecasted streamflow as the sole predictor, reflecting the value of hydrological simulations without external predictors (blue line in Fig. 5b).
- (2) E2: BJP model uses a combined predictor set including DRIVE-forecasted streamflow, observed streamflow on the initial date, and S2S precipitation forecasts, to capture both initial state information and anticipated meteorological drivers (green line in Fig. 5b).

This multi-input design was implemented in BJP by extending its predictor matrix and applying the same Bayesian joint probability modeling approach across all predictors.

To systematically evaluate performance and highlight the added value of the coupled approach, we designed two control schemes:

- (1) BJP: BJP model with streamflow on the initial date and S2S precipitation (red line in Fig. 5b).
- (2) DRIVE: DRIVE forecasting mode with S2S precipitation input (orange line in Fig. 5b).

The ensemble generation process for DRIVE was based on all available S2S ensemble members. Each member was run independently through DRIVE's hydrological and routing modules, which consist of VIC for water and energy balance at 0.125° resolution and DRTR for routing via the kinematic wave equation. Model outputs were then averaged to provide a robust predictor for the BJP model, minimizing the effect of individual ensemble uncertainty.

We adopted a leave-one-day-out cross-validation over the full period from June 2019 to December 2022. Each day was sequentially used as the validation set, with all other days for training, to evaluate overall model performance across diverse hydrological conditions. This comprehensive approach covers normal and flood events throughout the entire timeframe, providing a detailed assessment of the model's robustness and generalization ability. This ensemble approach amalgamates various modeling techniques to improve streamflow prediction capabilities compared to using a single model alone.

Validation metrics

The efficacy of model simulations is evaluated using both deterministic and probabilistic forecast indicators. For the deterministic assessment, the Nash–Sutcliffe Efficiency (NSE) is used to quantify the proximity between observed and simulated data. The NSE ranges from negative infinity to 1, with values closer to 1 signifying superior predictive precision. It is calculated as:

$$NSE = 1 - \sum_{i=1}^n (O_i - P_i)^2 / (O_i - \bar{O})^2 \quad (6)$$

where O_i is the observed value; P_i is the predicted value; \bar{O} is the mean of the observed values; and n is the number of observations.

In addition to this continuous metric, categorical verification metrics were used to assess the capability of the forecasting methods in detecting flood events. We defined as streamflow values exceeding the 90th percentile plus 0.5 times the standard deviation²¹. Binary flood labels were assigned to both observed and predicted series based on this threshold. Using the resulting contingency table, two skill scores were computed:

$$\text{Probability of Detection (POD)} = \frac{H}{H + M} \quad (7)$$

$$\text{False Alarm Ratio (FAR)} = \frac{F}{H + F} \quad (8)$$

Here, POD quantifies the fraction of observed flood events that were correctly forecast, FAR reflects the proportion of predicted flood events that did not occur. H refers to the number of correctly predicted flood events (hits), M is the number of missed flood events, and F denotes the number of false alarms.

Probabilistic forecasts are assessed utilizing the Continuous Ranked Probability Score (CRPS)⁸⁰ and Probability Integral Transform (PIT)⁸¹, which measure the discrepancy between the predicted cumulative distribution and the observed outcome. CRPS is calculated as:

$$\text{CRPS}(F, y) = \int [F(x) - H(y - x)]^2 dx \quad (9)$$

where $F(x)$ is the predicted cumulative distribution function (CDF) and $H(y - x)$ is the Heaviside step function. Lower CRPS values indicate better agreement between forecasted and observed distributions. The model predictions were also compared with reference predictions using skill scores:

$$\text{CRPS Skill} = \frac{\overline{\text{CRPS}}_{\text{ref}} - \overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{ref}}} \times 100\% \quad (10)$$

Higher CRPS values indicate greater forecasting accuracy.

For observed values x_1, x_2, \dots, x_n , suppose the model provides corresponding cumulative distribution functions F_1, F_2, \dots, F_n . With well-calibrated, the Probability Integral Transform (PIT) values generally follow a uniform distribution on the interval [0, 1]:

$$z_i = F_i(x_i), \text{ for } i = 1, 2, \dots, n \quad (11)$$

To evaluate whether the PIT values from different schemes are drawn from the same distribution, we perform the two-sample Kolmogorov–Smirnov (KS) test. The KS test statistic is defined as:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)| \quad (12)$$

where $F_n(x)$ and $G_m(x)$ represent the empirical cumulative distribution functions (ECDFs) of the two samples. The test measures the maximum distance between them to evaluate whether the two distributions differ significantly.

Data availability

The Climate Hazards Group InfraRed Precipitation with Station data can be accessed through the Climate Hazards Center website (<https://www.chc.ucsb.edu/data/chirps>). The Forecast precipitation data are obtained from the ECMWF public dataset S2S prediction and is available online at (<https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-kwbc/levtype=sfc/type=cf/>). ERA-5 reanalysis data is available online at (<https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview>). The streamflow data were obtained from the open source at National Hydrological and Rainfall Information Network, Ministry of Water Resources of China (http://xxfb.mwr.cn/sq_djdh.html).

Code availability

The codes used for our calculations are available on request from authors.

Received: 6 May 2025; Accepted: 12 November 2025;

Published online: 28 November 2025

References

1. United Nations Office for Disaster Risk Reduction (UNDRR). Human Cost of Disasters (UN, 2020). <https://doi.org/10.18356/79b92774-en>.
2. Rentschler, J., Salhab, M. & Jafino, B. A. Flood exposure and poverty in 188 countries. *Nat. Commun.* **13**, 3527 (2022).
3. Blöschl, G. et al. Current European flood-rich period exceptional compared with past 500 years. *Nature* **583**, 560–566 (2020).
4. Hallegatte, S., Green, C., Nicholls, R. J. & Corfee-Morlot, J. Future flood losses in major coastal cities. *Nat. Clim. Change* **3**, 802–806 (2013).
5. Winsemius, H. C. et al. Global drivers of future river flood risk. *Nat. Clim. Change* **6**, 381–385 (2016).
6. Cloke, H. L. & Pappenberger, F. Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009).
7. Jain, S. K. et al. A Brief review of flood forecasting techniques and their applications. *Int. J. River Basin Manag.* **16**, 329–344 (2018).
8. White, C. J. et al. Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorol. Appl.* **24**, 315–325 (2017).
9. Phakula, S., Landman, W. A. & Engelbrecht, C. J. Literature survey of subseasonal-to-seasonal predictions in the southern hemisphere. *Meteorol. Appl.* **31**, e2170 (2024).
10. Vitart, F., Robertson, A. W. & Anderson, D. L. T. Subseasonal to seasonal prediction project: bridging the gap between weather and climate. *Bull. World Meteorol. Org.* **61**, 23 (2012).
11. Monhart, S. et al. Skill of subseasonal forecasts in Europe: effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.* **123**, 7999–8016 (2018).
12. Magnusson, L. & Källén, E. Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Weather Rev.* **141**, 3142–3153 (2013).
13. Greuell, W., Franssen, W. H. P. & Hutjes, R. W. A. Seasonal streamflow forecasts for Europe – Part 2: sources of skill. *Hydrol. Earth Syst. Sci.* **23**, 371–391 (2019).
14. Pechlivanidis, I. G., Crochemore, L., Rosberg, J. & Bosshard, T. What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resour. Res.* **56**, e2019WR026987 (2020).
15. Cao, Q., Shukla, S., DeFlorio, M. J., Ralph, F. M. & Lettenmaier, D. P. Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in Coastal Western U.S. watersheds. *J. Hydrometeorol.* **22**, 1535–1552 (2021).
16. Monhart, S., Zappa, M., Spirig, C., Schär, C. & Bogner, K. Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: benefits of the NWP approach. *Hydrol. Earth Syst. Sci.* **23**, 493–513 (2019).
17. Ehsan Bhuiyan, M. A. et al. Assessment of precipitation error propagation in multi-model global water resource reanalysis. *Hydrol. Earth Syst. Sci.* **23**, 1973–1994 (2019).
18. Maggioni, V. et al. Investigating the applicability of error correction ensembles of satellite rainfall products in river flow simulations. *J. Hydrometeorol.* **14**, 1194–1211 (2013).
19. Nanding, N. et al. Assessment of precipitation error propagation in discharge simulations over the contiguous United States. *J. Hydrometeorol.* **22**, 1987–2008 (2021).
20. Shukla, S., Sheffield, J., Wood, E. F. & Lettenmaier, D. P. On the sources of global land surface hydrologic predictability. *Hydrol. Earth Syst. Sci.* **17**, 2781–2796 (2013).
21. Yan, Y. et al. Exploring the ENSO impact on Basin-scale floods using hydrological simulations and TRMM precipitation. *Geophys. Res. Lett.* **47**, e2020GL089476 (2020).

22. Wu, H., Adler, R. F., Tian, Y., Gu, G. & Huffman, G. J. Evaluation of quantitative precipitation estimations through hydrological modeling in IFloodS river basins. *J. Hydrometeorol.* **18**, 529–553 (2017).
23. Sankarasubramanian, A., Vogel, R. M. & Limbrunner, J. F. Climate elasticity of streamflow in the United States. *Water Resour. Res.* **37**, 1771–1781 (2001).
24. Zhang, Y., Viglione, A. & Blöschl, G. Temporal scaling of streamflow elasticity to precipitation: a global analysis. *Water Resour. Res.* **58**, e2021WR030601 (2022).
25. Li, W. et al. Evaluation and bias correction of S2S precipitation for hydrological extremes. *J. Hydrometeorol.* **20**, 1887–1906 (2019).
26. Zhang, L., Gao, S. & Yang, T. Adapting subseasonal-to-seasonal (S2S) precipitation forecast at watersheds for hydrologic ensemble streamflow forecasting with a machine learning-based post-processing approach. *J. Hydrol.* **631**, 130643 (2024).
27. Schick, S., Rössler, O. & Weingartner, R. An evaluation of model output statistics for subseasonal streamflow forecasting in European catchments. *J. Hydrometeorol.* **20**, 1399–1416 (2019).
28. Zhao, T., Schepen, A. & Wang, Q. J. Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach. *J. Hydrol.* **541**, 839–849 (2016).
29. Konapala, G., Kao, S. C., Painter, S. L. & Lu, D. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environ. Res. Lett.* **15**, 104022 (2020).
30. Lu, D., Konapala, G., Painter, S. L., Kao, S. C. & Gangrade, S. Streamflow simulation in data-scarce basins using Bayesian and physics-informed machine learning models. *J. Hydrometeorol.* **22**, 1421–1438 (2021).
31. Frame, J. M. et al. Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *J. Am. Water Resour. Assoc.* **57**, 885–905 (2021).
32. Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E. & Schepen, A. Reliable long-range ensemble streamflow forecasts: combining calibrated climate forecasts with a conceptual runoff model and a staged error model. *Water Resour. Res.* **52**, 8238–8259 (2016).
33. Cheng, M., Fang, F., Kinouchi, T., Navon, I. M. & Pain, C. C. Long lead-time daily and monthly streamflow forecasting using machine learning methods. *J. Hydrol.* **590**, 125376 (2020).
34. Hapuarachchi, H. A. P. et al. Development of a national 7-day ensemble streamflow forecasting service for Australia. *Hydrol. Earth Syst. Sci.* **26**, 4801–4821 (2022).
35. Wu, C. S., Yang, S. L. & Lei, Y. Quantifying the anthropogenic and climatic impacts on water discharge and sediment load in the Pearl River (Zhujiang), China (1954–2009). *J. Hydrol.* **452–453**, 190–204 (2012).
36. Duan, W. et al. Floods and associated socioeconomic damages in China over the last century. *Nat. Hazards* **82**, 401–413 (2016).
37. Wu, H. et al. From China's Heavy Precipitation in 2020 to a "Glocal" Hydrometeorological Solution for Flood Risk Prediction. *Adv. Atmos. Sci.* **38**, 1–7 (2021).
38. Tripathy-Lang, A. *Finding "Glocal" Solutions to Flooding Problems*, 102, (Eos, 2021).
39. Sexton C. *Glocal Approach is Needed to Mitigate Flood Damage* (Earth.com, 2021). <https://www.earth.com/news/glocal-approach-is-needed-to-mitigate-flood-damage/>.
40. Wu, H. *International Research Team Calls for 'Glocal' Approach to Help Mitigate Flooding Damage* (UNDRR PreventionWeb, 2020). <https://www.preventionweb.net/news/international-research-team-calls-glocal-approach-help-mitigate-flooding-damage>.
41. Ritter, A. & Muñoz-Carpena, R. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. *J. Hydrol.* **480**, 33–45 (2013).
42. Mariotti, A., Ruti, P. M. & Rixen, M. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Clim. Atmos. Sci.* **1**, 4 (2018).
43. Toth, Z. Estimation of atmospheric predictability by circulation analogs. *Mon. Weather Rev.* **119**, 65–72 (1991).
44. DelSole, T., Kumar, A. & Jha, B. Potential seasonal predictability: Comparison between empirical and dynamical model estimates. *Geophys. Res. Lett.* **40**, 3200–3206 (2013).
45. Dirmeyer, P. A. et al. Model estimates of land-driven predictability in a changing climate from CCSM4. *J. Clim.* **26**, 8495–8512 (2013).
46. Krishnamurthy, V. Predictability of weather and climate. *Earth Space Sci.* **6**, 1043–1056 (2019).
47. Nanding, N. et al. Uncertainty assessment of radar-raingauge merged rainfall estimates in river discharge simulations. *J. Hydrol.* **603**, 127093 (2021).
48. Jiang, L. et al. Combining Climate Forcings and Initial Hydrological Conditions for Enhanced Subseasonal-to-Seasonal Flood Forecasting. *J. Hydrometeorol.* <https://doi.org/10.1175/JHM-D-25-0015.1> (2025).
49. Koster, R. D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P. & Reichle, R. H. Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. *Nat. Geosci.* **3**, 613–616 (2010).
50. Xu, R. et al. Prediction of streamflow based on the long-term response of streamflow to climatic factors in the source region of the Yellow River. *J. Hydrol. Reg. Stud.* **52**, 101681 (2024).
51. Domínguez-Tuda, M. & Gutiérrez-Jurado, H. A. Global analysis of the hydrologic sensitivity to climate variability. *J. Hydrol.* **603**, 126720 (2021).
52. Gnann, S. et al. The influence of topography on the global terrestrial water cycle. *Rev. Geophys.* **63**, e2023RG000810 (2025).
53. Wu, J. et al. Assessing water quality in the Pearl River for the last decade based on clustering: characteristic, evolution and policy implications. *Water Res.* **244**, 120492 (2023).
54. Wu, Z., Milliman, J., Zhao, D., Zhou, J. & Yao, C. Recent geomorphic change in LingDing Bay, China, in response to Economic and Urban Growth on the Pearl River Delta, Southern China. *Glob. Planet. Change* **123**, 1–12 (2014).
55. Funk, C. et al. The climate hazards infrared precipitation with stations — a new environmental record for monitoring extremes. *Sci. Data* **2**, 150066 (2015).
56. Vitart, F. et al. The subseasonal to seasonal (S2S) prediction project database. *B. Am. Meteorol. Soc.* **98**, 163–173 (2017).
57. 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).
58. Liang, X., Lettenmaier, D. P., Wood, E. F. & Burges, S. J. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res. Atmos.* **99**, 14415–14428 (1994).
59. Wu, H. et al. A new global river network database for macroscale hydrologic modeling. *Water Resour. Res.* **48**, W09701 (2012).
60. Wu, H., Adler, R. F., Hong, Y., Tian, Y. & Policelli, F. Evaluation of global flood detection using satellite-based rainfall and a hydrologic model. *J. Hydrometeorol.* **13**, 1268–1284 (2012).
61. Wu, H. et al. Real-time global flood estimation using satellite-based precipitation and a coupled land surface and routing model. *Water Resour. Res.* **50**, 2693–2717 (2014).
62. Liang, X., Lettenmaier, D. P. & Wood, E. F. One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model. *J. Geophys. Res. Atmos.* **101**, 21403–21422 (1996).
63. Cherkauer, K. A., Bowling, L. C. & Lettenmaier, D. P. Variable infiltration capacity cold land process model updates. *Glob. Planet. Change* **38**, 151–159 (2003).

64. Yan, Y. et al. Climatology and interannual variability of floods during the TRMM era (1998–2013). *J. Clim.* **33**, 3289–3305 (2020).
 65. Lettenmaier, D. P. et al. Inroads of remote sensing into hydrologic science during the WRR era. *Water Resour. Res.* **51**, 7309–7342 (2015).
 66. Peters-Lidard, C. D. et al. 100 years of progress in hydrology. *Meteorol. Monogr.* **59**, 25.1–25.51 (2018).
 67. Grimaldi, S., Schumann, G. P., Shokri, A., Walker, J. P. & Pauwels, V. R. N. Challenges, opportunities, and pitfalls for global coupled hydrologic-hydraulic modeling of floods. *Water Resour. Res.* **55**, 5277–5300 (2019).
 68. Wu, H. et al. Evaluation of real-time global flood modeling with satellite surface inundation observations from SMAP. *Remote Sens. Environ.* **233**, 111360 (2019).
 69. Huang, Z. et al. Paired satellite and NWP precipitation for global flood forecasting. *J. Hydrometeorol.* **24**, 2191–2205 (2023).
 70. Huang, Z. et al. Multisourced flood inventories over the contiguous United States for actual and natural conditions. *B. Am. Meteorol. Soc.* **102**, E1133–E1149 (2021).
 71. Wang, Q. J., Robertson, D. E. & Chiew, F. H. S. A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.* **45**, W05407 (2009).
 72. Wang, Q. J. & Robertson, D. E. Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.* **47**, W02546 (2011).
 73. Robertson, D. E., Shrestha, D. L. & Wang, Q. J. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.* **17**, 3587–3603 (2013).
 74. Khan, M. Z. K., Sharma, A., Mehrotra, R., Schepen, A. & Wang, Q. J. Does improved SSTA prediction ensure better seasonal rainfall forecasts?. *Water Resour. Res.* **51**, 3370–3383 (2015).
 75. Shrestha, D. L., Robertson, D. E., Bennett, J. C. & Wang, Q. J. Improving precipitation forecasts by generating ensembles through postprocessing. *Mon. Weather Rev.* **143**, 3642–3663 (2015).
 76. Schepen, A., Zhao, T., Wang, Q. J., Zhou, S. & Feikema, P. Optimising seasonal streamflow forecast lead time for operational decision making in Australia. *Hydrol. Earth Syst. Sci.* **20**, 4117–4128 (2016).
 77. Wang, Q. J. et al. An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. *Environ. Model. Softw.* **122**, 104550 (2019).
 78. Schepen, A., Wang, Q. J. & Everingham, Y. Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Mon. Weather Rev.* **144**, 2421–2441 (2016).
 79. Gelman, A. et al. *Bayesian Data Analysis*, 3rd ed. (CRC Press, 2014).
 80. Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast* **15**, 559–570 (2000).
 81. Gneiting, T. & Katzfuss, M. Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1**, 125–151 (2014).
- and the Southern Marine Science and Engineering Laboratory (Zuhai) at Sun Yat-sen University for their scientific support. We are especially grateful to Prof. Wang for his valuable guidance on the application of the BJP model. We also thank the anonymous reviewers for their constructive comments and suggestions, which greatly improved the quality of this manuscript. This study was supported by the National Key R&D Program of China (Grants: 2024YFC3013302), the National Natural Science Foundation of China (Grants: 42275019, 42088101), CMA's Open Fund Project for Heavy Rain of China (Grants: BYKJ2024Z10) and also partially supported by Hainan R&D Program (CXFZ2022J074, SCSF202203) and Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies (Grants: 2020B1212060025).

Author contributions

L.L., H.W. conceived the study. L.L. carried out the analyses. L.L. wrote the draft manuscript. H.W. managed funding, and supervised the project. All authors, L.L., H.W., L.J., Y.M., J.S. K., L.A., Z.H., Y.H., S.C., S.D., Y.H. and W.W. discussed the results and edited the paper at multiple stages.

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/s41612-025-01273-9>.

Correspondence and requests for materials should be addressed to Huan Wu.

Reprints and permissions information is available at <http://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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2025

Acknowledgements

The authors acknowledge the School of Atmospheric Sciences, Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies,