communications earth & environment

Article

A Nature Portfolio journal

https://doi.org/10.1038/s43247-025-02324-y

Hybrid approaches enhance hydrological model usability for local streamflow prediction

Yiheng Du & Ilias G. Pechlivanidis

Hydrological models are essential for predicting water flux dynamics, including extremes, and managing water resources, yet traditional process-based large-scale models often struggle with accuracy and process understanding due to their inability to represent complex, non-linear hydrometeorological processes, limiting their effectiveness in local conditions. Here we explore hybrid methods combining process-based modelling and statistical or machine learning post-processors to improve streamflow predictive accuracy, including extremes, across Europe's hydro-climatic gradient. We investigate various post-processing methods, such as random forest, long short-term memory model, quantile mapping and generalised linear model, demonstrating notable improvements in model performance, in terms of reducing errors in total volume and extremes and increasing robustness across diverse climatic and geographic conditions. We further show that hydrologic similarity is one of the key drivers that control the hybrid approach's improvements, together with other basin characteristics, such as mean precipitation and mean temperature. Our results also reveal spatial complementarity among the post-processing methods, with no absolute superiority identified from a single method, pointing towards multi-model averaging approaches for the future evolution of hybrid hydrological modelling.

Hydrological modelling has advanced the understanding of the water cycle by simulating the movement, dynamics and quality of water, allowing scientists and policymakers to monitor and predict complex hydrological processes and their interactions with climatic and environmental factors[1–3]. Large-scale hydrological models (LSHM) are applied to provide valuable insights into complex transboundary river systems that are difficult to directly monitor and describe the river system functions and responses to different inputs and environmental factors[4,5]. However, LSHMs, especially at national, continental or global levels, face considerable challenges when applied to local scales, referring to locations within the river system which are critical for water management and decision-making[6,7]. One of the primary issues is the inherent uncertainties and errors in model setup and parameter identification, leading to poor performance and incomplete or even misinformed understanding of the fluxes[8,9]. Strong hydro-climatic gradients across the large domain, driven by varying climate conditions, topography, and anthropogenic influences like irrigation and reservoir regulation, has introduced additional challenges. Moreover, the lack of sufficient gauging in river systems, particularly in remote areas, further complicates LSHM setups and parameterisations, which traditionally depend on long streamflow time series[10,11]. Additionally, the lack of a "perfect" meteorological dataset poses another barrier, with no global product accurately capturing the meteorological dynamics at the local scale, particularly for precipitation[12–14]. Other hydrometeorological fluxes, such as evapotranspiration, groundwater recharge and soil moisture, remain critical in closing the water and energy balance of the river systems, yet poorly quantified in the water cycle[15,16]. These challenges collectively highlight the need for beyond state-of-the-art frameworks to enhance the regional applicability of LSHMs, especially in the context of varying environmental and climatic conditions.

Post-processing in hydrological and meteorological modelling has proved capability for enhancing local performance and process representation. The refinement of model outputs to better represent the observations improves the model reliability and applicability for local decision making[17–20]. Among the various techniques, statistical and machine learning (ML) methods have been increasingly recognized for their potential to tailor hydrological model outputs. Statistical methods (i.e., quantile mapping) are commonly employed to bias adjust and downscale model outputs[21,22]. Meanwhile, ML-based methods (i.e., neural networks, decision trees, ensemble learning) are particularly capable at handling large and diverse datasets and extracting meaningful patterns, and have emerged as powerful

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden. e-mail: yiheng.du@smhi.se; ilias.pechlivanidis@smhi.se

tools for capturing complex, nonlinear relationships within data, allowing more advanced prediction capabilities[17,23,24]. Both statistical and ML-based approaches are capable of reducing uncertainties and increasing accuracy, and therefore setting a pathway for more reliable local applications of hydrological models[25].

The misuse of post-processing and their non-explainability through, for instance, overtrained parameterisation or black box modelling, induces the lack of interpretability and transparency, which makes the understanding of the underlying processes less clear, potentially limiting the ability to trust the results in decision-making scenarios[26–28]. Conventional post-processing methods, which primarily depend on mathematical algorithms, frequently fail to account for the key influences of topography, soil type, vegetation, and regional climate patterns—factors closely associated with the dynamics of river systems[29–31]. Understanding these physiographic characteristics in the post-processing context, will not only ensure that the post-processing techniques indeed improve the model performance, but also provide insights on the underlying processes, indicating how they bridge the gap between generic model outputs and the varied local conditions they aim to represent.

Here, we enhance the quality of streamflow simulations from LSHM at the local scale across the pan-European domain, and improve the understanding of model enhancement to allow for more reliable applications for local decision-making, by answering the following scientific questions: (1) How do hybrid process-based and statistical/ML methods enhance local model performance across various streamflow characteristics? (2) How does the performance of different post-processing methods vary across Europe's hydro-climatic gradient? and (3) What are the key drivers controlling the hybrid model performance enhancement across Europe? To address these questions, we establish a hybrid framework (Fig. 1a) for post-processing the outputs from the E-HYPE process-based LSHM across the entire European domain. This framework employs two statistical methods (Generalised Linear Model, GLM; Quantile Mapping, QM; Methods section) and two ML-based methods (Random Forest, RF; Long Short-Term Memory model, LSTM; Methods section), with comprehensive evaluation metrics and performance attribution, allowing a thorough assessment and process understanding. Our analysis, covering over 2000 gauging stations across a wide range of hydrological regimes (Fig. 1b), shows that the two ML methods yield higher improvements than the statistical methods, particularly in capturing extreme streamflow characteristics. However, no single method consistently outperforms across the entire domain, rather a spatial complementarity occurs, which is primarily influenced by catchment characteristics, including hydrological regimes (represented by predefined hydrological clusters[30,32], details in Methods, Supplementary Fig. 1 and Supplementary Table 1) and climate conditions among the investigated potential drivers. From an operational perspective linked to either early warning systems or climate services, this effort strongly enhances LSHM usability for streamflow predictions at local conditions and carry critical implications for water-dependent sectors (e.g., agriculture, hydropower, drinking water etc.).

Results

Hybrid modelling improves representation of streamflow characteristics at local scale

We applied four post-processing methods, including both classical statistical and state-of-art ML methods, to correct streamflow simulations across the pan-European region. Our evaluation, focusing on total volume as well as high and low flow extremes, reveals that integrating any of these methods yields substantial improvement in the performance of the underlying process-based E-HYPE model (Fig. 2), while also revealing notable differences in their effectiveness. All methods perform almost equally across all performance groups with regard to total volume; however the differences between the methods are more apparent for high and low extremes with the ML methods achieving better performance than the statistical methods at all groups below 0.5 (below the *fair* performance group defined in Fig. 3, the same for other groups presented in italic), especially at the group below 0 (*very poor* and *unsatisfactory* groups), where QM gives relatively the lowest performance.

We next investigate the changes in the performance groups between the stations before and after post-processing and identify stations where the highest/lowest improvements are achieved (Fig. 3). Notably, our analysis reveals that the two ML methods not only increase the number of stations achieving very good and good performance but also yield larger improvement jumps across performance groups compared to statistical methods. LSTM and RF are particularly effective at some stations enhancing performance from an initial fair performance to a very good group, whereas the statistical approaches mainly enhance performance within the fair-to-good range. This confirms that ML methods can compensate for model structural errors (e.g., due to anthropogenic interventions) which are challenging to represent, while the statistical methods mainly account for uncertainties in forcing inputs and model parameters[33,34].

Overall, RF performs similarly to LSTM with their performance having small differences over the *very good* and *good* groups with regard to high and

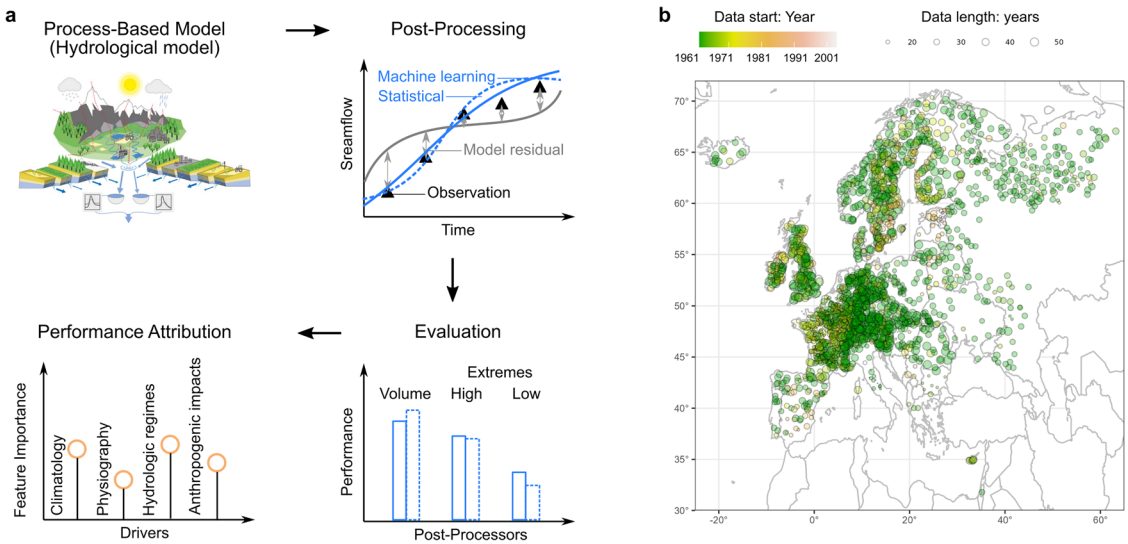

Fig. 1 | Hybrid framework for post-processing process-based LSHM and data availability of the observations. a Schematic of the hybrid framework, detailing the process-based model, post-processing, evaluation, and performance attribution steps. **b** Spatial distribution of the stations used in the study, annotated with the start year and duration of the observational data.

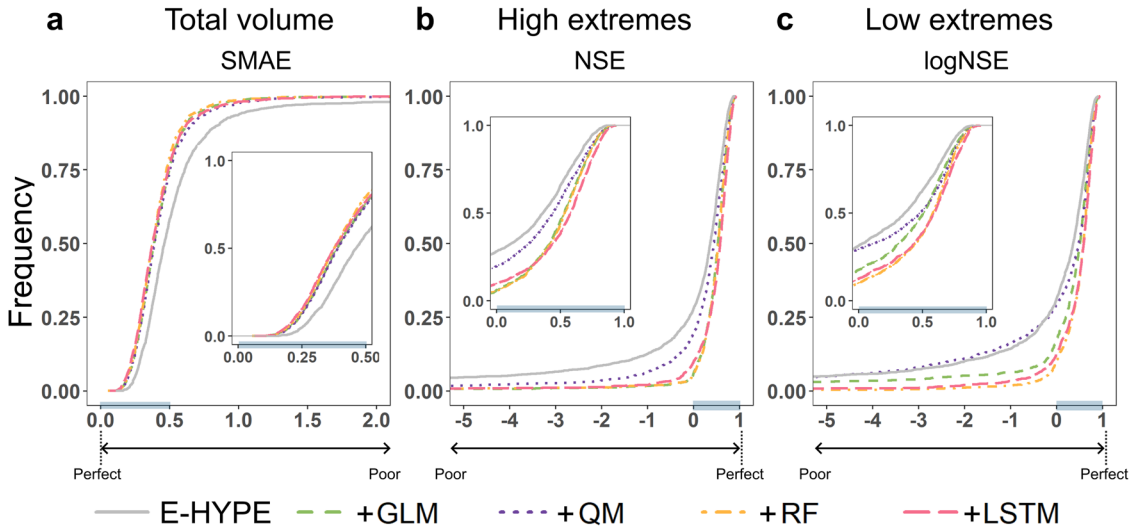

Fig. 2 | Performance comparison of process-based (E-HYPE) and hybrid models (E-HYPE integrated with GLM, QM, RF and LSTM) in predicting streamflow total volume (SMAE), high extremes (NSE) and low extremes (logNSE). The cumulative distribution of model performance is shown using the SMAE (**a**), NSE (**b**), and logNSE (**c**) metrics (see Methods). Perfect performance corresponds to 0 for SMAE and 1 for NSE and logNSE. The grey line represents E-HYPE, while colored lines with varying styles denote hybrid models with different post-processing methods. Performance improves as the lines approach the perfect value marker on the *x*-axis. The *x*-axis represents the metric values, and the *y*-axis indicates the proportion of stations with performance not exceeding the corresponding metric level. The inset plot provides a zoomed-in view of the most common range (highlighted on the *x*-axis) for clarity.

low extremes. LSTM is designed to handle sequential data and complex, nonlinear relationships, making the model adept at capturing temporal dependencies. This is crucial in hydrological modelling, where past conditions significantly influence future events. RF, on the other hand, is more suited to capturing complex, non-linear relationships between features without assuming temporal dependencies, which could explain the differences between the two ML methods. In addition, GLM typically does not account for such sequential dependencies, as its linear assumptions consequently do not capture nonlinear dynamics effectively. QM can overall improve the total volume, yet the method can lead to performance deterioration for extremes in certain catchments (Fig. 3). Notably, for the low streamflow extremes, the performance at approximately 2% of stations deteriorates and ends up in the *unsatisfactory* category, which leads to an "unexpected" expansion of the *unsatisfactory* group after the QM post-processing. Although QM has been widely applied in hydro-meteorological time series, the method mainly adjusts the statistical variability of the data and consequently the volume. Whilst QM does not show sensitivity to temporal dynamics, which is the reason for occasionally deteriorating performance in extremes which are time sensitive.

No single best hybrid method: spatial complementarity of post-processing potential

Building on the overall improvement achieved by the hybrid modelling framework, we now assess its spatial effectiveness, examining how different post-processing methods perform across regions and whether a universally applicable method exists. The added value (skill) achieved by the post-processing over the process-based LSHM is provided for each station (Fig. 4b). A consistent pattern emerges across all hybrid methods, with post-processing achieving higher skills in central, southern and eastern Europe, over which raw E-HYPE performance for high streamflow extremes is considered at least poor, in comparison to the other regions (Fig. 4a). Similar patterns with high skill values for both total volume and low extremes further confirms the overall capability of the hybrid modelling framework; however, spatial variations of skills across the post-processing methods are also evident. For instance, in the United Kingdom, only a small improvement is achieved from the two statistical post-processing methods (Fig. 4b), while the two ML methods result in considerable skills, especially LSTM. This can be attributed to LSTM's superior capability in detecting complex and nonlinear relationships within the dataset (e.g., driven by the chalk streams and the river-aquifer interactions), which is less strong in the QM method.

We next identify the best performing model based on the highest skill achieved at each station (Fig. 4c), and conclude the methods' spatial complementarity. The majority of the stations are mostly improved by LSTM, with over 50% demonstrating this in terms of high streamflow extremes, and about 20% for total volume and low streamflow extremes, with the river systems mainly located in central and western Europe. RF excels the other methods at approximately 20% of the stations, making it the second most effective post-processing method overall, especially in northern Europe and along the Mediterranean coastlines. Both statistical methods (GLM and QM) show their superiority at various stations within the domain, with performance varying according to different evaluation metrics. QM excels particularly in handling low extremes in the region west of the Urals Mountains in the Russian Federation, where the streamflow is mainly snow dominated. Despite these performance improvements, in few river systems, no hybrid method adds value. These stations are spread across the European domain with the river systems being characterised by small upstream area (mostly less than 500 km^2) and quick response to rainfall input based on analysis of their hydrological regime (Supplementary Fig. 2 and Supplementary Table 1), even if in some of them baseflow is a strong contributor, and hence small changes in the streamflow dynamics can affect the model performance.

Overall, the analysis suggests that there is no universally superior hybrid model, with each incorporated post-processing method presenting varying degrees of skill across different spatial locations and under different streamflow properties (total volume and extremes). This could be addressed with a model averaging approach, for instance, based on Bayesian concepts[35,36] and/or copula-based frameworks[37,38], allowing for integrating multiple models, deriving advantages of each model and compensating for the individual limitations. The spatial variability of best performing models also highlights the importance of diagnostically selecting appropriate post-processing methods accounting specific local characteristics and particular signatures of river system behaviour. This sets the need for explainable hybrid frameworks by investigating the key factors of post-processing improvements.

Hydrological regime as a key driver to model performance enhancement

We next introduce different potentially key factors with regard to climatology, physiography, hydrological similarity and anthropogenic impact,

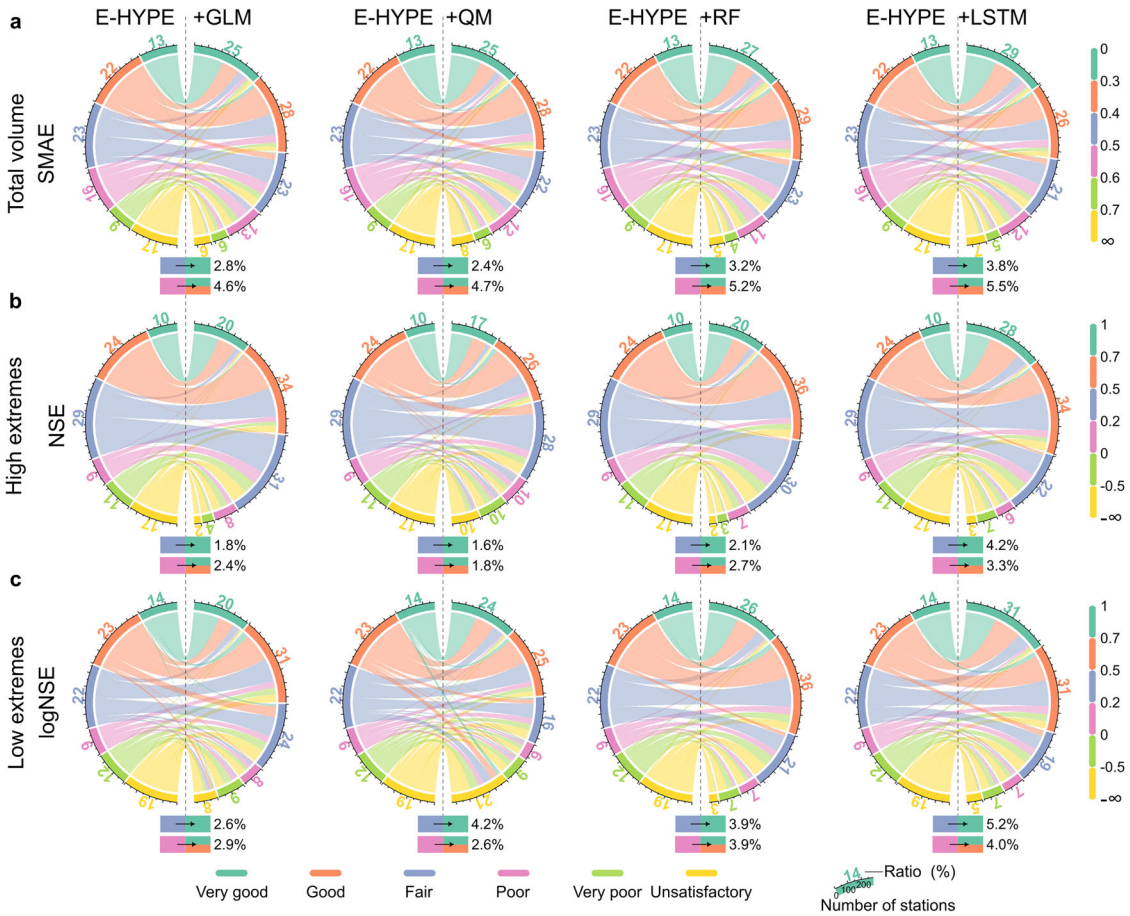

Fig. 3 | Chord diagram showing performance transitions before and after post-processing. The performance for process-based (E-HYPE, left side of the chord) and hybrid models (E-HYPE integrated with GLM, QM, RF and LSTM, right side of the chord) in predicting streamflow total volume (SMAE; **a**), high extremes (NSE; **b**), and low extremes (logNSE; **c**) are presented. The diagram visualizes how stations transition across six performance groups. The width of each chord represents the proportion of stations shifting between performance groups, highlighting improvements or deteriorations due to post-processing. Portions of stations that experienced performance jumps (i.e., from fair to very good, and from poor to good/very good), are displayed below each chord diagram.

and filter them by removing interdependency (Methods; Fig. 5a). Hydrological similarity has been widely considered for model parameterisation and regionalisation, yet its impact in hybrid modelling is not sufficiently explored. Here we use predefined clusters of hydrologically similar regimes (Supplementary Table 1 and Supplementary Fig. 1) across Europe based on a set of hydrological signatures[30]. The Classification and Regression Tree (CART) method provides the feature importance each of these factors has on model performance, including both the process-based and hybrid models. The overall importance is further summarised by the comprehensive ranking index (RI) across the models (Methods).

The same dominant factors are identified with regard to total volume (SMAE; Fig. 5b) and high streamflow extremes (NSE; Fig. 5c), the leading factors are the hydrological similarity represented by hydrological clusters, and climatic conditions represented by mean precipitation and mean temperature as shown by the higher feature importance and ranking index (Cluster, Prec and Temp; Fig. 5). In addition, we show how skill changes as a function of its influencing drivers, using as an example the skill of LSTM for high streamflow extremes (Fig. 5c). LSTM enhances model performance, with higher skills in drier and warmer conditions; the skill increases with increased mean temperature and decreases with increased mean precipitation. Moreover, the degree of improvement varies across the hydrological clusters, as indicated by the differences in the distribution shape and median values of the skills. Similar patterns are observed for the other post-processing methods and evaluation metrics, as shown in the Appendix (Supplementary Fig. 3). For low streamflow extremes (logNSE; Fig. 5d), different dominant factors arise, including the hydrological cluster, elevation, and dryness index. In particular, the hydrological cluster ranks among the most influential drivers, reflecting that low streamflows are much less influenced by precipitation and are instead strongly correlated to river systems' memory, which is well represented by the hydrological signatures of the clusters[30,39].

The observed patterns between model skill and key driving factors remain consistent across different post-processing methods (Fig. 5 and Supplementary Fig. 3), suggesting that the skill improvements are prone to these factors, rather than the choice of post-processing method alone. While different methods may vary in their capacity to enhance model performance, their responses to underlying hydrological and climatic controls are similar, highlighting the importance of considering these factors when selecting or explaining results from post-processing methods. This conclusion also offers insights into future frameworks focusing on optimising hybrid hydrological model performance in both gauged and ungauged conditions. Similar to parameter regionalisation in hydrological modelling[40], here the revealing of strong influence from local basin characteristics suggests that post-processing methods may also be effectively adapted across different river systems by considering their hydro-climatic similarity.

Discussion

Here we established a strong connection between hydrological regimes and the effectiveness of model enhancement through post-processing. This extends the current knowledge about the influence of hydrological similarity in streamflow simulation and regionalisation[41,42] into a post-processing

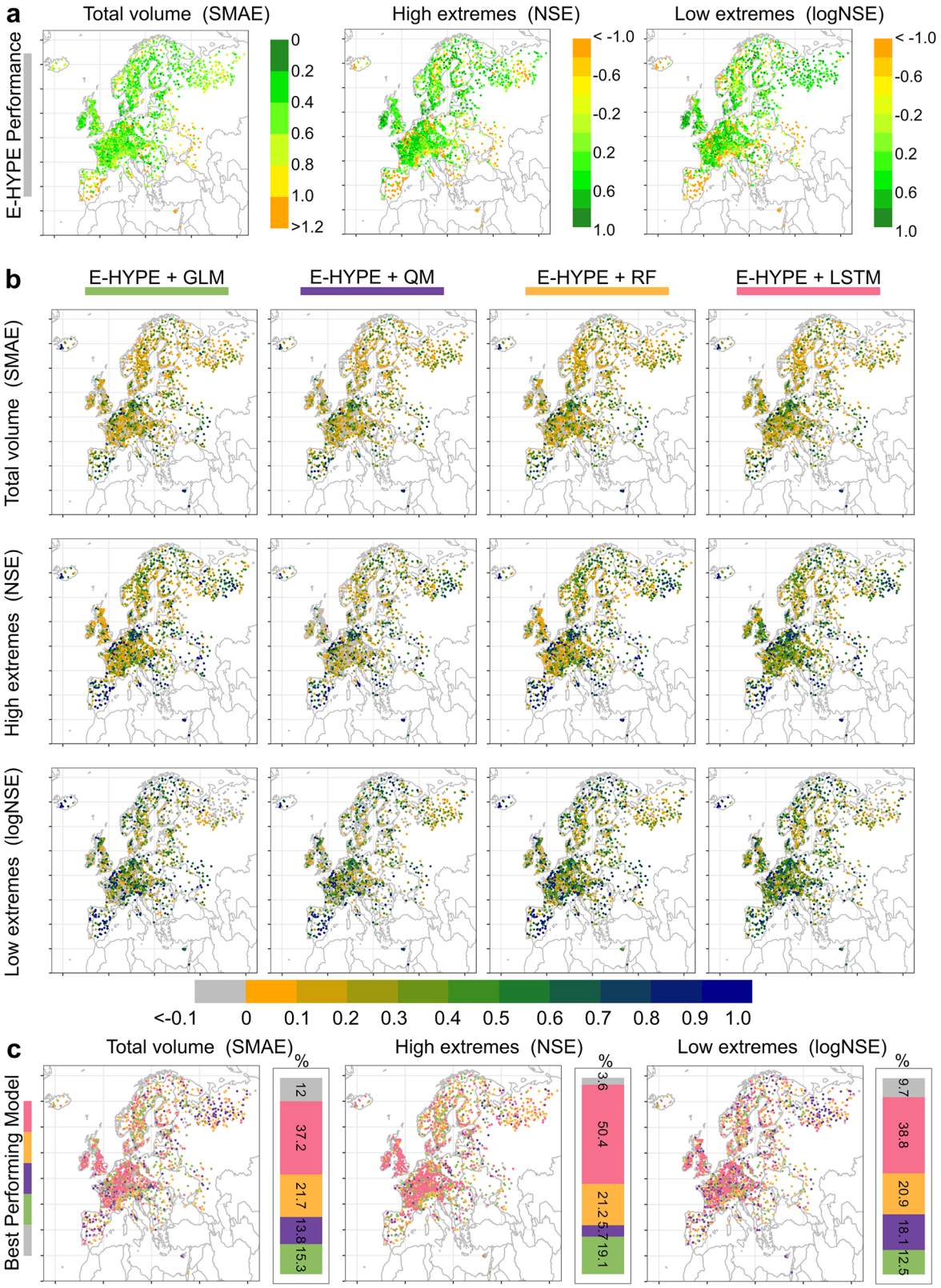

Fig. 4 | Spatial distribution of E-HYPE performance, post-processing skill, and best-performing models across different streamflow characteristics. a Raw performance of the process-based E-HYPE model. **b** Skill improvement from each post-processing method, with colors ranging from yellow to blue indicating performance improvements, while grey represents no improvement after post-processing. **c** Best-performing post-processing method at each station and the proportion of stations where each method achieves the highest skill. Colors represent different models: process-based E-HYPE (grey), and hybrid using GLM (green), QM (purple), RF (orange) and LSTM (pink).

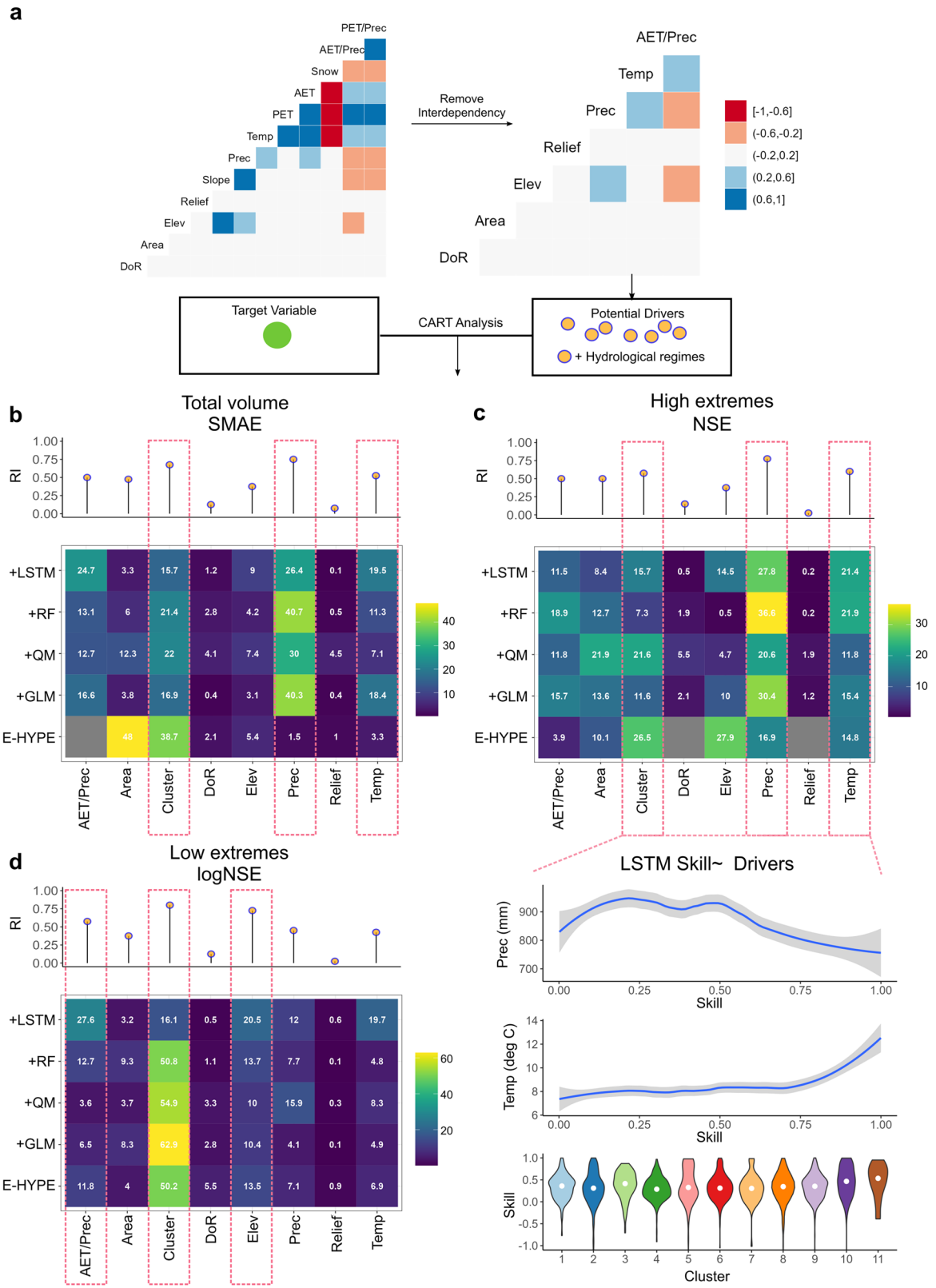

Fig. 5 | Drivers influencing model performance enhancement. a Potential drivers included and filtered by interdependency (see Table 2 for acronyms). **b–d** Feature importance from CART analysis for the process-based and hybrid models for the total volume (SMAE; **b**), high extremes (NSE; **c**), and low conditions (log NSE; **d**). An example of how skill changes as a function of its influencing drivers is also presented (**c**), showing the skill of E-HYPE-LSTM for high streamflow extremes.

context[31], where the correction of process-based LSHM outputs at the local scale, whether analysed in terms of total volume, high or low streamflow extremes, is influenced by basins' hydrological characteristics. Whilst it is important to note that this connection persists regardless of the post-processing methods employed. Therefore, this new insight provides support for the regionalisation of the post-processors, which is important for the broader application of hydrological models[29,30,43]. A way forward would be the establishment of a multi-basin post-processing approach building on the current method by integrating data from multiple river systems to train a single, regionalized model. While the current framework trains individual models at each gauged basin, the multi-basin approach can incorporate basin characteristics, such as climatic conditions, physiographic attributes, and hydrological regimes, as static input features, enabling the post-processing method to capture shared hydrological behaviours across basins and allowing it to generalize beyond the training locations. Consequently, basins under ungauged or data sparse conditions can benefit from the learned patterns in hydrologically similar gauged systems. Furthermore, this approach lays a solid scientific foundation for expanding the insights gained from pilot studies to broader applications, particularly in vulnerable areas with limited resources. This aligns with the vision that using large-scale hydrological services to identify solutions at the local scale is essential for maximising global impact[44,45].

Another key outcome of our study is the potential to produce more accurate forecasts in an operational setting. Hydrological forecast predictability relies on two primary components: the initialization of hydrological conditions at the onset of a forecast, and the hydrological model forcing with (bias-adjusted) meteorological forecasts[30,43]. However, the biases in these two components are inherited from the reference simulation and the quality of the meteorological forecasts which deteriorates as a function of lead time[46,47]. These limitations can both be addressed by training the post-processing models using reforecasts and the corresponding observations for each lead time[48], and consequently providing lead time-specific correction factors. These trained models are then applied to new forecasts, improving accuracy across the entire forecast horizon. Overall, such investigations support global scientific and operational efforts to ensure equitable access to reliable hydrological data, information and services, including the HELPING scientific decade launched by the International Association of Hydrological Sciences[49] and the Early Warnings for All (EW4ALL) initiative launched by the United Nations[50]. Both global community efforts aim to protect everyone from hydrological hazards (floods and droughts), and this achievement relies on accurate hydrological forecasts as a foundation for action, calling for the operationalization of model enhancement efforts.

Methods

Hybrid modelling framework and data

Our hybrid hydrological modelling framework (Fig. 1) combines the output from the process-based continental E-HYPE model with post-processing methods and adjusts the simulated streamflow to better align with local observations by capturing complex patterns or discrepancies that the process-based LSHM alone may not account for. This hybrid approach leverages the strengths of both process-based modelling (understanding natural processes) and data-driven techniques (capturing complex, site-specific patterns), with the aim to result in improved model performance.

The hybrid models are benchmarked against the E-HYPE hydrological model which is driven by meteorological forcing, i.e., temperature and precipitation, to produce streamflow simulations across the pan-European domain. E-HYPE is a semi-distributed process-based LSHM of water quantity and quality based on the HYPE (The HYdrological Predictions for the Environment) model structure. The pan-European setup simulates components of the water cycle at daily time steps, i.e., snow accumulation and melting, evapotranspiration, soil moisture, streamflow generation, groundwater recharge, and routing through rivers and lakes. The historical model performance in terms of streamflow reaches a median Nash-Sutcliffe Efficiency (NSE) of 0.53 over more than 500 streamflow stations across Europe[51,52].

Simulated streamflow ($m^3\ s^{-1}$) was obtained for the period 1961–2023 by forcing E-HYPE with the HydroGFD v3.2 meteorological reanalysis data[53]. Streamflow observations were collected in the pan-European domain from various data sources, including Global Runoff Data Centre, European Water Archive, and national authorities, reaching 2072 stations[51]. To ensure the sufficiency of training samples, we selected only the stations with at least 10 years of observations (Fig. 1). The final dataset shows a comprehensive spatial coverage of the stations across the entire European domain, with a higher concentration in central Europe and relatively fewer stations in the southern (e.g., Spain) and the eastern part of the continent.

Post-processing method description

In total four methods were used to post-process the E-HYPE LSHM output to better align with local observations at each individual station; two statistical (Generalised Linear Model and Quantile Mapping) and two ML-based (Random Forest and Long Short-Term Memory), which are briefly described below. The models are implemented using the R packages randomForest and qmap, along with the Python package TensorFlow. Details on data processing and model training are provided in the code availability section, ensuring reproducibility and transparency.

Generalised Linear Model (GLM): a statistical technique that extends linear regression to allow for non-normal distributions of error terms[54]. It allows the inclusion of different types of predictor variables and the modelling of response variables that follow non-normal distributions, such as Gaussian, to provide a flexible framework for understanding the relationships between variables.

Quantile Mapping (QM): a statistical technique used for bias correction by adjusting the distribution of the variable of interest (here simulated streamflow) to match the target variable (here observed streamflow) distribution, in order to correct systematic biases in model outputs[55]. The tricubic spline method is adopted here to allow for a smooth adjustment of the cumulative distribution functions, to improve the biases in the tails of the distribution.

Random Forest (RF): a supervised, non-parametric method, where an ensemble of uncorrelated trees yields prediction for classification or regression. Multiple trees are built based on bootstrapping samples from the training data. After all the trees are grown, the forests produce the final results by averaging predictions from the trees[56]. The same model configuration, regarding maximum node numbers (10) and minimum node size, is maintained across all stations, to ensure comparability throughout the study domain, allowing the analysis of potential influencing factors.

Long Short-Term Memory (LSTM): a model for time series, which is capable of learning long-term dependencies[57]. For post-processing purposes, previous research has proved that the lookback length can be reduced as model performance remains reliably consistent across diverse temporal scales[58]. Our designed lookback length for LSTM in the hybrid framework is 3-day, as the seasonal dynamics are already represented in the process-based model, which also confirmed its capability of capturing temporal dependencies present in streamflow data by initially experimenting values between 1 to 215 lookback days. This model is structured with three layers containing different numbers of cells (i.e., 100-50-20), allowing an effective process and remembering information over extended periods.

To prevent overfitting, a portion of the training set (10%) is reserved for validation, while the model training includes a monitoring mechanism where if the validation loss does not decrease over 10 consecutive steps, an early stopping criterion is triggered. Normalisation is applied to the input data to scale the range of data points, allowing smoother training process and more stable convergence. To address data imbalances, particularly concerning extreme values critical for hydrological services, a sample weight technique is implemented. This method assigns weights to samples, emphasising the importance of accurately predicting extreme events, which are often underrepresented in the dataset but hold importance for hydrological analyses and applications. Weights are calculated based on percentiles in the observations, where the 10th, 33rd, 66th and 90th percentiles divide the samples into five groups, representing low extremes, lower than

normal, normal, higher than normal, and high extremes. Samples within each group share a total weight of 0.2. The root mean square error is used as the loss function for the LSTM model during optimisation.

The post-processing models take simulated streamflow from the E-HYPE hydrological model as input, with the target variable representing either the observed streamflow (Eq. 1) or the relative residual between observed and simulated values (Eq. 2). Both observed streamflow and relative residuals were tested as target variables across the methods. An exception is the QM method, which exclusively uses observed streamflow as the target.

$$target_{obs} = y_{obs} \tag{1}$$

$$target_{residual=}(y_{obs} - y_{sim})/(y_{sim} + \varepsilon) \tag{2}$$

where ε is a small constant value introduced to prevent division by zero, particularly in scenarios of low streamflow, ensuring the target variable remains within a reasonable range. By setting the target thresholds to be no smaller than -1, this approach also effectively mitigates the common issue of generating negative streamflow values when using residuals as the target variable.

In the Results section, the target variable yielding the highest performance for each method is selected and presented (Supplementary Table 2, calculation of the metrics can be found in Table 1), ensuring that the analysis highlights the most effective implementation of each approach.

Each station is corrected independently, with separate model calibration for each location. For model training, the dataset was subsequently divided into training and testing periods, by applying an 80–20% data split. The model evaluation is conducted on the testing periods.

Overall, the hypotheses for this experiment include using the identical model structure for all stations, e.g., the same number of layers, cells, and hyperparameters, without individual optimization for each station. Nevertheless, this generalised approach enables us to compare and provide an overall assessment for the methods across the domain, which well aligns with the objective of this study.

Model evaluation

To evaluate the added value from post-processing, three evaluation metrics were used to assess the potential improvements with regard to errors in total volume, high and low streamflow extremes (Table 1), as represented by the Mean Absolute Error[59] (MAE), NSE[60] and its logarithmic form[61] (logNSE), respectively. In particular, the Scaled Mean Absolute Error (SMAE) is applied to adjust MAE in relation to the average streamflow observed at each station, thus allowing the comparison of MAE values across stations that have varying streamflow magnitudes.

Improvement at each station is further denoted by calculating the skill, which quantifies the efficacy of post-processing methods (*pp*, Eq. 3) relative to raw E-HYPE simulations (*ref*, Eq. 3), with positive (negative) skill values indicating improvements (deterioration). A skill value approaching 1 signifies a greater enhancement in predictive performance, highlighting the effectiveness of the post-processing techniques in refining hydrological simulations. The skill (over the historical simulation period) is expressed as:

$$Skill = \frac{Score_{pp} - Score_{ref}}{Score_{perfect} - Score_{ref}} \tag{3}$$

The cumulative distribution plot (Fig. 2) presents the proportion of stations that fall below a given performance threshold for the three evaluation metrics. This allows for an inter-comparison between the different post-processing methods.

The chord diagram (Fig. 3) provides a detailed comparison by tracking the transitions of stations between performance groups before and after post-processing. By depicting the "flow" of stations from one group to another, this visualisation helps clarify the extent to which post-processing methods improve, degrade, or maintain performance across different stations. The groups are determined subjectively but still driven by expert knowledge from previous analyses[62]. However, we note that these are not universally applicable[63,64] and are determined specifically for this study.

To further evaluate model performance across stations, we analyse the spatial distribution of skill by identifying the best-performing method at each station (Fig. 4). The results are visualised using a color-coded map, where each station is assigned a color based on the method that yields the highest performance. Additionally, we calculate the proportion of stations where each method performs best (Fig. 4). These ratios are presented in a bar plot alongside the map, offering a comprehensive view of how different methods perform across the entire study area. This combined visualisation helps highlight spatial patterns in model performance and provides insights into the effectiveness of different post-processing methods.

Attributing hybrid model enhancement to hydrological processes

The CARTs method is used to identify the most important drivers of model performance and to explain the complex, non-linear relationships between them[65]. The algorithm splits the data into subsets based on the values of the input features that result in the largest reduction in heterogeneity of the target variable (i.e., model performance). This process continues until further splitting does not significantly improve the algorithm's accuracy or until predefined stopping criteria are met, such as a minimum number of leaf nodes. To avoid overfitting, the technique of pruning is used by removing branches that have little to no contribution to the algorithm's predictive power, aiming to find the optimal balance between the tree's complexity and its accuracy.

The drivers' importance is calculated by summing changes in the probability of splitting on every driver and dividing the sum by the number of branch nodes[30]. This importance score is then standardised, spanning from 0 to 100 for comparability. The association between hydrological model performance and potential drivers is investigated by calculating the feature importance of each potential driver (Table 2). We note that some drivers are highly interdependent and could therefore introduce uncertainty

Table 1 | The evaluation metrics used to quantify the potential model performance improvements for different characteristics of the streamflow time series

Characteristic of the streamflow signal	Evaluation metric	Abbreviation	Equation
Total volume	Scaled mean absolute error	SMAE	$MAE = \frac{\sum_{t=1}^{T}\lvert y_o^t - y_m^t\rvert}{T}$ $SMAE = \frac{MAE}{\bar{y}_o}$
High streamflow extreme	Nash-sutcliffe efficiency	NSE	$NSE = 1 - \frac{\sum_{t=1}^{T}(y_o^t - y_m^t)^2}{\sum_{t=1}^{T}(y_o^t - y_0)^2}$
Low streamflow extreme	Logarithmic nash-sutcliffe efficiency	logNSE	$\log NSE = 1 - \frac{\sum_{t=1}^{T}(\log(y_o^t) - \log(y_m^t))^2}{\sum_{t=1}^{T}(\log(y_o^t) - \log(\bar{y_0}))^2}$

y_o^t and y_m^t denotes the observation and model simulation at each timestep t, respectively, where t ranges from 1 to T.

Table 2 | Drivers considered to influence model performance, including topography, climate, anthropogenic impact and hydrological regimes

Name	Abbreviation	Unit	Selected / Replaced by (->)
Precipitation	Prec	mm	✓
Temperature	Temp	°C	✓
Snow depth	Snow	cm	-> Temperature, Precipitation
Actual evapotranspiration	AET	mm	-> Temperature
Potential evapotranspiration	PET	mm	-> Temperature
Dryness index	PET/Prec	–	-> Temperature
Evaporative index	AET/Prec	–	✓
Upstream area	Area	km^2	✓
Elevation	Elev	m	✓
Relief ratio	Relief	–	✓
Slope	Slope	%	-> Precipitation
Degree of regulation	DoR	%	✓
Hydrological clusters	Cluster	–	✓

The column "Selected / Replaced by" denotes if the corresponding variable is selected and kept after removing the interdependency (✓), or replaced by other highly correlated variables (->).

to the CART analysis. Therefore, the highly interdependent drivers (Pearson correlation coefficient greater than 0.6) are removed, and finally 8 potential drivers are kept for the CART analysis.

Following the concept of feature importance, a comprehensive ranking index[66] is used to enable the evaluation and comparison of potential drivers' influence across the process-based and hybrid models. The ranking index (RI) is mathematically expressed as:

$$RI = 1 - \frac{1}{nm}\sum_{i=1}^{n} rank_i \quad (4)$$

where m represents the total number of potential drivers, which here is 8, and n denotes the number of models, here set at 5 (the process-based model and four hybrid models). $rank_i$ indicates the assigned rank of each potential driver, with 1 being the most critical and 8 the least. Thus, an RI value approaching 1 indicates a more accurate and effective simulation outcome.

With RI, the analysis identifies the three most influential drivers across both the process-based model and the different hybrid models. This approach can reveal the underlying drivers of the model performance and provide information on where post-processing methods can significantly refine the model's accuracy.

To assess feature importance, we present results as heatmaps (Fig. 5), where color intensity represents feature importance on a scale from 0 to 100%. The corresponding rankings are visualised as points in the marginal plots, providing a clear comparison of relative feature contributions across the models.

In analysing model skill as a function of key driving factors (Fig. 5 and Supplementary Fig. 3), we illustrate trends using a locally estimated scatterplot smoothing curve (LOESS). This approach captures the general pattern of how model skill varies with numerical driving factors (e.g., temperature or precipitation), providing insight into the underlying relationships. For categorical driving factors as hydrological clusters, model skill is decomposed by cluster group and visualized using violin plots. These plots illustrate the distribution of skill within each cluster, highlighting variations across hydrological regimes and emphasising the role of river system characteristics in shaping model performance.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We performed the investigations at the SMHI Hydrology Research unit, where work benefits from joint efforts in developing models and concepts by the whole team. The HYPE model code is available from the HYPEweb portal (https://hypeweb.smhi.se/model-water/). Streamflow data, including E-HYPE model simulations and observations, can be shared upon request, following the SMHI data-sharing policy. Evaluation metrics and model skill assessments can be accessed in the repository on Zenodo [https://zenodo.org/records/14938526].

Code availability

All plots and post-processing models were generated using the R programming language. The code, including scripts for data processing, model training and evaluation, and visualization, can be found in the Zenodo repository [https://zenodo.org/records/14938526].

Received: 10 October 2024; Accepted: 22 April 2025;
Published online: 30 April 2025

References

1. Koutsoyiannis, D. Hydrology and change. *Hydrol. Sci. J.* **58**, 1177–1197 (2013).
2. Liu, Y., Gupta, H., Springer, E. & Wagener, T. Linking science with environmental decision making: experiences from an integrated modeling approach to supporting sustainable water resources management. *Environ. Model. Softw.* **23**, 846–858 (2008).
3. Montanari, A. et al. "Panta Rhei—Everything flows": change in hydrology and society—The IAHS scientific decade 2013–2022. *Hydrol. Sci. J.* **58**, 1256–1275 (2013).
4. Du, T. L. T. et al. Streamflow prediction in highly regulated, transboundary watersheds using multi-basin modeling and remote sensing imagery. *Water Resour. Res.* **58**, e2021WR031191 (2022).
5. Kumar, R., Samaniego, L. & Attinger, S. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resour. Res.* **49**, 360–379 (2013).
6. Pechlivanidis, I. G. & Arheimer, B. Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case. *Hydrol. Earth Syst. Sci.* **19**, 4559–4579 (2015).
7. Pimentel, R. et al. Which potential evapotranspiration formula to use in hydrological modeling world-wide?. *Water Resour. Res.* **59**, e2022WR033447 (2023).
8. Massei, N. et al. Understanding and predicting large-scale hydrological variability in a changing environment. *Proc. IAHS* **383**, 141–149 (2020).
9. Yoshida, T. et al. Inference of parameters for a global hydrological model: identifiability and predictive uncertainties of climate-based parameters. *Water Resour. Res.* **58**, e2021WR030660 (2022).
10. Beven, K. A brief history of information and disinformation in hydrological data and the impact on the evaluation of hydrological models. *Hydrol. Sci. J.* **69**, 519–527 (2024).
11. Hrachowitz, M. et al. A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrol. Sci. J.* **58**, 1198–1255 (2013).
12. Gebrechorkos, S. H. et al. Global scale evaluation of precipitation datasets for hydrological modelling. *Hydrol. Earth Syst. Sci. Discuss.* 1–33 https://doi.org/10.5194/hess-2023-251 (2023).
13. Sun, Q. et al. A review of global precipitation data sets: data sources, estimation, and intercomparisons. *Rev. Geophys.* **56**, 79–107 (2018).
14. Tarek, M., Brissette, F. P. & Arsenault, R. Large-scale analysis of global gridded precipitation and temperature datasets for climate change impact studies. *J. Hydrometeorol.* **21**, 2623–2640 (2020).

15. Dorigo, W. et al. Closing the water cycle from observations across scales: where do we stand?. *Bull. Am. Meteorol. Soc.* **102**, E1897–E1935 (2021).
16. Zhang, K. et al. A global dataset of terrestrial evapotranspiration and soil moisture dynamics from 1982 to 2020. *Sci. Data* **11**, 445 (2024).
17. Liu, S., Wang, J., Wang, H. & Wu, Y. Post-processing of hydrological model simulations using the convolutional neural network and support vector regression. *Hydrol. Res.* **53**, 605–621 (2022).
18. Ma, X., Liu, H., Dong, Q., Chen, Q. & Cai, N. Statistical post-processing of multiple meteorological elements using the multimodel integration embedded method. *Atmospheric Res.* **301**, 107269 (2024).
19. Slater, L. J. et al. Hybrid forecasting: blending climate predictions with AI models. *Hydrol Earth Syst Sci* **27**, 1865–1889 (2023).
20. Stachura, G. et al. Machine learning based post-processing of model-derived near-surface air temperature – A multimodel approach. *Q. J. R. Meteorol. Soc.* **150**, 618–631 (2024).
21. Li, W. et al. A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *WIREs Water* **4**, e1246 (2017).
22. Madadgar, S., Moradkhani, H. & Garen, D. Towards improved post-processing of hydrologic forecast ensembles. *Hydrol. Process.* **28**, 104–122 (2014).
23. Lee, D.-G. & Ahn, K.-H. A stacking ensemble model for hydrological post-processing to improve streamflow forecasts at medium-range timescales over South Korea. *J. Hydrol.* **600**, 126681 (2021).
24. Papacharalampous, G., Tyralis, H., Pechlivanidis, I. G., Grimaldi, S. & Volpi, E. Massive feature extraction for explaining and foretelling hydroclimatic time series forecastability at the global scale. *Geosci. Front.* **13**, 101349 (2022).
25. Troin, M., Arsenault, R., Wood, A. W., Brissette, F. & Martel, J.-L. Generating ensemble streamflow forecasts: a review of methods and approaches over the past 40 years. *Water Resour. Res.* **57**, e2020WR028392 (2021).
26. Chakraborty, D., Başağaoğlu, H. & Winterle, J. Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Syst. Appl.* **170**, 114498 (2021).
27. Lee, E. & Kam, J. Deciphering the black box of deep learning for multi-purpose dam operation modeling via explainable scenarios. *J. Hydrol.* **626**, 130177 (2023).
28. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
29. Girons Lopez, M., Crochemore, L. & Pechlivanidis, I. G. Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden. *Hydrol. Earth Syst. Sci.* **25**, 1189–1209 (2021).
30. 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).
31. Tang, S. et al. Optimal postprocessing strategies with LSTM for global streamflow prediction in ungauged basins. *Water Resour. Res.* **59**, e2022WR034352 (2023).
32. Du, Y., Clemenzi, I. & Pechlivanidis, I. G. Hydrological regimes explain the seasonal predictability of streamflow extremes. *Environ. Res. Lett.* **18**, 094060 (2023).
33. 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).
34. Li, D., Marshall, L., Liang, Z., Sharma, A. & Zhou, Y. Characterizing distributed hydrological model residual errors using a probabilistic long short-term memory network. *J. Hydrol.* **603**, 126888 (2021).
35. Duan, Q., Ajami, N. K., Gao, X. & Sorooshian, S. Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**, 1371–1386 (2007).
36. Höge, M., Guthke, A. & Nowak, W. The hydrologist's guide to Bayesian model selection, averaging and combination. *J. Hydrol.* **572**, 96–107 (2019).
37. Madadgar, S. & Moradkhani, H. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resour. Res.* **50**, 9586–9603 (2014).
38. Sattari, A., Jafarzadegan, K. & Moradkhani, H. Enhancing streamflow predictions with machine learning and Copula-Embedded Bayesian model averaging. *J. Hydrol.* **643**, 131986 (2024).
39. Sutanto, S. J. & Van Lanen, H. A. J. Catchment memory explains hydrological drought forecast performance. *Sci. Rep.* **12**, 2689 (2022).
40. Guo, Y., Zhang, Y., Zhang, L. & Wang, Z. Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: a comprehensive review. *WIREs Water* **8**, e1487 (2021).
41. Betterle, A. & Botter, G. Does catchment nestedness enhance hydrological similarity?. *Geophys. Res. Lett.* **48**, e2021GL094148 (2021).
42. Beck, H. E. et al. Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *J. Geophys. Res. Atmospheres* **125**, e2019JD031485 (2020).
43. Musuuza, J. L., Crochemore, L. & Pechlivanidis, I. G. Evaluation of earth observations and in situ data assimilation for seasonal hydrological forecasting. *Water Resour. Res.* **59**, e2022WR033655 (2023).
44. Dasgupta, A. Connecting hydrological modelling and forecasting from global to local scales: Perspectives from an international joint virtual workshop. J. Flood Risk Manag. 18, e12880 (2023).
45. Jackson-Blake, L. A. et al. Opportunities for seasonal forecasting to support water management outside the tropics. *Hydrol. Earth Syst. Sci.* **26**, 1389–1406 (2022).
46. Crochemore, L., Ramos, M.-H. & Pechlivanidis, I. G. Can continental models convey useful seasonal hydrologic information at the catchment scale?. *Water Resour. Res.* **56**, e2019WR025700 (2020).
47. Girons Lopez, M., Bosshard, T., Crochemore, L. & Pechlivanidis, I. G. Leveraging GCM-based forecasts for enhanced seasonal streamflow prediction in diverse hydrological regimes. *J. Hydrol.* **650**, 132504 (2025).
48. Chang, A. Y.-Y. et al. Exploring the use of european weather regimes for improving user-relevant hydrological forecasts at the subseasonal scale in Switzerland. https://doi.org/10.1175/JHM-D-21-0245.1 (2023).
49. Arheimer, B. et al. The IAHS science for solutions decade, with hydrology engaging local people IN one global world (HELPING). *Hydrol. Sci. J.* **69**, 1417–1435 (2024).
50. United Nations. Early warnings for all: executive action plan 2023-2027. https://www.preventionweb.net/publication/early-warnings-all-executive-action-plan-2023-2027 (2022).
51. Hundecha, Y., Arheimer, B., Donnelly, C. & Pechlivanidis, I. A regional parameter estimation scheme for a pan-European multi-basin model. *J. Hydrol. Reg. Stud.* **6**, 90–111 (2016).
52. Brendel, C., Capell, R. & Bartosova, A. Rational gaze: presenting the open-source HYPEtools R package for analysis, visualization, and interpretation of hydrological models and datasets. *Environ. Model. Softw.* **178**, 106094 (2024).
53. Berg, P., Almén, F. & Bozhinova, D. HydroGFD3.0 (Hydrological Global Forcing Data): a 25 km global precipitation and temperature data set updated in near-real time. *Earth Syst. Sci. Data* **13**, 1531–1545 (2021).
54. Madsen, H. & Thyregod, P. *Introduction to General and Generalized Linear Models*. (CRC Press, 2010).
55. Gudmundsson, L., Bremnes, J. B., Haugen, J. E. & Engen-Skaugen, T. Technical note: downscaling RCM precipitation to the station scale

using statistical transformations – A comparison of methods. *Hydrol. Earth Syst. Sci.* **16**, 3383–3390 (2012).
56. Pham, L. T., Luo, L. & Finley, A. Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. *Hydrol. Earth Syst. Sci.* **25**, 2997–3015 (2021).
57. Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **22**, 6005–6022 (2018).
58. Liu, J., Koch, J., Stisen, S., Troldborg, L. & Schneider, R. J. M. A national scale hybrid model for enhanced streamflow estimation – Consolidating a physically based hydrological model with long short-term memory networks. *Hydrol. Earth Syst. Sci. Discuss*. 1–34 https://doi.org/10.5194/hess-2023-235 (2023).
59. Willmott, C. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **30**, 79–82 (2005).
60. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.* **10**, 282–290 (1970).
61. Lamontagne, J. R., Barber, C. A. & Vogel, R. M. Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resour. Res.* **56**, e2020WR027101 (2020).
62. Crochemore, L. et al. Comparing expert judgement and numerical criteria for hydrograph evaluation. *Hydrol. Sci. J.* **60**, 402–423 (2015).
63. Clark, M. P. et al. The abuse of popular performance metrics in hydrologic modeling. *Water Resour. Res.* **57**, e2020WR029001 (2021).
64. Moriasi, D. N., Gitau, M. W., Pai, N. & Daggupati, P. Hydrologic and water quality models: performance measures and evaluation criteria. *Am. Soc. Agric. Biol. Eng.* **58**, 1763–1785 (2015).
65. Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees*. https://doi.org/10.1201/9781315139470 (Chapman and Hall/CRC, New York, 2017).
66. Jiang, Z., Li, W., Xu, J. & Li, L. Extreme precipitation indices over China in CMIP5 models. Part I: model evaluation. *J. Clim.* **28**, 8603–8619 (2015).

Acknowledgements

This study was funded by the EU Horizon 2020 project CLINT (Climate Intelligence: Extreme events detection, attribution and adaptation design using machine learning) under Grant Agreement 101003876 and by the EU Horizon 2020 project I-CISK (Innovating climate services through integrating scientific and local knowledge) under Grant Agreement 101037293. Funding was also received from the EU Horizon 2020 project MedEWSa (Mediterranean and pan-European forecast and Early Warning System against natural hazards) under Grant Agreement 101121192.

Author contributions

Authorship has been assigned to researchers who participated in this study in compliance with global research ethics & inclusion standards. Y.D. contributed to the basic idea, study design, code development, model runs, result analysis and figures, interpretation of results, and writing of the manuscript. I.G.P. was responsible for the project management and funding acquisition and contributed to the basic idea, study design, result analysis and figures, interpretation of results, and writing of the manuscript.

Funding

Open access funding provided by Swedish Meteorological and Hydrological Institute.

Competing interests

The authors of this work declare no conflict of interest. Our funders had no role in the choice of research project; design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-025-02324-y.

Correspondence and requests for materials should be addressed to Yiheng Du or Ilias G. Pechlivanidis.

Peer review information *Communications Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Somaparna Ghosh A peer review file is available.