Journal of Hydroinformatics

Journal of Hydroinformatics Vol 25 No 5, 1648 doi: 10.2166/hydro.2023.217

Global streamflow modelling using process-informed machine learning

Michele Magni [a,*], Edwin H. Sutanudjaja[a], Youchen Shen[a,b] and Derek Karssenberg[a]

[a] Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
[b] Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands
*Corresponding author. E-mail: m.magni@uu.nl

MM, 0000-0002-6184-4413

ABSTRACT

We present a novel hybrid framework that incorporates information from the process-based global hydrological model PCR-GLOBWB, to reduce prediction errors in streamflow simulations. In addition to catchment attributes and meteorological data, our methodology employs simulated streamflow and state variables from PCR-GLOBWB as predictors of observed river discharge. These outputs are used in a random forest, trained on a global database of streamflow measurements, to improve estimates of simulated river discharge across the globe. PCR-GLOBWB was run for the years 1979–2019 at 30 arcmin and its inputs and outputs were upscaled from daily to monthly time steps. A single random forest model was trained with these state variables, meteorological data and catchment attributes, as predictors of observed streamflow at 2,286 stations worldwide. Model performance was evaluated using Kling–Gupta efficiency (KGE). Results based on cross-validation show that the model is capable of discerning between a variety of hydroclimatic conditions and river flow dynamics, improving KGE of PCR-GLOBWB simulations at more than 80% of testing locations and increasing median KGE from −0.03 in uncalibrated runs to 0.51 after post-processing. Performance boosts are usually independent of the availability of streamflow data, making our method a potential candidate in addressing prediction in poorly gauged and ungauged basins.

Key words: global hydrology, hybrid streamflow modelling, machine learning, post-processing, random forests

HIGHLIGHTS

- A hybrid framework for global streamflow modelling is developed, connecting PCR-GLOBWB with random forest.
- The framework enables the correction of global-scale streamflow predictions with parsimonious parametrization.
- Random forests improve streamflow predictions better when additionally fed with outputs from the hydrological model, as opposed to only using meteorological forcing and catchment attributes.

1. INTRODUCTION

River discharge is a highly human-relevant component of the hydrological cycle, as it influences and is influenced by human practices and anthropogenic climate change (Haddeland *et al.* 2013). Accurate streamflow predictions from rainfall-runoff models are needed for efficient water management, flood and drought risk assessment and for the direction of mitigation efforts (Sharma & Machiwal 2021), as well as improving water access and the understanding of human–water interactions (Montanari *et al.* 2013).

The growing complexity of process-based hydrological models (PBHM) over the past few decades, however, has not considerably increased accuracy in streamflow prediction (Liu & Gupta 2007). As more processes are included in the representation of streamflow generation, the uncertainties that characterize each model component have propagated errors to the simulated runoff (Liu & Gupta 2007; Montanari *et al.* 2009; Beven 2012; Li *et al.* 2016; Moges *et al.* 2020). Observational data used to evaluate the model are also subject to uncertainty (McMillan *et al.* 2018).

In contrast to classic calibration methods (Liu & Gupta 2007; Todini 2007), alternative approaches to address hydrological uncertainty focus on residuals, by estimating the overall error of a simulation as the difference between modelled and observed streamflow (Liu & Gupta 2007; Evin *et al.* 2014; Li *et al.* 2016). Such methods are also referred to as 'aggregated error models', as they characterize total uncertainty without attempting to disentangle individual sources of error (Evin *et al.*

2014; Matthews *et al.* 2022). Error models may be characterized as 'joint' or 'post-processor', depending on the timing of streamflow error correction (Evin *et al.* 2014). Here we focus on post-processing, where the hydrological model is first run, its output compared to observational data and then corrected with a selected error model to improve streamflow forecasts (Bogner *et al.* 2016; Hopson *et al.* 2018; Matthews *et al.* 2022), thereby ignoring interactions between hydrological and error model parameters (Evin *et al.* 2014; Valdez *et al.* 2022).

In the past two decades, statistical learning (SL) has gained increasing attention in the hydrological community (Nearing *et al.* 2021; Xu & Liang 2021; Mosaffa *et al.* 2022), sparked by its many different implementations in the geosciences at large (Dramsch 2020). SL applications in streamflow prediction mostly relate to the use of various forms of artificial neural networks (ANNs) (Abrahart *et al.* 2012). An extensive list of SL algorithms for hybrid runoff modelling can be found in Mosavi *et al.* (2018). PBHM and data-driven models have complementary strengths and weaknesses, and combining them in various ways provides opportunities to enhance the knowledge of both modelling approaches (Herath *et al.* 2021; Xu & Liang 2021). This integration of theoretical knowledge and observational data enables the extraction of greater volumes of information (Ghaith *et al.* 2020; Nearing *et al.* 2021).

In more recent years, the research has been directed towards hybrid streamflow models employing SL algorithms that are more computationally efficient than ANNs (Tyralis *et al.* 2019). Specifically, random forest (RF) (Breiman 2001), an ensemble machine learning algorithm (Zounemat-Kermani *et al.* 2021) that uses decision trees as base learners, has become increasingly popular in the hydrological sciences (Tyralis *et al.* 2019). RF has been successfully applied in streamflow modelling (Li *et al.* 2019; Tyralis *et al.* 2019; Pham *et al.* 2021; Roy *et al.* 2023), with performances comparable to those of more complex ANNs (Li *et al.* 2019; Desai & Ouarda 2021; Hauswirth *et al.* 2021; Ghazikhani *et al.* 2022), albeit being extremely parsimonious in terms of parameters, computationally efficient and of easier interpretability (Tyralis *et al.* 2019). All previous studies, however, are limited to the use of observed streamflow, meteorological inputs and a few other variables as predictors of river discharge.

Shen *et al.* (2022a) developed an innovative hybrid modelling framework, consisting of an RF-based post-processing approach to correct predictions from the global hydrological model (GHM) PCRaster Global Water Balance (PCR-GLOBWB) (Sutanudjaja *et al.* 2018). Meteorological input as well as intermediate hydrological state variables from PCR-GLOBWB were used as predictors for the RF to estimate residuals in streamflow predictions, which were then applied to correct simulated discharge at the daily scale. The RF was trained and tested at three individual stations in the Rhine basin characterized by a variety of physiographic features and streamflow generation processes, showing excellent improvements in model performance. The RF post-processor achieved comparable results when PCR-GLOBWB was not manually calibrated prior to error correction, implying that the SL algorithm was able to properly account for the various sources of uncertainty in the uncalibrated hydrological model, without the need for previous human intervention.

This approach reconciles the use of PBHM and data-driven models (Todini 2007), by exploiting the theoretical understanding that is engrained in the intermediate outputs of the hydrological model. This makes it potentially capable to characterize universal catchment behaviours (Sivapalan 2005), by extrapolating transferable knowledge on the main sources of error of a GHM, across a variety of climates and catchment characteristics. Unlike other methodologies, proper availability of observational data and adjustments to the single-station framework can thus enable larger-scale correction of discharge for predictions in ungauged basins (PUB) (Sivalapan *et al.* 2003; Hrachowitz *et al.* 2013).

A potential application of such a framework is thus to correct simulations of the global hydrological cycle, of which there are very rare and only recent examples. Kraft *et al.* (2022) developed a joint hybrid framework, but they exclusively rely on a combination of long short-term memory networks and simple ANNs to enhance model results, with little insight into the processes involved in improving streamflow predictions. The method by Shen *et al.* (2022a), on the other hand, could only be applied to singular stations, undermining its potential to correct simulations at any catchment over the globe.

The main idea underlying this study was to test if information on errors acquired by training the model in gauged catchments could be used as proxy to correct streamflow simulations at previously unseen locations. In this context, the current study aims at expanding the work of Shen *et al.* (2022a) to the global scale. The work presented here is thus intended to answer the following research questions: can a hybrid framework extrapolate beyond individual catchment characteristics to properly address streamflow prediction in ungauged basins? Does a hybrid framework, which includes hydrological model outputs as predictors of observed streamflow, outperform an SL methodology that only relies on catchment attributes and meteorological inputs?

Reaching this goal requires including predictors of river discharge that involve basin characteristics of a pixel (topography, soil and groundwater characteristics, land use and climatic indices) to be able to generalize global watershed behaviours. It also requires that a global dataset of streamflow observations is built, such that a variety of hydroclimatic regions is captured during model training. This will also enable the understanding of where future hydrological modelling efforts should be directed to address the main sources of uncertainty in global-scale streamflow models.

In the following, we present a global dataset of monthly streamflow observations (section 2) and the novel hybrid methodology employed to generate post-processed values of river discharge across the globe (section 3). Section 4 describes the results as they happen during workflow (hyperparameter tuning, RF training and cross-validation), while section 5 discusses a few key points concerning the potential use of our framework for prediction in ungauged basins, its current limitations and future research directions. Conclusions are given in section 6.

2. DATA

River discharge data were downloaded from the Global Runoff Data Centre (GRDC 2022). Stations were selected with the following criteria: spatially, a minimum upstream area of 10,000 km^2, such that catchments occupied at least ~4 pixel windows (PCR-GLOBWB was run at a resolution of 30 arcmin, or ~50 km at the equator) and enough physical processes were assumed to be represented by the hydrological model; temporally, GRDC stations had to have at least 1 year in the period 1979–2019, such that a few timesteps could be compared between simulated and observed discharge. We considered both managed and unmanaged catchments.

The selection brought to the final dataset used in this study, consisting of 2,286 stations with variable availability of monthly data for the years 1979–2019 (Figure 1), for a total of 641,735 observations. The average catchment size used in this study was 128,480 km^2 (1st q. = 16,500 km^2, median = 32,550 km^2, 3rd q. = 82,147 km^2, max = 4,680,000 km^2). On average, stations were missing 43% of data in the 1979–2019 period, with the richest regions in observational records being North and South America, Australia, South-West Africa and Europe.

3. METHODS

The following sections describe the numerical models used to characterize streamflow predictions at the global scale, and the different predictor configurations that were tested. Figure 2 shows the general scheme of the hybrid post-processing strategy. The inputs to PCR-GLOBWB are the meteorological forcing (time-series, in blue) and the model parameters maps (static, in

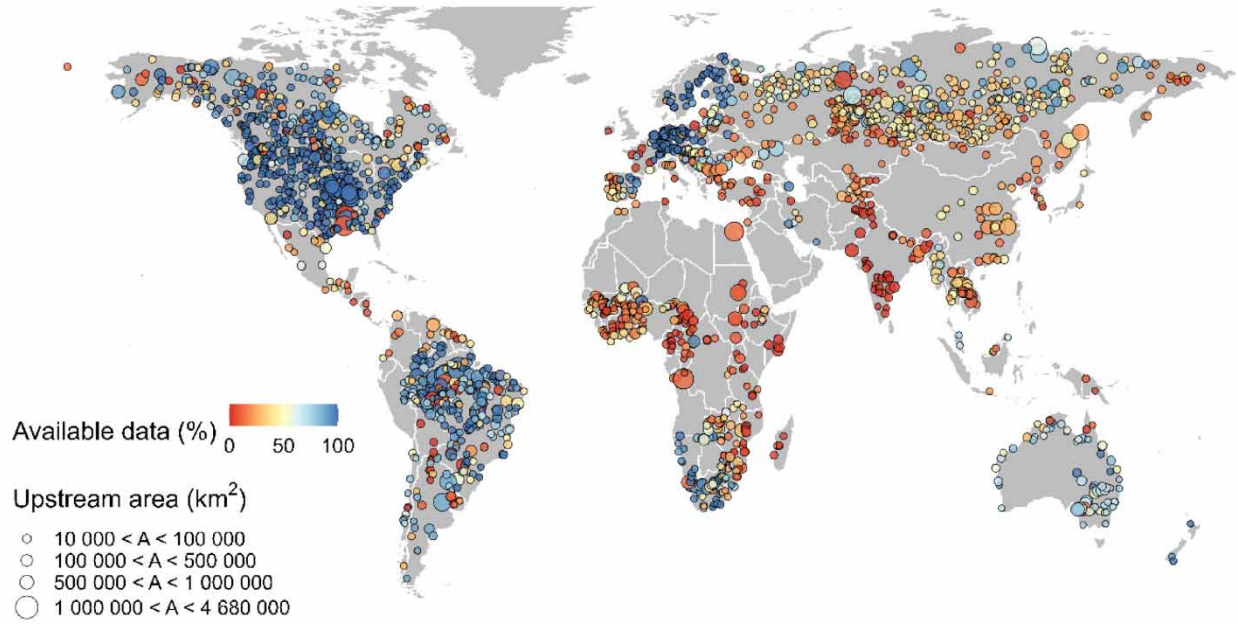

Figure 1 | Availability (%) of monthly river discharge data for the years 1979–2019, for GRDC stations with a minimum upstream area of 10,000 km^2. The size of the circles is proportional to the upstream area (km^2).

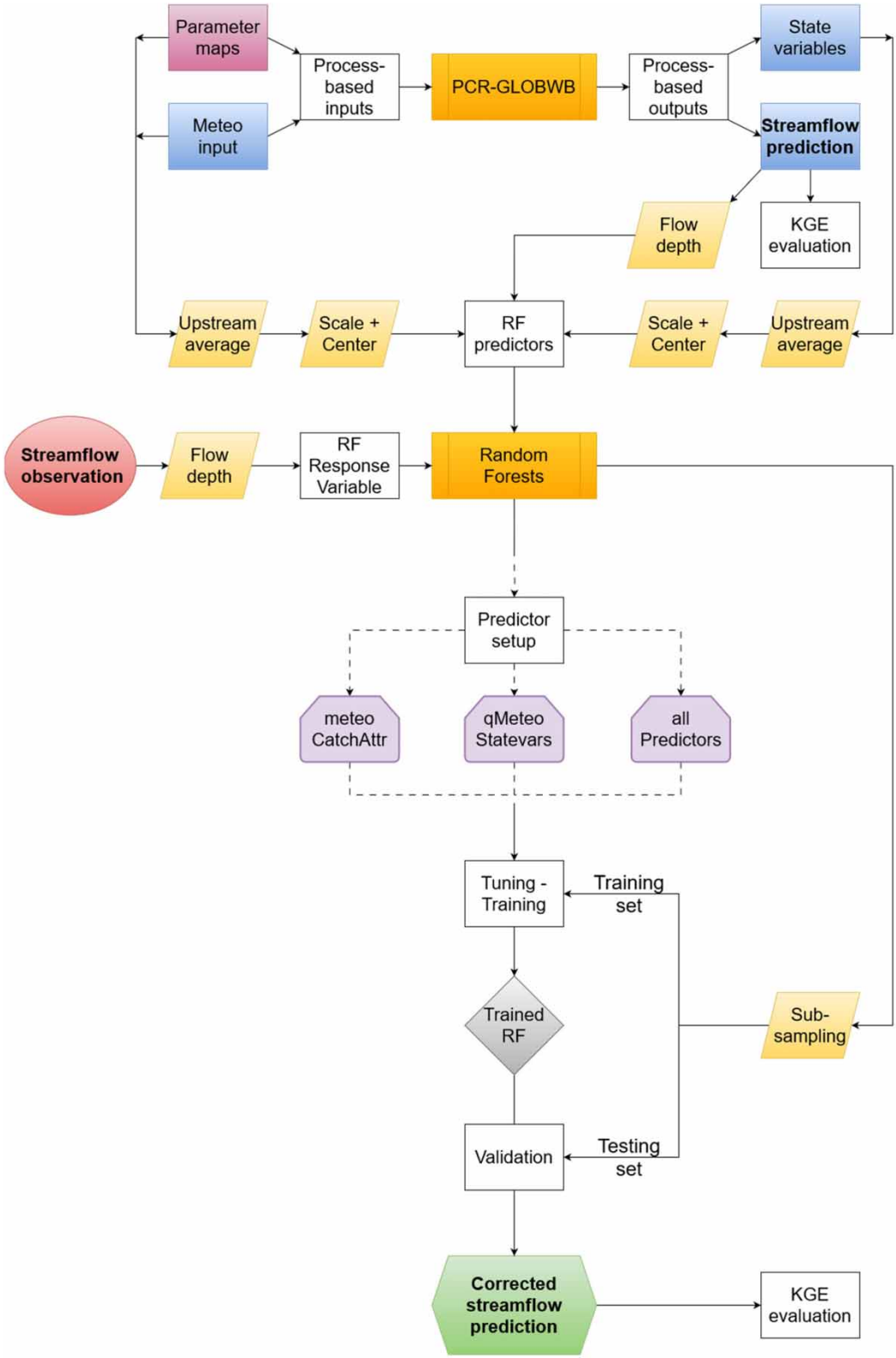

Figure 2 | Workflow of the modelling framework (modified after Shen *et al.* 2022a). The initial modelling phase is characterized by the extraction of predictor variables from PCR-GLOBWB as inputs for the RF. Prior to training, only one of the predictor setups can be chosen (singularity represented by the dashed lines). The available data are split into training and testing sets (with a ratio of approximately 70/30), which are used to evaluate overall model performance.

purple) (as in Sutanudjaja *et al.* 2018), which are then normalized to the upstream area and used as predictors for the RF. Static catchment attributes are thus extracted from the parameter maps. The same procedure is executed on the PCR-GLOBWB hydrological state variables, while streamflow simulations and observations are converted to flow depth based on the upstream area at the pixel location. Streamflow observations (red circle) are the response variable for the RF. A single RF is then trained using these predictors and a corrected value of streamflow at previously unseen locations is generated. More details on the overall workflow are given in the following.

3.1. PCR-GLOBWB

PCR-GLOBWB (Sutanudjaja *et al.* 2018) is a grid-based GHM. GHMs attempt to simulate the global hydrological cycle and associated processes, providing valuable spatiotemporal estimates of global water resources (Sood & Smakhtin 2015). PCR-GLOBWB represents anthropogenic influences on the water cycle (de Graaf *et al.* 2014; Wada *et al.* 2014) alongside natural hydrological processes (van Beek *et al.* 2011) and is characterized by high flexibility due to its modular structure (Supplementary Material S1).

Meteorological forcing for the hydrological model consists of three variables, namely precipitation (P), temperature (T) and reference potential evapotranspiration (PET), based on the W5E5 v2 (Lange *et al.* 2021). Here, PCR-GLOBWB was run without calibration at daily timesteps for the years 1979–2019, in its version with a spatial resolution of 30arcmin (~50 km at the equator), using its standard input parameters (Sutanudjaja *et al.* 2019) and with the kinematic wave solution for water routing. Model inputs and outputs (I/O) were then upscaled to monthly averages.

3.2. Random forests

RF (Breiman 2001) is an ensemble tree-based algorithm that samples predictor variables in each split node of an independent tree and aggregates all the trees for prediction. Supplementary Material S2 shows the in-depth structure of a random forest. The number of trees in the forest is controlled by the parameters *ntree*. The depth of each tree, i.e. the number of iterations that are executed before tree growth is stopped, is regulated by the *nodesize* parameter. The ensemble algorithm in the RF can deal with the high correlations occurring between the predictors by randomly selecting a subset of candidate variables for each split. This selection is effectively carried out through the *mtry* parameter, which can thus take on values between 1 and the total number of predictors. This bootstrapping technique avoids highly influential variables dominating all trees. Each tree is then developed by minimizing a loss function (here, Mean-Squared Error (MSE), t_i). After a certain value is determined during hyperparameter calibration, a larger *ntree* will lead to a decrease in computational efficiency without a justifiable increase in predictive performance. On the other hand, larger *mtry* and *nodesize* values will usually cause overfitting on top of larger computational times.

In RF, only a certain subset of the training data is used to grow the forest (*Bootstrap n*). The unused data, referred to as Out-Of-Bag (OOB) observations, are exploited to estimate the generalization error. Each OOB observation has an OOB prediction, from which an overall OOB root-mean-squared error (RMSE) of the n observations can be obtained instead of using cross-validation, which requires larger computation efforts and may lead to biased error estimates (Breiman 2001). Moreover, internal estimates of variable importance in reducing predictive error are calculated as mean decrease in node impurity (Grömping 2009; Gregorutti *et al.* 2016; Wright & Ziegler 2017). These are useful in understanding which predictors contribute the most to overall model performance and to increase the interpretability of RF.

3.3. Hybrid modelling setup

3.3.1. Predictors selection and pre-processing

Predictor variables were exclusively extracted from PCR-GLOBWB inputs and outputs (I/O), allowing for coherency in data-feeding to the RF to correct uncertainty in the hydrological model, i.e. no external dataset was used. To predict the observed discharge at time t (response variable), we used both time-variant and static predictors. No lagged variable was employed. Time-variant predictors consist of all PCR-GLOBWB I/O, while static predictors are extracted from PCR-GLOBWB parameter maps (see Supplementary Material S3 for a few examples) and are used to inform the algorithm with catchment characteristics.

All predictors were first averaged to the upstream area using the PCRaster Python framework (Karssenberg *et al.* 2010). They were then extracted from the netCDF (network Common Data Form) files where PCR-GLOBWB I/O is stored using the pixels corresponding to each station of the GRDC. These then had its own predictor table to be used for training or

validation. Static predictor maps' values were scaled and centred globally prior to extraction, such that they had a mean of 0 and a standard deviation of 1. For meteorological inputs and state variables, this was done after feature extraction. Simulated and observed runoff ($m^3\ s^{-1}$) were converted to flow depth (m day^{-1}), using data on the upstream area.

The final selection of predictors and a brief description of their functioning in PCR-GLOBWB are shown in Supplementary Material S4. For further details on their inner workings, the reader is referred to Sutanudjaja *et al.* (2018). Supplementary Material S5 shows the correlation analysis that was conducted prior to training to verify which predictors were highly correlated (Gregorutti *et al.* 2016).

3.3.2. Hybrid modelling configurations

Shen *et al.* (2022a) showed that the RF improves the performance of the uncalibrated hydrological model, with comparable results when PCR-GLOBWB is manually calibrated prior to feeding data to the RF. They also proved that the hybrid runoff framework always achieves better performance both than the uncalibrated and calibrated PCR-GLOBWB. On the other hand, the calibration of a GHM requires massive efforts, which were beyond the scope of this study. For these reasons, we decided not to calibrate PCR-GLOBWB. Overall, our approach saves time in setting up the model framework, in addition to enabling the algorithm to compensate for errors in hydrological model parameters and input without prior human intervention.

In Shen *et al.* (2022a), the response variable in the RF post-processor was the residual between the simulated and the observed discharge. However, early trials on previously untested GRDC stations showed that this approach would cause the creation of negative discharge values and a generalized incapability of the RF to properly correct runoff simulations. A preliminary test proved that directly using streamflow observations as a response variable in the RF, instead of the residual, enhanced the overall performance of the post-processor, therefore also reducing data pre-processing and model complexity. This stems from the fact that discharge observations can only take positive values, whereas residual dynamics are characterized with both positive and negative values. Additional model developments are described in Supplementary Material S6.

With these preconditions, three different model setups were tested (Table 1), including the following predictor combinations: (1) *qMeteoStatevars* is the setup developed by Shen *et al.* (2022a), which uses uncalibrated discharge (*q*), meteorological input (*meteo*) and uncalibrated hydrological state variables (*Statevars*). Here we developed two additional configurations: (2) *meteoCatchAttr* only employs *meteo* and static catchment attributes (*CatchAttr*). This configuration is used as a baseline to understand if a fully SL-based framework generalizes catchment behaviours, while excluding all PCR-GLOBWB output from the predictors. (3) Finally, in this study, we combine the previous two configurations, by adding predictor values of static catchment attributes, extracted from global maps of hydrological parameters, to *qMeteoStatevars*. This hybrid configuration (*allPredictors*) ultimately shows if the use of outputs from a GHM can complement predictors of input catchment parameters and meteorological forcing to improve streamflow predictions.

3.3.3. Model training and evaluation

RF was implemented in R with the *ranger* package (Wright & Ziegler 2017), a fast implementation of RF for high-dimensional data. The model requires tuning of the three RF hyperparameters (section 3.2): *ntree*, *mtry* and *nodesize*. To do this, the OOB RMSE was used for hyperparameter tuning in the RF, to minimize overfitting and maximize its generalization capabilities, while also optimizing performance.

The hyperparameters were optimized for each post-processor configuration (section 3.3.2) with the following approach: first, forests were grown with a fixed *ntree* of 200, to quickly understand the response of *mtry*. *mtry* was initially searched with a spacing of 5, ranging between 1 and the possible number of predictors; after this, the search was reduced to ~10

Table 1 | Predictors used in the different post-processing setups

	Catchment attributes	Meteorological input	PCR-GLOBWB state variables	PCR-GLOBWB discharge	Total predictors
qMeteoStatevars		v	v	v	25
meteoCatchAttr	V	v			30
allPredictors	v	v	v	v	52

units within the *mtry* with the two lowest RMSEs, and *mtry* was calculated with a spacing of 1. Once the optimal *mtry* was determined for each configuration, this was kept constant to understand the response of *ntree*. *nodesize* was kept constant with a value of 5 as in Shen *et al.* (2022a). The hyperparameter set that gave the smallest OOB RMSE with the highest computational efficiency was used to train each RF post-processor.

To train and validate the RF, location-based split sampling was executed, generating five random subsamples of the whole dataset. Each subsample was divided into training stations and testing stations. All data from the training stations were accumulated into a single table that was used to tune the hyperparameters and train the RF. The training table contained ~2/3 of all available observations (~ 427,823 timesteps). During the training phase, the algorithm fits the predictors to the response variable (discharge observations), to generate a predictive model, also referred to as a 'trained RF'.

Each of the five trained RF was then applied on each station from their respective testing dataset and values of streamflow were predicted at each timestep. Improvements in Kling–Gupta Efficiency (KGE) (Gupta *et al.* 2009) were calculated to validate the error-correcting capabilities of the trained RF, comparing the performance of uncalibrated PCR-GLOBWB to the fully SL-based and hybrid setups, for each subsample.

KGE was created to remedy the limitations of the Nash–Sutcliffe Efficiency (NSE) (Nash & Sutcliffe 1970). In fact, exclusively relying on NSE to evaluate a hydrological model can lead to untruthful estimates of its general performance, due to normalization of the MSE in NSE (Gupta *et al.* 2009). This causes the relative importance of bias evaluation to vary across basins and in highly variable observed flows, in addition to systematically underestimating flow variability (Gupta *et al.* 2009). The use of KGE, on the other hand, allows to disentangle the components that constitute the overall performance of a hydrological model, namely linear correlation, bias and variability.

KGE is thus characterized as follows:

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

$$r = \frac{cov(q_m, q_o)}{\sigma_m \cdot \sigma_o}; \tag{2}$$

$$\alpha = \frac{\sigma_m}{\sigma_o}; \tag{3}$$

$$\beta = \frac{\mu_m}{\mu_o}; \tag{4}$$

where r, α and β are, respectively, the linear correlation coefficient between modelled (m) and observed (o) streamflow (covariance over product of standard deviations), and the ratios of their respective standard deviations (variability) and means (bias).

A hydrological model with perfect performance has a KGE of 1, meaning that ideal values of r, α and β correspond to 1, such that the equation term under the square root is minimized to 0. r can take values between -1 and 1, α and β between $-\infty$ and $+\infty$. Improvements upon mean flow benchmark for prediction have been shown to undertake a value of $1-\sqrt{2}$ (~ -0.41) (Knoben *et al.* 2019), which is used in the following as a baseline to evaluate good model performance.

4. RESULTS

4.1. Hyperparameter tuning

Since stations with different availability of observational data were used, the training data are not uniformly distributed across the globe. The table in Supplementary Material S7 enumerates, for each subsample, how many stations were used for training and testing, respectively, and contributions to the total training data for each region of the World Meteorological Association (WMO 2006). The percentage was calculated by dividing the available timesteps in each region by the total timesteps used in a certain training subset. These ratios clearly show that some WMO regions are characterized by data scarcity more than others. This issue can potentially be tackled using the transferability of catchment behaviours within our framework.

Figure 3 shows the hyperparameter tuning procedure, which was executed for each model configuration, for each subsample. Figure 3(a)–3(c) coincides with the first tuning phase, where *ntrees* is kept constant at 200, while Figure 3(d)–3(f) shows the second phase, where the response of larger *ntrees* is computed keeping *mtry* constant at the previously obtained

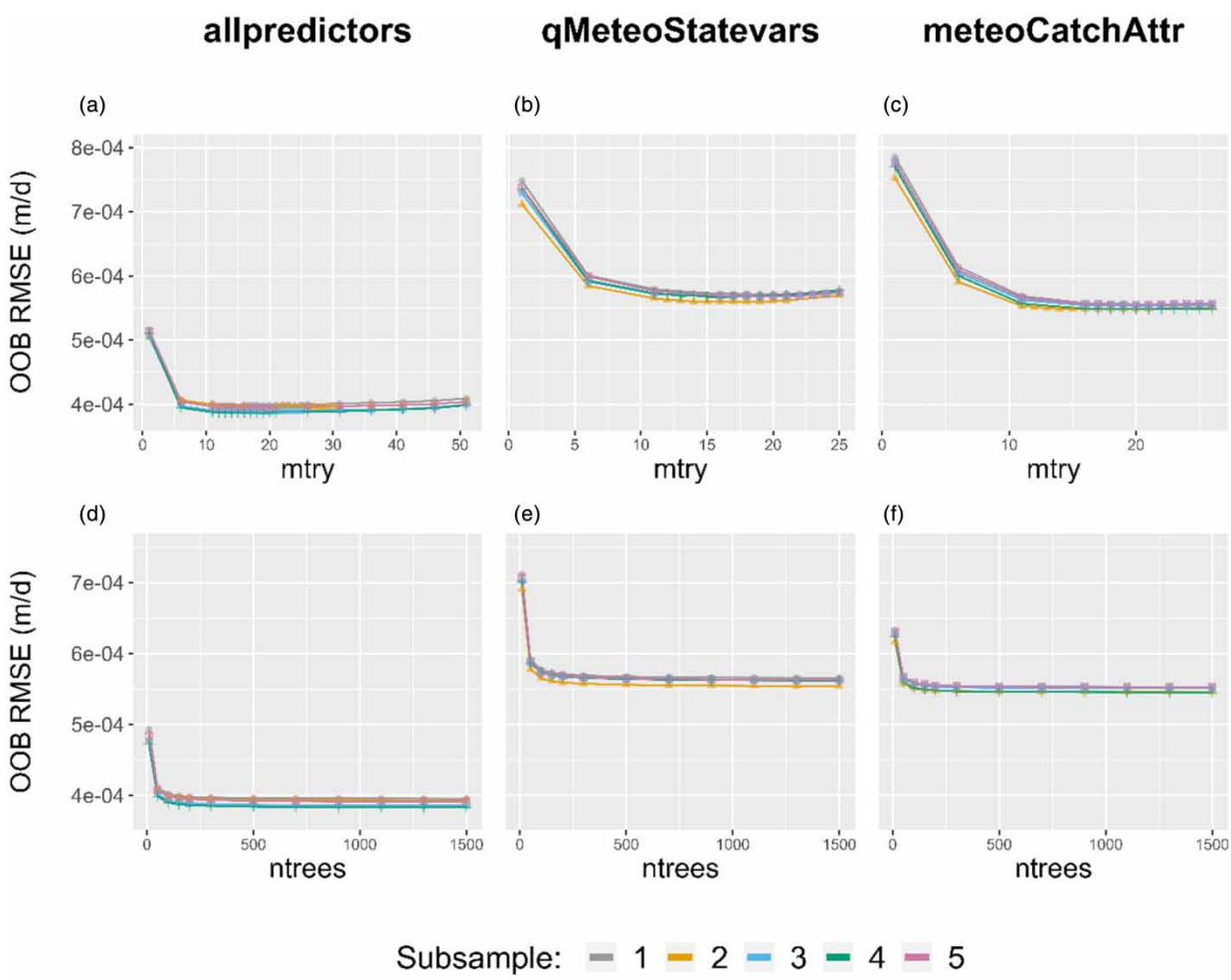

Figure 3 | RF hyperparameter tuning. (a–c) Tuning of *mtry*, using a fixed *ntree* of 200. (d–f) Tuning of *ntrees*, using the optimal *mtry* calculated in the previous step.

optimal value. Even though small variations in each subsample's OOB RMSE can be easily explained due to slight differences in the training datasets, values show comparable trends and error values within the same predictor configuration.

mtry is characterized by an initially decreasing OOB RMSE to a global optimum at a value of around half the possible predictors; higher values cause the RF to slightly overfit the training data and error starts increasing again. The *mtry* that gave the lowest OOB RMSE (Supplementary Material S8) was then used to study the response of *ntrees*. For what concerns *ntrees*, the OOB RMSE converges early for all predictor configurations at ~500 trees, a value that was ultimately used to train the various models. After this point, further decreases in OOB RMSE outweigh the computational necessities of the RF.

4.2. Training

The selected hyperparameters were thus used to train the RF, for each subsample and for each model configuration. Figure 4 shows the variable importance metrics for the three post-processor configurations, calculated as mean decrease in node impurity, averaged over the five random subsamples and with their respective standard deviations. The highest 20 scoring variables are shown, and their values are scaled to their square root to better show small changes.

The results clearly demonstrate that including static predictors influences the relationship between different variables at the node splits. For instance, in the *allPredictors* configuration (Figure 4(a)), half of the major 20 predictors are static ones, with the most significant ones being related to climate, river channel and land use characteristics. Moreover, having different configurations that include various predictor combinations allows to better understand the relative importance within

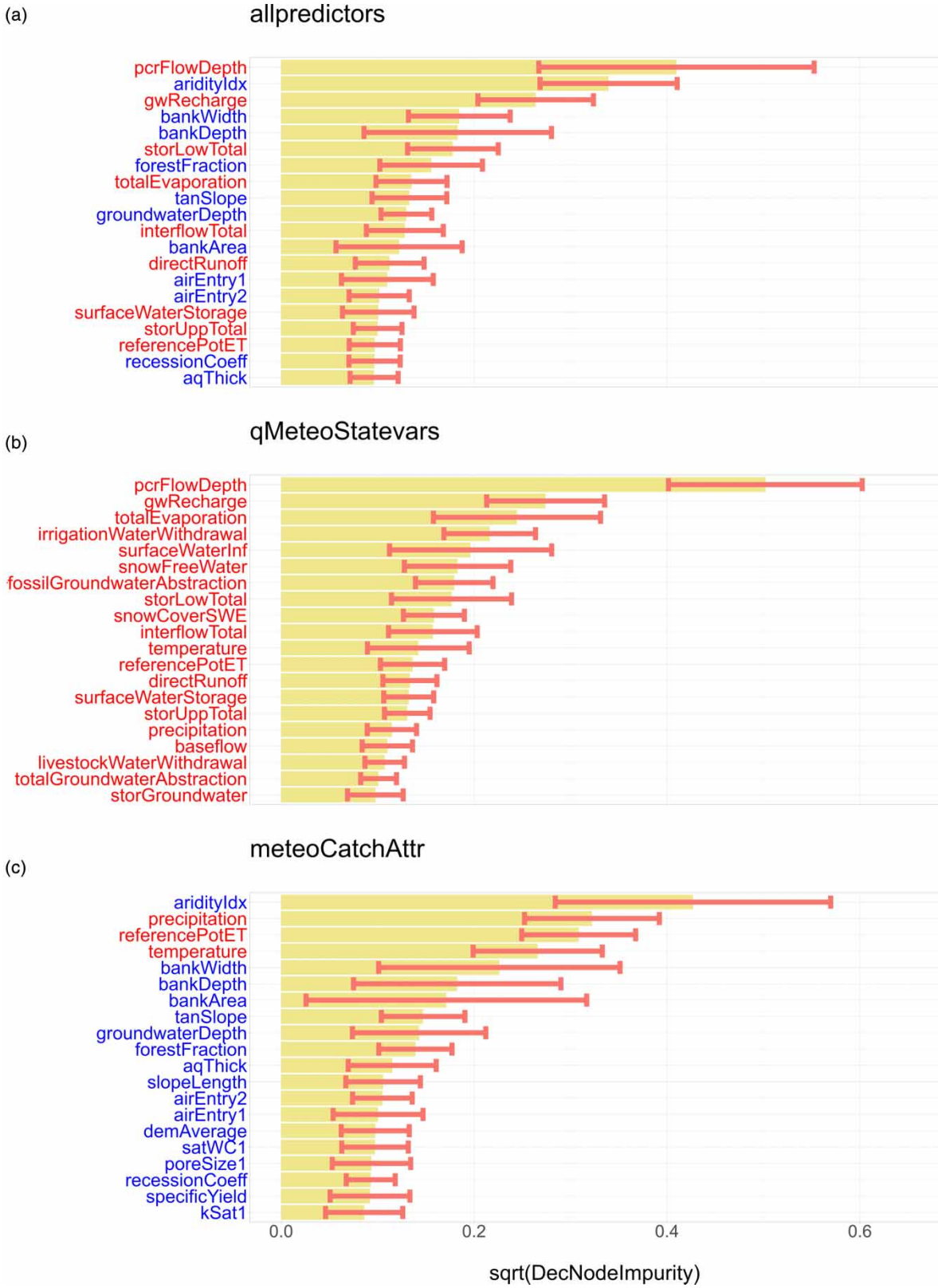

Figure 4 | Square-rooted mean decreases in node impurity for the three post-processor configurations, showing the top 20 variables as average value and standard deviation over the 5 training subsamples. Static predictors are shown in blue, time-variant predictors are coloured in red.

time-series (*qMeteoStatevars*, Figure 4(b)) and within meteorological and catchment parameter inputs (*meteoCatchAttr*, Figure 4(c)).

The most important time-variant variable is the flow depth from PCR-GLOBWB, showing that the process-based streamflow prediction is highly informative to the SL algorithm, even in its uncalibrated form. Other very significant transient predictors are the groundwater recharge rate, the water storage of the lower soil layer and the total evaporation.

These conclusions are, however, subjective to the training dataset, where contributions of discharge observations from GRDC stations are unbalanced towards some regions more than others. The bars in Figure 4 clearly show the relative uncertainty of each predictor, with some being more sensitive to the specific training sub-dataset, especially the flow depth, the aridity index, bank width and bank depth. Another variable with large uncertainty in importance is the irrigation water withdrawal (Figure 4(b)), most certainly dependent on the large differences in irrigated area upstream of a certain station around the globe.

4.3. Cross-validation

After training the RF, each of them was applied to their respective test dataset on a station-by-station basis. The results for the different subsamples are summarized in Figure 5, where each column corresponds to a subsample and each row corresponds to KGE and its three components (respectively, linear correlation, variability and bias). The uncalibrated runs from PCR-GLOBWB are shown in grey, while the other boxes correspond to the different predictor setups during post-processing.

The validation results show that the various configurations work consistently, independently of the training set, with small divergences related to the random subsampling and the base performance of the PCR-GLOBWB. Compared to the uncalibrated runs (median KGE $= -0.03$), the *qMeteoStatevars* (median KGE $= 0.37$) setup (from Shen *et al.* 2022a) improved correlation better than *meteoCatchAttr* (median KGE $= 0.34$), while this last one achieved greater performances in improving bias and variability. This can also be seen in the greater interquartile ranges in KGE for the *qMeteoStatevars* setup. Intuitively, the hydrological state variables act as a 'state update' of the hydrological system, while the predictors relating to catchment attributes and meteorological inputs enable a generalization of catchment behaviours.

However, blending all these predictors together in a hybrid setup ultimately achieved the overall best performance, as it is quite evident from the results of the *allPredictors* configuration. Including meteorological variables, hydrological state variables and static catchment attributes brought a clear improvement in KGE (median KGE $= 0.51$) as well as in correlation, bias and variability across all subsamples. The biggest performance boost in correlation comes from the hydrological state variables, while reduced variability and bias ratio are likely caused by the catchment attributes. This proves that including descriptors of catchment behaviours improves the reduction in predictive error of the RF-based post-processor. Supplementary Material S9 summarizes the values of KGE and its components cumulated over all subsamples (rightmost column in Figure 5).

Figure 6 shows the KGE values for all GRDC stations that appeared in a minimum of one testing dataset, for the uncalibrated PCR-GLOBWB (Figure 6(a)) and following RF modelling (Figure 6(b) for *meteoCatchAttr* and Figure 6(c) for *allPredictors*). Both the fully RF-based and the hybrid RF model improved streamflow predictions above mean benchmark (KGE > -0.41) at ~85% of all stations in the various testing subsets, against ~65% in the uncalibrated PCR-GLOBWB. Although a fully RF-based setup (*meteoCatchAttr*) already reaches satisfactory results, it is evident both from Figure 6 and from the distributions of the KGE values (Supplementary Material S10) that the hybrid configuration (*allPredictors*) ultimately brings about the best set of streamflow predictions, confirming the capture of significant information in the discharge and state variables modelled by PCR-GLOBWB.

Improvements were usually independent of the availability of training data (e.g. Central and Western Africa, South-East Asia, the Iberic peninsula and Russia) (Figure 1), demonstrating that a good donor catchment from a data-rich region was often available in the training dataset. On the other hand, it is also the case that not all data-rich regions necessarily received performance boosts, which may indicate the influence of other factors in poor behaviour generalization, e.g. arid environments and/or highly engineered catchments.

Figure 7 shows the hydrographs, their flow duration curves and the residual scatterplots of streamflow predictions at selected stations, comparing the *uncalibrated* PCR-GLOBWB runs with the hybrid runoff from the *allPredictors* scenario, for the period 1980–1989. Even though only 10 years are shown for better clarity, the model is designed to produce a continuous time-series also when observations are not available, which are, however, necessary to compare KGE and calculate residuals.

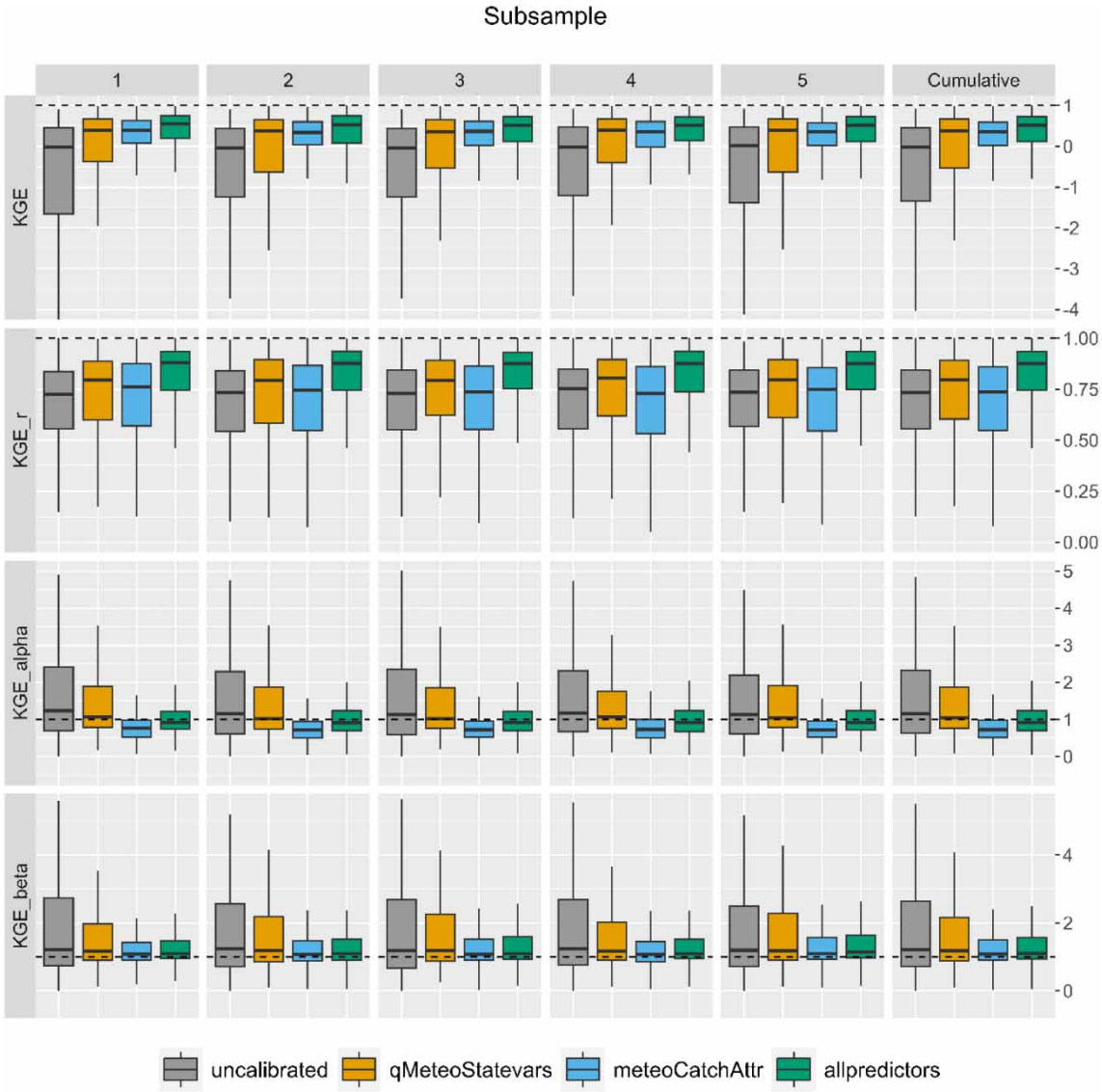

Figure 5 | Boxplots of (from the top) KGE and its three components (correlation (*r*), variability (alpha) and bias (beta)) for the validation sets of the five subsamples in the k-fold cross-validation. The last column on the right accumulates all values from all subsamples; if a station appeared in more than one subsample, average values were used. 'Uncalibrated' refers to the uncalibrated PCR-GLOBWB discharge simulations. The dashed lines indicate the value of 1, which implies a perfect representation of the streamflow component by the model.

In Figure 7(a) and 7(b), the hydrographs show that streamflow dynamics remain realistic after post-processing and are not negatively impacted by the RF. Moreover, the model is capable of correcting significant differences between real-world and simulated values, which can be better seen in the flow duration curves, highlighting how our methodology is effective at correcting flow quantiles. The residual plots show that the framework properly accounts for autocorrelation without the need for lagged state variables and correctly addresses heteroscedasticity, as increasing values of discharge magnitude are not anymore associated with a higher difference in error between observed and simulated values.

On average, streamflow simulations were improved after post-processing at ~80% of all stations in the validation subsets, for the *allPredictors* configuration, thereby increasing median KGE from −0.19 to 0.57 (as exemplified in Figure 7(a) and 7(b)). However, this implies that KGE degraded after passing through the RF in ~20% of testing stations for the best setup. At such locations, median KGE decreased from 0.50 (*uncalibrated*) to 0.23 (*allPredictors*). An example of this case is shown in Figure 7(c).

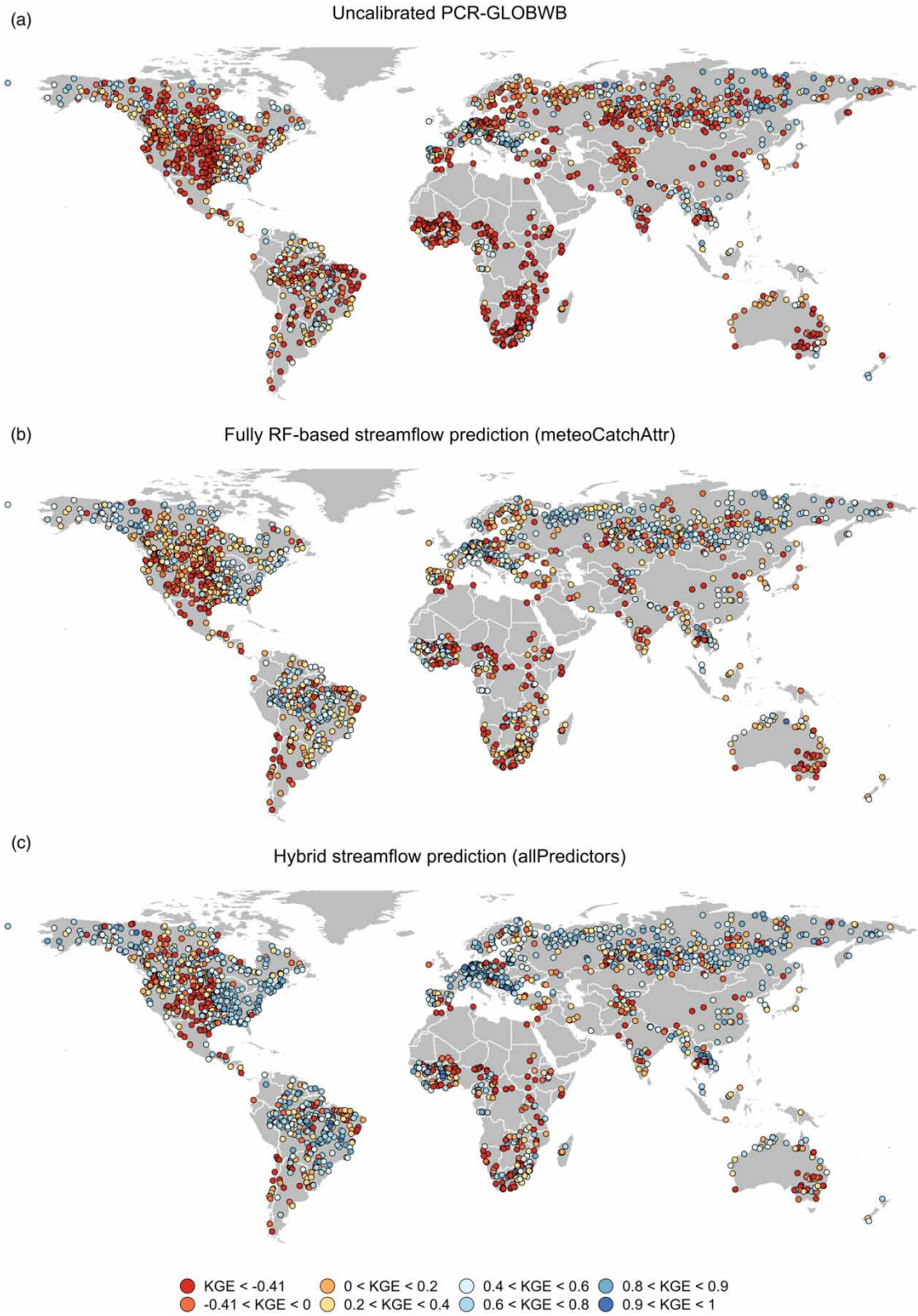

Figure 6 | KGE of the testing datasets, (a) for the uncalibrated PCR-GLOBWB; (b) for the RF streamflow prediction that only employs meteorological inputs and catchment attributes (*meteoCatchAttr*); and (c) for the hybrid streamflow model (*allPredictors*). If a station was used more than once for validation, the resulting KGE values were averaged over all subsamples in which it appeared. KGE < -0.41 is considered a bad performance.

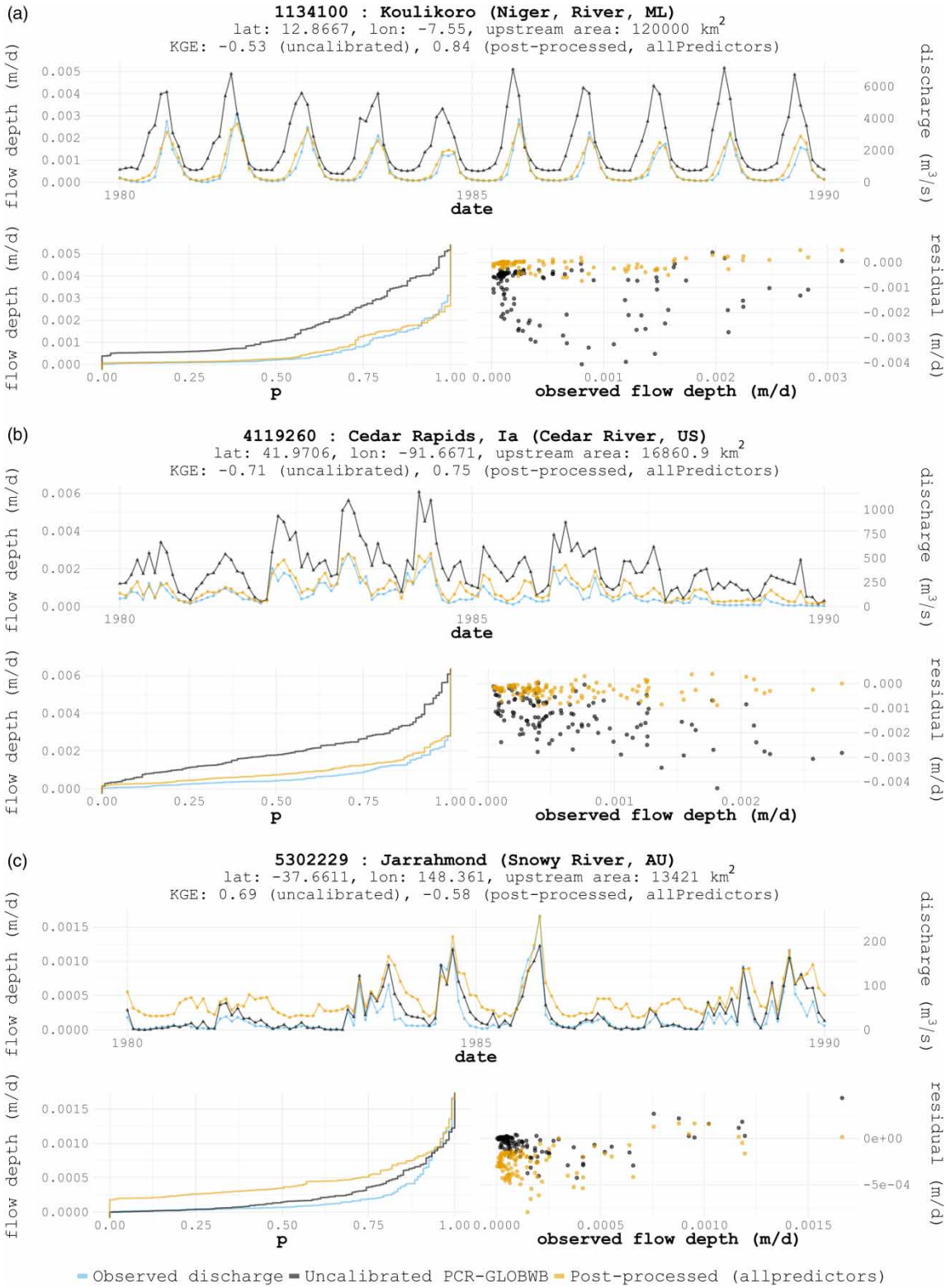

Figure 7 | Hydrographs, flow duration curves and residuals at three selected stations from the validation datasets, comparing streamflow simulations from uncalibrated PCR-GLOBWB (in black) and after post-processing (in yellow) against observed values of river discharge (in blue). *p* is the cumulative probability of streamflow magnitude. (a,b) Selected stations with good model performance. (c) Example of station with degrading KGE after post-processing.

5. DISCUSSION

5.1. Prediction in ungauged basins

Our results show that including hydrological state variables and simulated river discharge, in addition to meteorological inputs and catchment attributes, as predictors to the RF improves streamflow simulations better than a setup that excludes outputs from the hydrological model. This can be seen in the evident improvements to KGE values between the *meteoCatchAttr* and the *allPredictors* configurations, which, respectively, showed a median KGE of 0.34 and 0.51. These findings reconcile the long-standing divisions between fully data-driven and process-based hydrological models (Todini 2007), demonstrating that the two approaches can enhance each other's strengths and aid their respective weaknesses.

The current study aimed at properly addressing prediction in ungauged basins (PUB), which are characterized by scarce, inadequate, or absent data records of hydrological observations (Sivalapan *et al.* 2003). The issue of PUB, which is essentially a problem of extrapolation of universal hydrological laws (Sivapalan 2005), has been addressed with a variety of approaches (Wagener & Montanari 2011), though it is unclear which method works best, due to varying results across different areas (Singh *et al.* 2014).

Here, the use of SL algorithms enabled efficient extraction of information from large amounts of data, showing promising results in addressing PUB (Besaw *et al.* 2010; Kratzert *et al.* 2019a). The non-linear nature of SL methods is highly effective at regionalizing hydrological parameters and characterizing universal catchment behaviours, given sufficient training data (Kratzert *et al.* 2019b; Prieto *et al.* 2019; Potdar *et al.* 2021). This results in widespread generalization capabilities of the RF in ungauged basins, causing the spatial variability of physical catchment characteristics not to be an issue as with a classical calibration approach (Kling & Gupta 2009; Hrachowitz *et al.* 2013). This additionally enables post-processing across a variety of catchments and regions, a procedure that has been relatively little explored until now (Matthews *et al.* 2022).

In our case, the model was previously applied on a station-by-station approach, both for training and validation purposes (Shen *et al.* 2022a). This was done with a time-based split sampling, where the two halves of the available observations were, respectively, used for training and testing. However, this approach has been shown to produce inferior performance in the validation phase, compared to calibrating to the full available observations, which also eliminates subjective decisions on the calibration period (Shen *et al.* 2022b). Extending such a methodology to the global scale was not feasible, as many stations may only have a few timesteps of observations available compared to the runs of the hydrological model.

Thus, to be able to transfer hydrological knowledge between catchments, we extended the training dataset to include all streamflow observations in each training subset, i.e. all training stations contributed to the RF at all times when river discharge observations were available, to maximize extraction of global-scale knowledge on error in PCR-GLOBWB simulations. Even though this may generate a bias in the training sets, due to some regions contributing more than others, we assumed that an acceptable variety of catchment behaviours was represented.

In our modelling framework, the testing sets in each subsample are treated as ungauged basins, where the RF post-processor is applied to verify its capabilities in accurately predicting river discharge. The results presented here demonstrate a hybrid setup enables the RF post-processor to discern between a multitude of climates and catchment behaviours, making it a viable option to address PUB. This is encouraged by the fact that stations where runoff simulations improved were missing on average 41% of data over the whole modelling period (1979–2019), flagging that performance boosts were independent of the availability of the streamflow data at a particular location. Finally, the SL algorithm was also often capable of correcting for autocorrelation and heteroscedasticity, in addition to properly reproducing streamflow quantiles, whenever there was training data of sufficient quality.

5.2. Current limitations and future research

Notwithstanding the good results achieved by the RF-based post-processor, the modelling framework is still subject to a variety of limitations that require further examination and testing. For instance, 15–20% of stations in the various testing subsets are still characterized by a KGE < -0.41, often caused by an already poor performance in PCR-GLOBWB (e.g. central United States, south-east Australia). Since no single factor was found to strongly correlate to these negative performances, future analysis should focus on clustering these stations to understand if there are any shared causes of poor results.

We hypothesize these may stem from the inherent gaps of PCR-GLOBWB in capturing anthropogenic modifications of hydrological processes in arid environments (Smakhtin 2001; Shen & Chen 2010) or significant changes in the water balance of rivers due to inter-basin water transfers (Shumilova *et al.* 2018; Siddik *et al.* 2023). Although PCR-GLOBWB does include

a module for calculating the storage of lakes and reservoirs, the corresponding predictor variable (*surfaceWaterStorage*) was not found amongst the most important ones, which may be due to its highly simplified nature.

Another potentially related limitation in our study is the lack of a calibrated version of PCR-GLOBWB at the global scale as further benchmark against our hybrid model. This may constitute additional material for future research, which should primarily focus on the calibration of stations that have lower KGE than −0.41, both in the uncalibrated PCR-GLOBWB and after post-processing. Such a calibration could enhance the information contained in PCR-GLOBWB outputs, thereby potentially improving the generalization capabilities of the hybrid framework and runoff prediction in (semi-)ungauged catchments.

Moreover, the RF post-processor degraded PCR-GLOBWB performance at ~20% of stations, a sign that the SL algorithm does not always recognize when the hydrological model is already in a good state. At these stations, median KGE was downgraded from 0.50 to 0.23, which is, however, negligible in comparison to the increase from −0.19 to 0.57 at locations where performance was boosted. Results may also be biased on the training datasets, which would surely benefit from a higher length of the available observations. This can be achieved by extending PCR-GLOBWB runs to years prior to 1979, since many stations are characterized by a decline in observations in more recent years, as also noted by Ruhi *et al.* (2018).

Results would also benefit from a larger number of catchments, which can be included by running PCR-GLOBWB in its version at a resolution of 5 arcmin (~10 km at the equator), enabling the use of additional catchments for training (5272 GRDC stations have an upstream area between 400 and 10,000 km^2). This would also potentially increase the scaling capabilities of the RF post-processor by using a dataset with a higher variety of catchment sizes (Peters-Lidard *et al.* 2017; Sidle 2021; Tsai *et al.* 2021).

Further developments of the modelling framework may also focus on including additional predictors of catchment characteristics, especially extending feature extraction from PCR-GLOBWB parameter maps that were not included in this study, but also potentially through the use of external datasets at matching resolutions, e.g. HydroATLAS (Linke *et al.* 2019). Even though the exclusion of lagged time-variant predictors did not hinder good model results, it may still be significant to execute a feature extraction from PCR-GLOBWB time-series, as delineated in Papacharalampous *et al.* (2021) and Papacharalampous & Tyralis (2022). In addition to these potential experiments, the current framework could be adjusted and transferred to post-process streamflow simulations from another regional or global-scale process-based hydrological model, if tabular time-series of (at least) input meteorological variables, comparable catchment parameters and simulated discharge are available.

Finally, it is important that other tree-based algorithms are explored for hybrid streamflow modelling, especially more efficient ones and capable of handling the presence of missing values in the input predictors, e.g. XGBoost (Chen & Guestrin 2016), but also entirely new hydrological theories based on non-linearity and systems complexity, to develop more parsimonious process descriptions for predictive hydrological modelling, at multiple scales and in different biomes (Sivalapan *et al.* 2003).

6. CONCLUSION

In this study, we have presented a novel hybrid post-processing framework to address uncertainty in streamflow modelling. Static catchment attributes, meteorological input, hydrological state variables and simulated runoff from the global hydrological model PCR-GLOBWB were used as predictors of observed river discharge at the global scale for a random forest regressor.

Results show that the framework is capable of correcting bias and heteroscedasticity, and improving linear correlation between observed and simulated streamflow, by distinguishing different hydroclimatic conditions and a variety of river flow responses. While the median KGE for the *uncalibrated* PCR-GLOBWB was −0.03, *meteoCatchAttr* and *qMeteoStatevars* improved the median KGE to 0.34 and 0.37, respectively. However, combining all selected features as input to the RF (*allPredictors* setup) ultimately improved KGE at ~80% of the locations in the validation datasets. The cumulative median KGE for the hybrid setup (*allPredictors*) was 0.51, showing that the use of predictor values from the hydrological model outputs improved streamflow predictions at unseen locations.

These findings prove that the SL post-processor captures hidden non-linear dynamics between model components, to properly correct the uncertainties in a process-based hydrological model. Moreover, it reduces the need for calibration and the poor performances associated with the indiscriminate transfer of parameters sets between catchments, thus making it a potential candidate for improving streamflow prediction in ungauged basins.

Future research efforts to improve the current framework should focus on: (1) testing additional predictors of catchment and/or time-series behaviour; (2) increasing the spatial resolution of PCR-GLOBWB and the length of available simulations, to expand the current dataset of streamflow observations; (3) calibrating PCR-GLOBWB and (4) cluster analysing stations where the hybrid framework is not capable of improving performance above the mean benchmark, to understand any potentially shared reason of poor generalization.

Overall, the hybrid framework presented here works consistently across scales and climates, holding promises to bridge the gap between process-based and data-driven methods in streamflow prediction, by enhancing the information contained in both approaches. However, it also highlights the deficiencies of the two, such as ignorance about physical processes, poor representation of human impacts on catchments and bias on available data, further implying the urgency of exploring bolder hydrological theories to properly capture runoff formation at multiple scales.

ACKNOWLEDGEMENTS

The results presented in this work were made possible by the CPU cluster *velocity* at Utrecht University. The authors would like to sincerely thank ICT developer Dr. Oliver Schmitz for the patience and understanding while we were running intensive computations. We also thank the anonymous reviewers that took the time to read our work and helped improve the overall quality of the manuscript.

DATA AVAILABILITY STATEMENT

Please refer to the online version of this article for figures in colour. Uncompressed figures can be found at https://doi.org/10.5281/zenodo.8220029. Additional relevant data to the manuscript are included in the Supplementary Materials. Data and source code of the post-processing framework developed here are available at https://github.com/m7351gn/PCR-GLOBWB-RF. Model code of PCR-GLOBWB can be found at https://github.com/UU-Hydro/PCR-GLOBWB_model. Model inputs and outputs of the hybrid framework can be found at https://doi.org/10.5281/zenodo.7890583 and https://doi.org/10.5281/zenodo.7891352, respectively. All hydrographs produced during validation can be found as a supplement at https://doi.org/10.5281/zenodo.7893903.

AUTHORSHIP STATEMENT

D.K., E.H.S. and M.M. conceptualized and designed the study. E.H.S. ran PCR-GLOBWB simulations and provided process-based outputs. M.M. and E.H.S. contributed to data acquisition. M.M. and Y.S. developed the modelling framework. M.M. drafted the manuscript, which was edited and approved with input by all authors.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E. & Wilby, R. L. 2012 Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography: Earth and Environment* **36** (4), 480–513. https://doi.org/10.1177/0309133312444943.

Besaw, L. E., Rizzo, D. M., Bierman, P. R. & Hackett, W. R. 2010 Advances in ungauged streamflow prediction using artificial neural networks. *Journal of Hydrology* **386** (1–4), 27–37. https://doi.org/10.1016/j.jhydrol.2010.02.037.

Beven, K. J. 2012 Parameter estimation and predictive uncertainty. In: *Rainfall-Runoff Modelling: The Primer*. Wiley, New York, pp. 231–287.

Bogner, K., Liechti, K. & Zappa, M. 2016 Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. *Water* **8** (4), 115. https://doi.org/10.3390/w8040115.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32. https://doi.org/10.1023/a:1010933404324.

Chen, T. & Guestrin, C. 2016 XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785.

de Graaf, I., van Beek, L., Wada, Y. & Bierkens, M. 2014 Dynamic attribution of global water demand to surface water and groundwater resources: effects of abstractions and return flows on river discharges. *Advances in Water Resources* **64**, 21–33. https://doi.org/10.1016/j.advwatres.2013.12.002.

Desai, S. & Ouarda, T. B. 2021 Regional hydrological frequency analysis at ungauged sites with random forest regression. *Journal of Hydrology* **594**, 125861. https://doi.org/10.1016/j.jhydrol.2020.125861.

Dramsch, J. S. 2020 70 years of machine learning in geoscience in review. *Machine Learning in Geosciences* 1–55. https://doi.org/10.1016/bs.agph.2020.08.002.

Evin, G., Thyer, M., Kavetski, D., McInerney, D. & Kuczera, G. 2014 Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research* **50** (3), 2350–2375. https://doi.org/10.1002/2013wr014185.

Ghaith, M., Siam, A., Li, Z. & El-Dakhakhni, W. 2020 Hybrid hydrological data-driven approach for daily streamflow forecasting. *Journal of Hydrologic Engineering* **25** (2). https://doi.org/10.1061/(asce)he.1943-5584.0001866.

Ghazikhani, A., Babaeian, I., Gheibi, M., Hajiaghaei-Keshteli, M. & Fathollahi-Fard, A. M. 2022 A smart post-Processing system for forecasting the climate precipitation based on machine learning computations. *Sustainability* **14** (11), 6624. https://doi.org/10.3390/su14116624.

GRDC, Global Runoff Data Centre. The Global Runoff Database and River Discharge Data, 56068 Koblenz, Germany. Available from: https://www.bafg.de/GRDC (accessed 20 May 2022).

Gregorutti, B., Michel, B. & Saint-Pierre, P. 2016 Correlation and variable importance in random forests. *Statistics and Computing* **27** (3), 659–678. https://doi.org/10.1007/s11222-016-9646-1.

Grömping, U. 2009 Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician* **63** (4), 308–319. https://doi.org/10.1198/tast.2009.08199.

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

Haddeland, I., Heinke, J., Biemans, H., Eisner, S., Flörke, M., Hanasaki, N., Konzmann, M., Ludwig, F., Masaki, Y., Schewe, J., Stacke, T., Tessler, Z. D., Wada, Y. & Wisser, D. 2013 Global water resources affected by human interventions and climate change. *Proceedings of the National Academy of Sciences* **111** (9), 3251–3256. https://doi.org/10.1073/pnas.1222475110.

Hauswirth, S. M., Bierkens, M. F., Beijk, V. & Wanders, N. 2021 The potential of data driven approaches for quantifying hydrological extremes. *Advances in Water Resources* **155**, 104017. https://doi.org/10.1016/j.advwatres.2021.104017.

Herath, H. M. V. V., Chadalawada, J. & Babovic, V. 2021 Hydrologically informed machine learning for rainfall–runoff modelling: towards distributed modelling. *Hydrology and Earth System Sciences* **25** (8), 4373–4401. https://doi.org/10.5194/hess-25-4373-2021.

Hopson, T. M., Wood, A. W. & Weerts, A. H. 2018 Motivation and overview of hydrological ensemble post-processing. In: *Handbook of Hydrometeorological Ensemble Forecasting*, pp. 1–11. https://doi.org/10.1007/978-3-642-40457-3_36-2

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E. & Cudennec, C. 2013 A decade of predictions in ungauged basins (PUB) – a review. *Hydrological Sciences Journal* **58** (6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183.

Karssenberg, D., Schmitz, O., Salamon, P., de Jong, K. & Bierkens, M. F. 2010 A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environmental Modelling & Software* **25** (4), 489–502. https://doi.org/10.1016/j.envsoft.2009.10.004.

Kling, H. & Gupta, H. 2009 On the development of regionalization relationships for lumped watershed models: the impact of ignoring sub-basin scale variability. *Journal of Hydrology* **373** (3–4), 337–351. https://doi.org/10.1016/j.jhydrol.2009.04.031.

Knoben, W. J. M., Freer, J. E. & Woods, R. A. 2019 Technical note: inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences* **23** (10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019.

Kraft, B., Jung, M., Körner, M., Koirala, S. & Reichstein, M. 2022 Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences* **26** (6), 1579–1614. https://doi.org/10.5194/hess-26-1579-2022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S. & Nearing, G. S. 2019a Toward improved predictions in ungauged basins: exploiting the power of machine learning. *Water Resources Research* **55** (12), 11344–11354. https://doi.org/10.1029/2019wr026065.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. 2019b Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences* **23** (12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Lange, S., Menz, C., Gleixner, S., Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Müller Schmied, H., Hersbach, H., Buontempo, C. & Cagnazzo, C. 2021 WFDE5 over land merged with ERA5 over the ocean (W5E5 v2.0). *ISIMIP Repository*. https://doi.org/10.48364/ISIMIP.342217.

Li, M., Wang, Q. J., Bennett, J. C. & Robertson, D. E. 2016 Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. *Hydrology and Earth System Sciences* **20** (9), 3561–3579. https://doi.org/10.5194/hess-20-3561-2016.

Li, X., Sha, J. & Wang, Z. L. 2019 Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal* **64** (15), 1857–1866. https://doi.org/10.1080/02626667.2019.1680846.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F. & Thieme, M. 2019 Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data* **6** (1). https://doi.org/10.1038/s41597-019-0300-6.

Liu, Y. & Gupta, H. V. 2007 Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. *Water Resources Research* **43** (7). https://doi.org/10.1029/2006wr005756.

Matthews, G., Barnard, C., Cloke, H., Dance, S. L., Jurlina, T., Mazzetti, C. & Prudhomme, C. 2022 Evaluating the impact of post-processing medium-range ensemble streamflow forecasts from the European Flood Awareness System. *Hydrology and Earth System Sciences* **26** (11), 2939–2968. https://doi.org/10.5194/hess-26-2939-2022.

McMillan, H. K., Westerberg, I. K. & Krueger, T. 2018 Hydrological data uncertainty and its implications. *WIREs Water* **5** (6). https://doi.org/10.1002/wat2.1319.

Moges, E., Demissie, Y., Larsen, L. & Yassin, F. 2020 Review: sources of hydrological model uncertainties and advances in their analysis. *Water* **13** (1), 28. https://doi.org/10.3390/w13010028.

Montanari, A., Shoemaker, C. A. & van de Giesen, N. 2009 Introduction to special section on uncertainty assessment in surface and subsurface hydrology: an overview of issues and challenges. *Water Resources Research* **45** (12). https://doi.org/10.1029/2009wr008471.

Montanari, A., Young, G., Savenije, H., Hughes, D., Wagener, T., Ren, L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D. A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z. & Belyaev, V. 2013 'Panta Rhei – Everything Flows': change in hydrology and society – the IAHS scientific decade 2013–2022. *Hydrological Sciences Journal* **58** (6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088.

Mosaffa, H., Sadeghi, M., Mallakpour, I., Naghdyzadegan Jahromi, M. & Pourghasemi, H. R. 2022 Application of machine learning algorithms in hydrology. *Computers in Earth and Environmental Sciences* 585–591. https://doi.org/10.1016/b978-0-323-89861-4.00027-0

Mosavi, A., Ozturk, P. & Chau, K. W. 2018 Flood prediction using machine learning models: literature review. *Water* **10** (11), 1536. https://doi.org/10.3390/w10111536.

Nash, J. & Sutcliffe, J. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10** (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C. & Gupta, H. V. 2021 What role does hydrological science play in the age of machine learning? *Water Resources Research* **57** (3). https://doi.org/10.1029/2020wr028091.

Papacharalampous, G. & Tyralis, H. 2022 Time series features for supporting hydrometeorological explorations and predictions in ungauged locations using large datasets. *Water* **14** (10), 1657. https://doi.org/10.3390/w14101657.

Papacharalampous, G., Tyralis, H., Papalexiou, S. M., Langousis, A., Khatami, S., Volpi, E. & Grimaldi, S. 2021 Global-scale massive feature extraction from monthly hydroclimatic time series: statistical characterizations, spatial patterns and hydrological similarity. *Science of The Total Environment* **767**, 144612. https://doi.org/10.1016/j.scitotenv.2020.144612.

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T. E. & Woods, R. 2017 Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences* **21** (7), 3701–3713. https://doi.org/10.5194/hess-21-3701-2017.

Pham, L. T., Luo, L. & Finley, A. 2021 Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. *Hydrology and Earth System Sciences* **25** (6), 2997–3015. https://doi.org/10.5194/hess-25-2997-2021.

Potdar, A. S., Kirstetter, P. E., Woods, D. & Saharia, M. 2021 Towards predicting flood event peak discharge in ungauged basins by learning universal hydrological behaviors with machine learning. *Journal of Hydrometeorology*. https://doi.org/10.1175/jhm-d-20-0302.1

Prieto, C., le Vine, N., Kavetski, D., García, E. & Medina, R. 2019 Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. *Water Resources Research* **55** (5), 4364–4392. https://doi.org/10.1029/2018wr023254.

Roy, A., Kasiviswanathan, K. S., Patidar, S., Adeloye, A. J., Soundharajan, B. & Ojha, C. S. P. 2023 A novel physics-aware machine learning-based dynamic error correction model for improving streamflow forecast accuracy. *Water Resources Research* **59** (2). https://doi.org/10.1029/2022wr033318.

Ruhi, A., Messager, M. L. & Olden, J. D. 2018 Tracking the pulse of the earth's fresh waters. *Nature Sustainability* **1** (4), 198–203. https://doi.org/10.1038/s41893-018-0047-7.

Sharma, P. & Machiwal, D. 2021 *Advances in Streamflow Forecasting: From Traditional to Modern Approaches*, 1st edn. Elsevier, Amsterdam, The Netherlands.

Shen, Y. & Chen, Y. 2010 Global perspective on hydrology, water balance, and water resources management in arid basins. *Hydrological Processes* **24**, 129–135. https://doi.org/10.1002/hyp.7428.

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E. H. & Karssenberg, D. 2022a Random forests-based error-correction of streamflow from a large-scale hydrological model: using model state variables to estimate error terms. *Computers and Geosciences* **159**. https://doi.org/10.1016/j.cageo.2021.105019.

Shen, H., Tolson, B. A. & Mai, J. 2022b Time to update the split-sample approach in hydrological model calibration. *Water Resources Research* **58** (3). https://doi.org/10.1029/2021wr031523.

Shumilova, O., Tockner, K., Thieme, M., Koska, A. & Zarfl, C. 2018 Global water transfer megaprojects: a potential solution for the water-food-energy nexus? *Frontiers in Environmental Science* **6**. https://doi.org/10.3389/fenvs.2018.00150.

Siddik, M. a. B., Dickson, K. E., Rising, J., Ruddell, B. L. & Marston, L. 2023 Interbasin water transfers in the United States and Canada. *Scientific Data* **10** (1). https://doi.org/10.1038/s41597-023-01935-4.

Sidle, R. C. 2021 Strategies for smarter catchment hydrology models: incorporating scaling and better process representation. *Geoscience Letters* **8** (1). https://doi.org/10.1186/s40562-021-00193-9.

Singh, R., Archfield, S. & Wagener, T. 2014 Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments – a comparative hydrology approach. *Journal of Hydrology* **517**, 985–996. https://doi.org/10.1016/j.jhydrol.2014.06.030.

Sivalapan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O'Connel, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S. & Zehe, E. 2003 IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal* **48** (6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421.

Sivapalan, M. 2005 Pattern, process and function: elements of a unified theory of hydrology at the catchment scale. *Encyclopedia of Hydrological Sciences*. https://doi.org/10.1002/0470848944.hsa012.

Smakhtin, V. 2001 Low flow hydrology: a review. *Journal of Hydrology* **240** (3–4), 147–186. https://doi.org/10.1016/s0022-1694(00)00340-1.

Sood, A. & Smakhtin, V. 2015 Global hydrological models: a review. *Hydrological Sciences Journal* **60** (4), 549–565. https://doi.org/10.1080/02626667.2014.950580.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D. & Bierkens, M. F. P. 2018 PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development* **11** (6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D. & Bierkens, M. F. P. 2019 Input files underlying the publication 'PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model'. *4TU.ResearchData*. Dataset. https://doi.org/10.4121/uuid:e3ead32c-0c7d-4762-a781-744dbdd9a94b.

Todini, E. 2007 Hydrological catchment modelling: past, present and future. *Hydrology and Earth System Sciences* **11** (1), 468–482. https://doi.org/10.5194/hess-11-468-2007.

Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J. & Shen, C. 2021 From calibration to parameter learning: harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications* **12** (1). https://doi.org/10.1038/s41467-021-26107-z

Tyralis, H., Papacharalampous, G. & Langousis, A. 2019 A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **11** (5), 910. https://doi.org/10.3390/w11050910.

Valdez, E. S., Anctil, F. & Ramos, M. H. 2022 Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrology and Earth System Sciences* **26** (1), 197–220. https://doi.org/10.5194/hess-26-197-2022.

van Beek, L. P. H., Wada, Y. & Bierkens, M. F. P. 2011 Global monthly water stress: 1. Water balance and water availability. *Water Resources Research* **47** (7). https://doi.org/10.1029/2010wr009791.

Wada, Y., Wisser, D. & Bierkens, M. F. P. 2014 Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. *Earth System Dynamics* **5** (1), 15–40. https://doi.org/10.5194/esd-5-15-2014.

Wagener, T. & Montanari, A. 2011 Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research* **47** (6). https://doi.org/10.1029/2010wr009469.

WMO, World Meteorological Organization. International Organizations 2006 *Web Archive*. Available from: https://www.loc.gov/item/lcwaN0010741/.

Wright, M. N. & Ziegler, A. 2017 Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* **77** (1). https://doi.org/10.18637/jss.v077.i01

Xu, T. & Liang, F. 2021 Machine learning for hydrologic sciences: an introductory overview. *Wiley Interdisciplinary Reviews: Water* **8** (5), e1533. https://doi.org/10.1002/wat2.1533.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M. & Hinkelmann, R. 2021 Ensemble machine learning paradigms in hydrology: a review. *Journal of Hydrology* **598**, 126266. https://doi.org/10.1016/j.jhydrol.2021.126266.

First received 12 December 2022; accepted in revised form 25 July 2023. Available online 4 August 2023