Untangling hybrid hydrological models with explainable artificial intelligence

Daniel Althoffa,*, Helizani Couto Bazameb and Jessica Garcia Nascimentob

a Agricultural Engineering Department, Federal University of Viçosa (UFV), Av. Peter Henry Rolfs, s.n., Viçosa, MG 36570-900, Brazil
b Biosystems Engineering Department, College of Agriculture ‘Luiz de Queiroz’, University of São Paulo (ESALQ/USP), Av. Pâdua Dias, 11, Piracicaba, SP 13418-900, Brazil

*Corresponding author. E-mail: daniel.althoff@ufv.br

Abstract

Hydrological models are valuable tools for developing streamflow predictions in unmonitored catchments to increase our understanding of hydrological processes. A recent effort has been made in the development of hybrid (conceptual/machine learning) models that can preserve some of the hydrological processes represented by conceptual models and can improve streamflow predictions. However, these studies have not explored how the data-driven component of hybrid models resolved runoff routing. In this study, explainable artificial intelligence (XAI) techniques are used to turn a ‘black-box’ model into a ‘glass box’ model. The hybrid models reduced the root-mean-square error of the simulated streamflow values by approximately 27, 50, and 24% for stations 17120000, 27380000, and 33680000, respectively, relative to the traditional method. XAI techniques helped unveil the importance of accounting for soil moisture in hydrological models. Differing from purely data-driven hydrological models, the inclusion of the production storage in the proposed hybrid model, which is responsible for estimating the water balance, reduced the short- and long-term dependencies of input variables for streamflow prediction. In addition, soil moisture controlled water percolation, which was the main predictor of streamflow. This finding is because soil moisture controls the underlying mechanisms of groundwater flow into river streams.

Key words: GR4J, individual conditional explanations, LIME, partial dependence profiles, regression trees

Highlights

- A machine learning model is coupled to the GR4J hydrological model.
- The hybrid hydrological model consists of a single soil moisture accounting storage.
- The performance improvement is significant under low-flow conditions.
- Explainable artificial intelligence techniques are used for the global and local interpretation of the data-driven component of the model.

INTRODUCTION

Water scarcities, climate variabilities, and climate extremes have been a great concern in the planning and management of water resources in recent decades (Ni et al. 2020; Wyatt et al. 2020). Increasing water demands constantly threatens the planet’s security and sustainability. In this scenario, the
knowledge of streamflow in watersheds is essential to determining the availability of water over time. However, historical data alone have a limited value in supporting decision-making (Humphrey et al. 2016), and restrictions on gauge data in terms of density and frequency can affect the efficient management of water resources.

Alternatively, hydrological models enable accurate and reliable streamflow estimation and forecasting and play a crucial role in the proper management and allocation of water resources. Streamflow is a response to the physical interactions among climate, geology, topography, and land use in a watershed, but it is mainly a final derivation of rainfall (Poff et al. 1997). Thus, to estimate streamflow, models can use meteorological forcings along with physical and/or climatic parameters of watersheds, as well as historical data, as guides to solve empirical relationships.

Hydrological modeling is a challenging task due to the nonlinear and nonstationary nature of streamflow, especially under extreme events (high- and low-flow conditions). To obtain a reliable estimate, considerable efforts have been made to develop and apply different methods to improve the accuracy of such models. Typically, hydrological predictions are generated using models that consider a single meteorological force or multiple forces for simulating historical climatic scenarios and those of the present or the next few days, weeks, months, seasons, or even years (Sabzipour et al. 2020).

The models are based on methods that utilize various concepts, equations, parameters, and structures to describe the rainfall–runoff relationship (Tian et al. 2013). These models can be classified into two types. The first type includes process-based models in which a hydrological model is established to calculate the response of rainfall based on empirical equations (conceptual) or physical processes (Ni et al. 2020). These require the calibration of parameters to estimate, for example, evapotranspiration, infiltration, percolation, surface runoff, and other processes occurring in a basin (Kumanlioglu & Fistikoglu 2019). Regarding its disadvantages, these models have limitations due to the physical conditioning of parameters, high computational power requirements for spatially explicit operations, or the use of purely conceptual parameters that are not related to the hydrological properties of the watersheds. The second type includes data-driven models, which are based solely on observations. The modeling process becomes simpler and easier to implement, as it is based on observations/history and does not require information about physical processes (Liu et al. 2015). An advantage of these models, for example, machine learning models, compared to the more traditional approaches, is their flexible structure, which allows them to capture arbitrarily nonlinear and complex input–output relationships from data without any restrictive assumptions about the functional form of the underlying process (Humphrey et al. 2016). However, despite its appeal, the performance of machine learning models, as well as different statistical models, is highly dependent on the availability and quality of the observed data and does not provide any information regarding its hydrological processes, such as percolation and soil moisture.

In recent years, great effort has been made by scientists to combine machine learning hydrological models with conceptual models in which some hydrological processes are empirically described (Robertson et al. 2013; Humphrey et al. 2016; Ren et al. 2018; Kumanlioglu & Fistikoglu 2019). For instance, Ren et al. (2018) combined the HBV (Hydrologiska Byråns Vattenbalansavdelning) (Bergström 1992) model with artificial Bayesian neural networks, which enabled a better assessment of the uncertainties of hydrological predictions simulated on a monthly scale. Humphrey et al. (2016) combined artificial Bayesian neural networks with the intermediate products of the GR4J (Génie Rural à 4 paramètres Journalier) (Perrin et al. 2003) model to resolve runoff routing and estimate streamflow on a daily scale. Kumanlioglu & Fistikoglu (2019) also combined neural networks with outputs from the soil moisture accounting storage of the GR4J model to resolve runoff routing. These studies revealed the great potential of hybrid models in improving hydrological prediction. They did not, however, explore how the data-driven component of the models resolved the runoff routing processed and instead simply accepted them as ‘black-box’ models.
A subdomain of machine learning, explainable artificial intelligence (XAI), has recently received significant attention for helping its users to better understand how their ‘black-box’ models operate (Maksymiuk et al. 2020). The use of XAI techniques can extend the interpretability of machine learning models; therefore, the results can be better understood by humans. Thus, in this study, the objectives were (i) to assess the performance improvement in streamflow simulation resulting from using a hybrid (process-based/data-driven) hydrological model over a conceptual model and (ii) to explore a range of XAI techniques to uncover how runoff routing is being resolved and turn ‘black-box’ models into ‘glass box’ models.

MATERIALS AND METHODS

Study areas

The study areas consist of catchments corresponding to three gauge stations within the Brazilian Cerrado biome (Figure 1). The catchment areas of the corresponding gauge stations were delimited using the TauDEM toolbox (Tarboton 2008) and a Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) with a 90-m spatial resolution (Jarvis et al. 2008). The gauge stations are part of the Brazilian National Hydrometeorological Network (in Portuguese: Rede Hidrometeorológica Nacional), and their corresponding ID codes are 17120000, 27380000, and 33680000.

As part of the Brazilian Cerrado biome, these tropical watersheds are characterized by a high-energy input throughout the year and seasonal rainfall. The catchment areas from stations 17120000 and 27380000 present climates that, according to Köppen’s climate classification (Köppen 1936; Alvares et al. 2013), can be classified predominantly as tropical savannah climates.
with dry winters ($A_w$). Conversely, the catchment area from station 33680000 has portions characterized by a tropical savannah climate with a dry winter, a tropical savannah climate with dry summer ($A_s$), and a semiarid climate (BSh). The land-use and land-cover classes of the catchment areas are described in Supplementary Material Section S1.

**Hydrological models**

In this section, we describe a ‘traditional’ conceptual rainfall–runoff model, the GR4J (Perrin et al. 2003), and the hybrid rainfall–runoff model used in this study. GR4J is a daily lumped conceptual hydrological model (Figure 2(a)) based on four free parameters ($x_1$, $x_2$, $x_3$, and $x_4$) and is used in this study as a benchmark for the hybrid model to improve upon. The hybrid model, termed GR-Hyb here, consists of a process-based component, which shares a free parameter ($x_1$) with the GR4J model, and a data-driven-based component, which simulates streamflow using output from the process-based component (Figure 2(b)).

**The GR4J hydrological model**

As a lumped conceptual model, GR4J uses empirical relationships to model hydrological processes aggregated to the catchment scale (Figure 2(a)). In terms of meteorological forcings, it only requires as input the catchment average potential evaporation ($E$) and precipitation ($P$). The conceptual

![Figure 2](http://iwaponline.com/h2open/article-pdf/4/1/13/866287/h2oj0040013.pdf)
The structure of the GR4J model consists of two storages: the production storage \( S \) and the routing storage \( R \). In Figure 2, \( x_1 \) represents the size of the routing storage (mm), \( x_2 \) controls the flux to groundwater (mm), \( x_3 \) represents the size of the routing storage (mm), and \( x_4 \) controls the time base of the UH1 and UH2 unit hydrographs (days). After precipitation interception, if any, the current level of the production storage \( S \) regulates the fraction of precipitation that fills the production storage \( P_s \), water evaporating from the storage (actual evaporation, \( E_s \)), and water percolating from the storage (Perc). As it represents the soil water balance, the production storage is also referred to as the soil moisture accounting storage (Pushpalatha et al. 2011). The sum of precipitation that does not fill the production storage and percolation (Pr) is divided into two flow components according to a fixed split. Of the Pr, 90% is routed by a unit hydrograph (UH1) to the nonlinear routing storage, while the remaining 10% is routed by a unit hydrograph (UH2) to streamflow as direct flow \( Q_d \) after exchange with groundwater. Groundwater exchange is a function of \( x_2 \) (f(x2)). A negative \( x_2 \) value represents water infiltrating to the aquifer, while a positive value represents water exiting the aquifer and contributing to the routing storage and direct flow. The routing storage level \( R \) governs outflow \( Q_r \) and is able to simulate long streamflow recessions (Perrin et al. 2003).

### The hybrid hydrological model

The GR-Hyb model presents two components: a process-based component and a data-driven component. The process-based component of GR-Hyb is identical to the soil moisture accounting component of GR4J and consists of a single free parameter \( x_1 \). The data-driven component, on the other hand, is used to replace the routing component of GR4J. Thus, the hybrid model is able to simulate the soil water balance at the catchment scale and uses variables derived from this balance as input to the data-driven model to resolve routing. The reasoning behind combining process-based and data-driven components is that streamflow is generally considered an outcome of superimposed hydrological processes, for example, superimposing runoff contributions from distinct precipitation events. In addition, antecedent soil moisture is an important factor considered when calculating the event-based runoff coefficient (Song & Wang 2019). For these conditions to be captured by a machine learning model, long-time series are generally required as input; for example, Kratzert et al. (2019) showed that streamflow predictions may be influenced by over 100 previous timesteps. To account for the soil water balance at the catchment scale, machine learning models are provided with more information on the antecedent conditions; therefore, the data-driven component has a better ability to simulate streamflow.

The main differences between the GR-Hyb model presented here and the one presented by Kumanlioglu & Fistikoglu (2019) are with respect to the data-driven component and their parameter optimization techniques. In this study, cubist regression (Quinlan 1992, 1993) is used instead of a neural network to model streamflow. Cubist regression is an ensemble-like model based on the concept of regression trees (Breiman et al. 1984). Cubist regression builds rule-based regression trees in which the split criterion is used to test which of the predictor variables reaching a node can minimize the intrasubset standard deviation of the target values after the split. It adopts a boosting technique that helps adjust the predictions for observations that were poorly predicted in previous trees. Its regression trees are, however, different than traditional trees. Instead of discrete predictions, the final node of a tree presents linear models built considering the variables used in previous splits of the tree. The final prediction is smoothed by considering predictions made by the linear models built recursively up the tree (in previous splits). Cubist regression is chosen in this study for two reasons. First, it allowed other authors to achieve state-of-art performance within a range of studies (Noi et al. 2017; Worland et al. 2018; Filgueiras et al. 2019; Althoff et al. 2020b), and second, it provided a good trade-off between performance and computational time (Althoff et al. 2020b), which was
an important aspect when considering the optimization task. For more information on cubist regressions, refer to Quinlan (1993).

For parameter optimization, the particle swarm optimization technique, developed by Kennedy & Eberhart (1995) to mimic a simplified social model (bird flocking, fish schooling, and swarming theory), was used instead of the genetic algorithm to optimize \( x_1 \). Both optimization techniques have performed well in a range of previous studies. However, Panda & Padhy (2008) argued that from an evolutionary point of view, the performance of the particle swarm optimization technique is better than that of the genetic algorithm. Optimizing a single parameter seems to be a simple task. However, the cubist model calibration is independent of physical meanings in the outputs of the process-based component; therefore, calibration on highly varied parameter values can introduce equifinality. In this case, the particle swarm optimization technique was best to avoid local optima falling and find the parameter value resulting in the best outcome. The optimization algorithm was implemented using the model-independent software (Zambrano-Bigiarini & Rojas 2013).

By accounting for soil moisture, the model decreases its dependency on inputs from previous time-steps, as required by purely data-driven hydrological models (Kratzert et al. 2019; Kumanlioglu & Fistikoglu 2019). Nevertheless, conceptual models usually depend on previous time-steps. For example, in GR4J, the time base of the unit hydrograph UH2 (2 \( \times \) \( x_4 \)) results in fluxes exceeding a few consecutive days. Direct flow is then obtained by superimposing the fluxes from individual events. In Perrin et al. (2003), \( x_4 \) ranged from 1.1 to 2.9 (80% confidence interval) for a large sample of catchments. Thus, two versions of the hybrid model were assessed in this study. The first model, denoted GR-Hyb-1, considered only inputs from the respective day for the machine learning model, i.e., to predict \( Q_t \), only \( \text{Perc}_t \), \( \text{Es}_t \), and \( \text{Pn}_t \) are considered. The second model, denoted GR-Hyb-2, additionally considered inputs from previous days up to three time-steps, i.e., \( \text{Perc}_t \), \( \text{Perc}_{t-1} \), \( \text{Perc}_{t-2} \), \( \text{Perc}_{t-3} \), \( \text{Es}_t \), \( \text{Es}_{t-1} \), \( \text{Es}_{t-2} \), \( \text{Es}_{t-3} \), \( \text{Pn}_t \), \( \text{Pn}_{t-1} \), \( \text{Pn}_{t-2} \), and \( \text{Pn}_{t-3} \) were used to predict \( Q_t \).

The three input variables were selected based on model parsimony. The soil water balance input is the net precipitation, and the outputs are the actual evaporation and percolation. However, part of the net precipitation used to fill the storage can be inferred by the final water percolation. Thus, we allow the model handle net precipitation alone to predict its contribution to runoff. The inputs from three previous time-steps were used in the GR-Hyb-2 version considering the \( x_4 \) ranges found by Perrin et al. (2003). \( x_4 \) represents the time base of the unit hydrograph and the possible number of time-steps, where fluxes are superimposed for runoff generation.

**Meteorological forcing data**

The study period considered for this study ranged from 1 June 2000 to 31 December 2014 (~14.5 years). The meteorological forcings required by the models are precipitation (\( P \)) and potential evaporation (\( E \)). The datasets used to represent these variables were the GPM IMERG Final Precipitation V06 1 day (Huffman et al. 2019) and the ETo-Brazil V3 (Althoff et al. 2020a). The IMERG dataset has been favorably assessed across the Brazilian territory, while ETo-Brazil is a daily gridded Penman–Monteith evapotranspiration dataset developed for Brazil using machine learning models. Although it differs from the potential evaporation (Miralles et al. 2020), the Penman–Monteith reference evapotranspiration is closely related, and the dataset was available for the entire study period and areas.

**Observed streamflow data**

Daily streamflow data for gauge stations 17120000, 27380000, and 33680000 were downloaded from HidroWeb (ANA 2020), which is a tool part of the Nacional System for Water Resources Information.
(in Portuguese: Sistema Nacional de Informações sobre Recursos Hídricos). The data were filtered for the same period as considered for the meteorological forcings.

Model calibration and validation

The observed streamflow and forcing data were split into warm-up, calibration, and validation periods. The warm-up period consisted of 2.5 years (2000–2002), while the remaining 12 years of records were split in a 75:25 ratio for calibration–validation, i.e., 9 years for calibration (2003–2011) and the 3 most recent years for validation (2012–2014). Considering the strong seasonality of streamflow in the region and the importance of water availability during the dry season, it is essential to choose an objective function that gives proper weight to low-flow conditions during calibration.

It is well documented that the commonly used Nash–Sutcliffe efficiency index (NSE) (Nash & Sutcliffe 1970) poorly assesses the model’s performance under low-flow conditions (Legates & McCabe 1999; Krause et al. 2005; Schaeffli & Gupta 2007; Gupta et al. 2009). Thus, we used the Kling–Gupta efficiency (KGE) index (Gupta et al. 2009):

\[
KGE = 1 - ED, \quad -\infty < KGE \leq 1
\]

with:

\[
ED = \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}
\]

where ED is the Euclidian distance, \(r\) is the linear correlation coefficient between the observed \((O_i)\) and predicted \((P_i)\) streamflow, \(\alpha\) is the variability ratio or ratio between the standard deviation of simulated values and the standard deviation of observed values \((\sigma_P/\sigma_O)\), and \(\beta\) is the ratio between the mean simulated and observed values \((\mu_P/\mu_O\) or \(\bar{P}/\bar{O}\)).

The performance of the model was also evaluated by considering the following metrics: percent bias (PBIAS, Equation (2)), mean absolute error (MAE, Equation (3)), root-mean-square error (RMSE, Equation (4)), NSE (Equation (5)), and relative NSE (rNSE, Equation (6)), i.e., NSE computed with relative values:

\[
PBIAS = \frac{\sum_{i=1}^{n} (P_i - O_i)}{\sum_{i=1}^{n} O_i}, \quad -\infty < PBIAS < +\infty
\]

\[
MAE = \frac{\sum_{i=1}^{n} |P_i - O_i|}{n}, \quad 0 \leq MAE < +\infty
\]

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}, \quad 0 \leq RMSE < +\infty
\]

\[
NSE = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2}, \quad -\infty < NSE \leq 1
\]

\[
rNSE = 1 - \frac{\sum_{i=1}^{n} \left( \frac{O_i - P_i}{O_i} \right)^2}{\sum_{i=1}^{n} \left( \frac{O_i - \bar{O}}{O_i} \right)^2}, \quad -\infty < rNSE \leq 1
\]
where $O_i$ and $P_i$ are the observed and predicted (simulated) streamflows at timestep $i$, the overbar denotes the mean for the entire period, for example, $\bar{O}$ is the mean observed streamflow, and $n$ denotes the total number of timesteps.

PBIAS is a simple average of the prediction error by the average observed streamflow and indicates whether a model is generally under- or overestimating observed values. MAE indicates the average magnitude of the prediction error, while RMSE is penalized by larger errors. NSE is a commonly used goodness-of-fit criterion in many studies, which penalizes high error dispersion in relation to observed streamflow variability. rNSE provides a better assessment of streamflow estimates under low-flow conditions because errors in low-flow conditions tend to be relatively larger. The ideal value for PBIAS, MAE, and RMSE is 0, while that for KGE, NSE, and rNSE is 1.

**Explanation tasks for machine learning models**

Because of their inner complexity, machine learning models are often considered ‘black-boxes’. Even though the hybrid models are built based on cubist regression, which is based on more interpretable units (decision trees), the number of rules and linear regressions and its boosting-like scheme used for nonlinear predictions comes at the expense of interpretability. Recent advances made within XAI, a subdomain of machine learning, can aid its users in better understanding and reporting how their ‘black-box’ models function on a global scale, as well as allowing researchers to zoom into specific cases (Maksymiuk et al. 2020). To gain some insights into the models developed, we apply two global and two local interpretation techniques.

**Global interpretation**

The most common methods used for global interpretation include (i) variable/feature importance (VI/FI) measures and (ii) partial dependence profiles (PDPs) (Friedman 2001). VI is an explanation task based on model parts and requires only the trained model. For the cubist regression, the VI is calculated based on variable usage statistics, i.e., percentage of times each variable is used in a condition and/or in a linear model (Kuhn et al. 2018). At each split of the tree, cubist regression saves a linear model after feature selection. This linear model may utilize any variable used in the current split or the splits above it. The usage statistics reflect all the models involved in the predictions. The PDP technique, on the other hand, is an explanation task based on the model profile. This method profiles the global behavior of models within the context of a single variable, i.e., it presents the expected prediction as a function of a specific variable (Maksymiuk et al. 2020). Thus, records from the validation period were used for the PDP. The PDP, unlike the VI technique, is a model-agnostic method that does not require the knowledge of the model’s internal structure. The VI and PDP techniques were implemented in the R programming language and environment (R Core Team 2020) and using the caret package (Kuhn et al. 2018) and the pdp package (Greenwell 2018).

**Local interpretation**

For the local interpretation, two model-agnostic techniques were used; one, the individual conditional explanation (ICE) (Goldstein et al. 2015) was used to predict profiles, and the other, the local interpretable model-agnostic explanation (LIME) (Ribeiro et al. 2016) was used to analyze specific cases. ICE, also known as the Ceteris Paribus for profiles, iteratively examines the influence of each explanatory variable while holding the values of all other variables unchanged. The main difference between ICE and PDP is that there is no aggregation over the entire dataset, as it presents the perspective over single observations (Maksymiuk et al. 2020). The ICE profiles were also generated for the records from the validation period and implemented in R using the pdp package (Greenwell 2018).
A closer look was given to five different cases over a year using the LIME algorithm. The LIME explanation task is a technique that predicts parts, i.e., it attempts to explain local predictions. It assumes that every complex model is linear on a local scale and can be explained by a simple model fit around a single observation. The LIME algorithm first creates a (i) large number of permutations, i.e., replicate feature data with slight value modifications. It computes a (ii) similarity distance measured between the original observations and permutations and applies the (iii) original model to predict the outcomes of the permuted data. It then (iv) selects features that best describe the predicted outcomes and (v) fits a simple model to the permuted data weighted by its similarity to the original observation. The feature weights of this simple model are used to explain local behavior. The dates chosen were 25 January 2012, 8 April 2012, 22 June 2012, 5 September 2012, and 19 November 2012, which are all part of the validation period. These dates were chosen to represent the different contexts of streamflow prediction throughout the year.

The LIME algorithm was implemented using the lime package (Pedersen & Benesty 2019). In the LIME algorithm, (i) the number of permutations was set to 5,000, (ii) the similarity was computed using Gower’s distance (Gower 1971), (iii) the trained cubist model was used to predict outcomes from permutations, (iv) the forward selection was used for feature selection, and (v) the ridge regression was fitted to the weighted permuted observations.

RESULTS AND DISCUSSION

Performance of the hydrological models

A brief overview of the hydrological context of the studied catchments is provided in Supplementary Material Section S2, and a discussion regarding calibrated parameters is provided in Supplementary Material Section S3. The time series and scatterplots between the observed and predicted streamflows are presented in Figures 3 and 4, respectively. In general, the hybrid models presented similar

![Figure 3](http://iwaponline.com/h2open/article-pdf/4/1/13/866287/h2oj0040013.pdf)  
**Figure 3** | Time series of observed and predicted streamflow during the validation period.
performances and were superior to the GR4J model, which demonstrates that the association of the GR4J model with machine learning techniques improved the prediction of daily flow in the three catchment areas analyzed.

The hybrid models performed remarkably better under low-flow conditions and in transitions between wet and dry seasons. For example, let us consider low-flow conditions roughly as flows below the 66% exceedance probability level and transition conditions between seasons as flows between the 33 and 66% exceedance probability levels. Under low-flow conditions, the GR-Hyb-2 model produced RMSEs of 0.11, 0.07, and 0.02 mm day\(^{-1}\) for catchments 17120000, 27380000, and 33680000, respectively, while GR4J produced RMSEs of 0.15, 0.30, and 0.03 mm day\(^{-1}\), respectively. During the transition periods, GR-Hyb-2 produced RMSEs of 0.18, 0.44, and 0.05 mm day\(^{-1}\) for catchments 17120000, 27380000, and 33680000, respectively, while GR4J produced RMSEs of 0.26, 0.51, and 0.06 mm day\(^{-1}\), respectively. The rNSE values are also displayed in Table 1. Overall, they also showed a general reduction in dispersion (Figure 4), i.e., higher values of the determination coefficient (\(R^2\)). The best alignment to the ideal 1:1 line (black dashed line) was also notable in which the slopes of both adjusted regressions (ordinary least squares and forced to the origin) were closer to 1.

A summary of the performance of the models is presented in Table 1. In the catchment areas of stations 17120000 and 27380000, the GR-Hyb-1 model produced PBIAS values of −4.5 and 3.2%, indicating an under- and overestimation, respectively, of the daily flows observed in these basins. In the catchment area of station 33680000, the GR4J model presented the lowest PBIAS value (−15.0%) and showed a tendency to underestimate the values of daily flow, especially that during the wet season. In contrast, the hybrid models tended to overestimate the streamflow in the same catchment. This overestimation occurred with a greater magnitude mainly during the wet period of 2014 (Figure 3). The hybrid models showed lower MAE and RMSE values in relation to the GR4J
model. The reduction in RMSE was approximately 27, 50, and 24% for stations 17120000, 27380000, and 33680000, respectively.

Based on the NSE, rNSE, and KGE performance criteria, the GR-Hyb-2 model presented their highest values in basin 17120000, while the GR-Hyb-1 model resulted in the best results (values closer to 1) for streamflow prediction at stations 27380000 and 33680000. A more objective way to analyze and compare the performance of these models is through a Taylor diagram (Taylor 2001; Figure 5). The metrics presented by this diagram were similar to the metrics present in the KGE calculation. With the exception of the bias, the graph also shows the same linear correlation (blue-dashed lines) and the variability ratio, which, in the case of the Taylor diagram, is referred to as the normalized standard deviation (black dash-dotted lines). In addition, the diagram also shows the centralized RMSE (green-dotted lines). The model that comes closest to 1 on the abscissa axis was considered the most suitable.

Table 1 | Goodness-of-fit criteria for the GR4J and hybrid hydrological models during validation

| Station ID | Model     | PBIAS | MAE  | RMSE  | NSE     | rNSE   | KGE    |
|------------|-----------|-------|------|-------|---------|-------|--------|
| 17120000   | GR4J      | – 11.7| 0.27 | 0.37  | 0.78    | 0.86  | 0.78   |
|            | GR-Hyb-1  | – 4.5 | 0.20 | 0.29  | 0.87    | 0.92  | 0.82   |
|            | GR-Hyb-2  | – 4.8 | 0.18 | 0.27  | 0.88    | 0.93  | 0.82   |
| 27380000   | GR4J      | 12.7  | 0.46 | 0.61  | 0.75    | – 8.89| 0.83   |
|            | GR-Hyb-1  | 3.2   | 0.21 | 0.32  | 0.93    | 0.37  | 0.95   |
|            | GR-Hyb-2  | 3.4   | 0.21 | 0.31  | 0.93    | 0.25  | 0.95   |
| 33680000   | GR4J      | – 15.0| 0.09 | 0.17  | 0.57    | 0.77  | 0.65   |
|            | GR-Hyb-1  | 18.0  | 0.07 | 0.13  | 0.74    | 0.86  | 0.78   |
|            | GR-Hyb-2  | 17.8  | 0.07 | 0.14  | 0.72    | 0.84  | 0.77   |

PBIAS, percent bias (%); MAE, mean absolute error (mm day$^{-1}$); RMSE, root-mean-square error (mm day$^{-1}$); NSE, Nash–Sutcliffe efficiency index; rNSE, relative Nash–Sutcliffe index; KGE, Kling–Gupta efficiency index.

Bold values highlight the best performing method with respect to each goodness-of-fit criterion for each station.

The proximity between the GR-Hyb-1 and GR-Hyb-2 models indicates that the inclusion of input variables with values from previous days was not an efficient strategy to better capture the temporal effects. Studies using purely statistical models show that this inclusion can be advantageous for flow estimates, especially when the streamflow variable is available (Adnan et al. 2020a, 2020b; Tyralis et al. 2020). In contrast, hybrid models have the advantage of having the water balance of storage.
$x_1$, which, in this study, seems to be sufficient to capture the temporal effects and antecedent water conditions of the catchment area. Thus, the greater complexity given to the model through the inclusion of observations from previous days is unnecessary.

**Explainable artificial intelligence**

It is clear from the VI plots (Figure 6) that percolation was by far the most important input variable to the cubist regression models. Even when input variables of lagged observations were included (GR-Hyb-2), the lagged percolation variables still presented a higher importance. The importance of the actual evaporation and net precipitation varied between the gauge stations and within the models applied to the same gauge station. For example, the actual evapotranspiration (VI: ~10%) ranked better than the net precipitation (VI: ~0%) at gauge station 17120000 in GR-Hyb-1 and was generally considered more important than net precipitation in GR-Hyb-2. On the other hand, despite the net precipitation (VI: ~50%) showing a high importance at gauge station 27380000 in GR-Hyb-1, its importance was lower than that of the actual evaporation in GR-Hyb-2, even for the variables with lagged observations.

Because the lagged net precipitation observations presented a higher importance in GR-Hyb-2 than the net precipitation of the current day, $P_{n_i}$ for stations 17120000 and 27380000 and $P_{n_{i-2}}$ for station 33680000 could be related to the unit hydrograph duration (parameter $x_4$, see Supplementary Material Section S3) in the GR4J model. For example, the calibrated $x_4$ parameter was higher for station 17120000 (6.25 days) and lower for stations 27380000 (2.45 days) and 33680000 (2.44 days). However, it was difficult to make such an affirmation since $x_4$ was not included in GR-Hyb-1 and GR-Hyb-2. In addition, despite the fact that GR-Hyb-1 used only input variables corresponding to the current day, both hybrid models were calibrated to similar $x_1$ values and performed correspondingly.
As both the GR-Hyb-1 and GR-Hyb-2 models showed similar performances, it seems reasonable to use the more parsimonious model, especially when the objective is to untangle the complexities of the ‘black-box’ models. For this reason, the remaining explanation tasks were applied only to the cubist regression of the GR-Hyb-1 models. The ICEs and PDPs obtained from GR-Hyb-1 are shown in Figure 6. Despite showing a large VI for some of the models (Figure 7), the actual evaporation and net precipitation parameters show a limited influence on the prediction of streamflow. These variables seem to have only weighted a small number of instances, as shown by the black lines with gentler slopes. However, the overall dependence of streamflow on these variables was insignificant, as the PDPs (red lines) show almost flat slopes. In contrast, percolation was the largest driver of streamflow, especially for stations 17120000 and 27380000, accounting altogether for the baseflow and interflow components. The closer analysis provided by LIME on five different cases over a year (Supplementary Material Figure S2) shows that percolation has a higher absolute weight on the prediction for all cases. For all catchments, the weight was positive for the predictions performed during the wet season and negative during the dry season and under low-flow conditions. Thus, streamflow is, in most part, controlled by the production storage.

![Figure 7](image_url)

**Figure 7** | ICE profiles (black lines) and PDPs (red lines) for the hybrid hydrological model (GR-Hyb-1).

It may be surprising that the precipitation from the respective day, which can result in direct runoff, played such a small role in the streamflow prediction. However, a study considering a large number of catchments across Europe also showed that runoff anomalies correlated well with soil moisture anomalies (Ghajarnia et al. 2020). Ghajarnia et al. (2020) explain that the role of soil moisture, represented here by the production storage, is mechanistically explainable as it implies changes in the groundwater table depth and, therefore, in the hydraulic gradients governing groundwater flow toward streams. These findings are important in untangling how runoff routing is resolved and the key role that soil moisture should play in hydrological and earth system modeling.
CONCLUSIONS

In this study, we compared the performance of a conceptual hydrological model to that of a hybrid hydrological model. The hybrid hydrological model consisted of a conceptual-based component, in which a single parameter is used to calibrate the capacity of the production storage (soil moisture accounting), and a data-driven-based component, which uses water balance outputs as inputs for machine learning modeling to predict streamflow. In the data-driven component, two types of input were used: one that only considered variables of the respective day and another that considered variables from past days (lagged input).

The machine learning model benefitted from the inclusion of the soil moisture accounting storage and showed that including observations from the past day, as done by many studies of purely data-driven hydrological models, was unnecessary. Thus, the conceptual-based component used to account for soil moisture reduced the short- or long-term dependencies of the input variables used to make predictions. In addition, the hybrid models showed clear improvements compared to the conceptual model, especially for predictions made during the transition period between wet and dry seasons and during low-flow conditions. This improvement is important for watersheds with strong seasonality, as is the case with the tropical watersheds observed in this study.

A highlight from this study is the use of XAI techniques to untangle how the machine learning component of a hybrid hydrological model resolves runoff routing. These techniques showed that soil moisture is the most important factor in predicting streamflow. Soil moisture is represented in the hybrid models by the water level in the production storage and is responsible for the groundwater flow contribution to streamflow. These findings agree with large-scale studies that have unveiled a strong correlation between soil moisture and runoff.

FUNDING

This study was financed in part by the Coordination of Improvement of Higher Education Personnel (CAPES) – finance code 001 – and by the National Council for Scientific and Technological Development (CNPQ) – grant no. 142273/2019-8.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories (ETo-Brazil repository: https://data.mendeley.com/datasets/sstjw74ryh/3 and IMERG repository: https://gpm1.gesdisc.eosdis.nasa.gov/data/GPM_L3/GPM_3IMERGDF.06/).

REFERENCES

Adnan, R. M., Liang, Z., Parmar, K. S., Soni, K. & Kisi, O. 2020a Modeling monthly streamflow in mountainous basin by MARS, GMDH-NN and DENFIS using hydroclimatic data. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05164-3.

Adnan, R. M., Petroselli, A., Heddam, S., Santos, C. A. G. & Kisi, O. 2020b Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. Stochastic Environmental Research and Risk Assessment 1–20. https://doi.org/10.1007/s00477-020-01910-0.
Althoff, D., Dias, S. H. B., Filgueiras, R. & Rodrigues, L. N. 2020a ETo-Brazil: A Daily Gridded Reference Evapotranspiration Data Set for Brazil – Repository. Mendeley Data, V3. https://doi.org/10.17632/satjw74ryh.3.

Althoff, D., Dias, S. H. B., Filgueiras, R. & Rodrigues, L. N. 2020b ETo-Brazil: a daily gridded reference evapotranspiration data set for Brazil (2000–2018). Water Resources Research 56(7), https://doi.org/10.1029/2020WR027562.

Alvares, C. A., Stape, J. L., Sentellhas, P. C., de Moraes Gonçalves, J. L. & Sparovek, G. 2013 Köppen’s climate classification map for Brazil. Meteorologische Zeitschrift 22(6), 711–728. https://doi.org/10.1127/0941-2948/2013/0507.

ANA 2020 HidroWeb: Sistema de Informações Hidrológicas (Version 3.1.1). Agência Nacional de Águas (ANA). Available from: http://www.snirh.gov.br/hidroweb/. Accessed: 10 April 2020.

Bergström, S. 1992 The HBV Model-Its Structure and Applications. SMHI Reports Hydrology No. RH 4. SMHI, Sweden, p. 32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984 Classification and Regression Trees. Wadsworth & Brooks/Cole, Monterey, CA.

Friedman, J. H. 2001 Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Ghajarnia, N., Kalantari, Z., Orth, R. & Destouni, G. 2020 Close co-variation between soil moisture and runoff emerging from multi-catchment data across Europe. Scientific Reports 10(1), 4817. https://doi.org/10.1038/s41598-020-61621-y.

Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. 2015 Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24(1), 44–65. https://doi.org/10.1080/10618600.2014.907095.

Greenwell, B. 2018 Pdp: Partial Dependence Plots (Version 0.7.0). Available from: https://CRAN.R-project.org/package=pdp. Accessed: 12 November 2020.

Ghajarnia, N., Kalantari, Z., Orth, R. & Destouni, G. 2020 Close co-variation between soil moisture and runoff emerging from multi-catchment data across Europe. Scientific Reports 10(1), 4817. https://doi.org/10.1038/s41598-020-61621-y.

Ghajarnia, N., Kalantari, Z., Orth, R. & Destouni, G. 2020 Close co-variation between soil moisture and runoff emerging from multi-catchment data across Europe. Scientific Reports 10(1), 4817. https://doi.org/10.1038/s41598-020-61621-y.

Greenwell, B. 2018 pdp: Partial Dependence Plots (Version 0.7.0). Available from: https://CRAN.R-project.org/package=pdp. Accessed: 12 November 2020.

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(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J. & Tan, J. 2019 GPM IMERG Final Precipitation L3 1 Day 0.1 degree x 0.1 degree V06, Edited by Andrey Savtchenko. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD. Retrieved from: doi:10.5067/GPM/IMERG/DAY/2016.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C. & Maier, H. R. 2016 A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. Journal of Hydrology 540, 623–640. https://doi.org/10.1016/j.jhydrol.2016.06.026.

Jarvis, A., Reuter, H. I., Nelson, A. & Guevara, E. 2008 Hole-filled SRTM for the Globe Version 4. The CGIAR CSI SRTM 90 m Data Base, 15, 25–54. Available from: http://srtm.csi.cgiar.org. Accessed: 4 February 2020.

Kennedy, J. & Eberhart, R. 1995 Particle swarm optimization. In Proceedings of ICNN'95 – International Conference on Neural Networks. Vol. 4, pp. 1942–1948.

Küpper, W. 1936 Das geographische System der klima. In: Handbuch der klimatologie. Gebrüder Borntraeger, Berlin, Vol. part C, pp. 1–44. Available from: https://upload.wikimedia.org/wikipedia/commons/3/3c/Das_geographische_System_der_Klimate_%281936%29.pdf

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S. & Klambauer, G. 2019 NeuralHydrology – interpreting LSTM in hydrology. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R., eds). Springer International Publishing, Cham, pp. 347–362. https://doi.org/10.1007/978-3030-83954-6_19.

Krause, P., Boyle, D. P. & Bäse, F. 2005 Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 9, 89–97. https://doi.org/10.5194/adgeosci-5-89-2005.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. & Hunt, T. 2018 caret: Classification and Regression Training (Version 6.0-81). Available from: https://CRAN.R-project.org/package=caret. Accessed: 14 March 2019.

Kumanlioglu, A. A. & Fistikoglu, O. 2019 Performance enhancement of a conceptual hydrological model by integrating artificial intelligence. Journal of Hydrologic Engineering 24 (11), 04019047. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001850.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of ‘goodness-of-fit’ measures in hydrologic and hydroclimatic model validation. Water Resources Research 35 (1), 233–241. https://doi.org/10.1029/1998WR900018.

Liu, Z., Zhou, P., Chen, X. & Guan, Y. 2015 A multivariate conditional model for streamflow prediction and spatial precipitation refinement. Journal of Geophysical Research: Atmospheres 120(19). https://doi.org/10.1002/2015JD023787.

Maksymyk, S., Giosiewska, A. & Bicek, P. 2020 Landscape of R Packages for Explainable Artificial Intelligence. ArXiv:2009.15248 [cs, stat]. Available from: http://arxiv.org/abs/2009.15248. Accessed: 12 November 2020.

Miralles, D. G., Brutsaert, W., Dolman, A. J. & Gash, J. H. 2020 On the use of the term ‘evapotranspiration’. Water Resources Research 56(11), e2020WR028055. https://doi.org/10.1029/2020WR028055.

Nash, J. E. & Sutcliffe, J. V. 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.
Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J. & Liu, J. 2020 Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. Journal of Hydrology 586, 124901. https://doi.org/10.1016/j.jhydrol.2020.124901.

Noi, P. T., Degener, J. & Kappas, M. 2017 Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. Remote Sensing 9(5), 398. https://doi.org/10.3390/rs9050398.

Panda, S. & Padhy, N. P. 2008 Comparison of particle swarm optimization and genetic algorithm for FACTS-based controller design. Applied Soft Computing 8(4), 1418–1427. https://doi.org/10.1016/j.asoc.2007.10.009.

Pedersen, T. L. & Benesty, M. 2019 lime: Local Interpretable Model – Agnostic Explanations (Version 0.5.1). Available from: https://CRAN.R-project.org/package=lime. Accessed: 12 November 2020.

Perrin, C., Michel, C. & Andréassian, V. 2003 Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279(1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Stromberg, S. & Stromberg, J. C. 1997 The natural flow regime. BioScience 47(11), 769–784. https://doi.org/10.2537/1313099.

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T. & Andréassian, V. 2011 A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. Journal of Hydrology 411(1–2), 66–76. https://doi.org/10.1016/j.jhydrol.2011.09.034.

Quinlan, J. R. 1992 Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore.

Quinlan, J. R. 1995 Combining instance-based and model-based learning. In: Machine Learning Proceedings 1993, San Francisco, CA. Morgan Kaufmann, pp. 236–243. https://doi.org/10.1016/B978-1-55860-307-3.50037-X.

R Core Team 2020 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/. Accessed: 13 January 2020.

Ren, W. W., Yang, T., Huang, C. S., Xu, C. Y. & Shao, Q. X. 2018 Improving monthly streamflow prediction in alpine regions: integrating HBV model with Bayesian neural network. Stochastic Environmental Research and Risk Assessment 32(12), 3381–3396. https://doi.org/10.1007/s00477-018-1553-x.

Ribeiro, M., Singh, S. & Guestrin, C. 2016 Why should I trust you?: explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA. Association for Computational Linguistics, pp. 97–101. https://doi.org/10.18653/v1/N16-3020.

Robertson, D. E., Pokhrel, P. & Wang, Q. J. 2013 Improving statistical forecasts of seasonal streamflows using hydrological model output. Hydrology and Earth System Sciences 17(2), 579–593. https://doi.org/10.5194/hess-17-579-2013.

Sabzipour, B., Arsenault, R. & Brissette, F. 2020 Evaluation of the potential of using subsets of historical climatological data for ensemble streamflow prediction (ESP) forecasting. Journal of Hydrology 125656. https://doi.org/10.1016/j.jhydrol.2020.125656.

Schaeffi, B. & Gupta, H. V. 2007 Do Nash values have value? Hydrological Processes 21(15), 2075–2080. https://doi.org/10.1002/hyp.6825.

Song, S. & Wang, W. 2019 Impacts of antecedent soil moisture on the rainfall-runoff transformation process based on high-resolution observations in soil tank experiments. Water 11(2), 296. https://doi.org/10.3390/w11020296.

Tarboton, D. G. 2008 Terrain Analysis Using Digital Elevation Models (TauDEM). Available from: http://hydrology.usu.edu/taudem/. Accessed: 30 April 2019.

Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres 106(D7), 7183–7192. https://doi.org/10.1029/2000JD900719.

Tian, Y., Xu, Y.-P. & Zhang, X.-J. 2013 Assessment of climate change impacts on river high flows through comparative use of GR4J, HBV and Xinanjiang models. Water Resources Management 27(8), 2871–2888. https://doi.org/10.1007/s11269-013-0321-4.

Tyralis, H., Papacharalampous, G. & Langoussis, A. 2020 Super ensemble learning for daily streamflow forecasting: large-scale demonstration and comparison with multiple machine learning algorithms. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05172-5.

Worland, S. C., Farmer, W. H. & Kiang, J. E. 2018 Improving predictions of hydrological low-flow indices in ungauged basins using machine learning. Environmental Modelling & Software 101, 169–182. https://doi.org/10.1016/j.envsoft.2017.12.021.

Wyatt, B. M., Ochsner, T. E., Krueger, E. S. & Jones, E. T. 2020 In-situ soil moisture data improve seasonal streamflow forecast accuracy in rainfall-dominated watersheds. Journal of Hydrology 590, 125404. https://doi.org/10.1016/j.jhydrol.2020.125404.

Zambrano-Bigiarini, M. & Rojas, R. 2015 A model-independent particle swarm optimisation software for model calibration. Environmental Modelling & Software 43, 5–25. https://doi.org/10.1016/j.envsoft.2013.01.004.

First received 9 December 2020; accepted in revised form 12 February 2021. Available online 8 March 2021